CIENCIA



# Inteligencia artificial para todos (parte 2): ¿qué ocurre dentro de los modelos y en qué debemos fijarnos?

**VALERIA MAEDA GUTIÉRREZ Y JUAN JOSÉ OROPEZA VALDEZ**

La Dra. Maeda-Gutiérrez es Docente-Investigadora en la Unidad Académica de Ingeniería Eléctrica de la Universidad Autónoma de Zacatecas (UAZ). Su trabajo se centra en el desarrollo y aplicación de modelos de Inteligencia Artificial, Procesamiento de Imágenes y Aprendizaje Profundo en el ámbito biomédico, con énfasis en el estudio de complicaciones microvasculares asociadas a la Diabetes Tipo 2. Es miembro del Sistema Nacional de Investigadoras e Investigadores (SNII) nivel C en el área de Medicina y Ciencias de la Salud.

El Dr. Oropeza-Valdez es Profesor de Tiempo Completo en la División de Investigación de la Facultad de Medicina de la Universidad Nacional Autónoma de México (UNAM), adscrito a la Unidad de Investigación en Obesidad del Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (INCMNSZ). Su investigación se centra en el estudio de la nefropatía diabética mediante enfoques de metabolómica, bioinformática e inteligencia artificial, para identificar biomarcadores, comprender los mecanismos metabólicos asociados a la progresión de la enfermedad y contribuir al desarrollo de estrategias diagnósticas y de medicina personalizada.

*Esta publicación fue revisada por el comité editorial de la Academia de Ciencias de Morelos.*

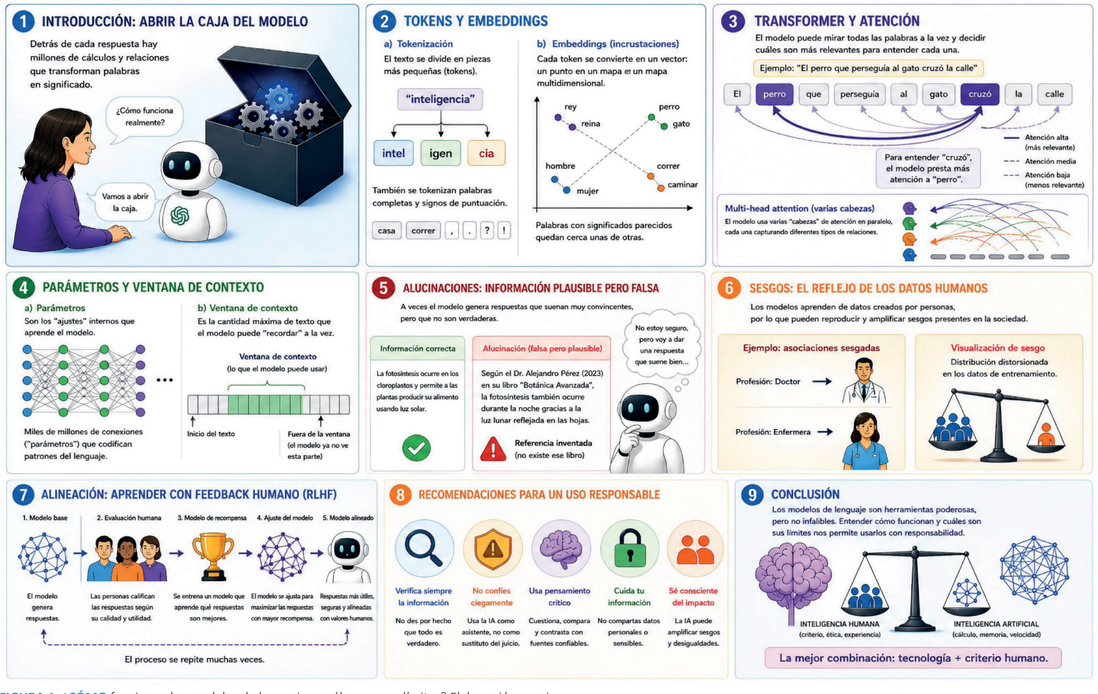


**FIGURA 1. ¿CÓMO** funcionan los modelos de lenguaje y cuáles son sus límites? Elaboración propia.

## Introducción

En un artículo anterior [1] explicamos qué son los Modelos Grandes de Lenguaje (en inglés, Large Language Models o LLMs), cómo aprenden y cómo generan texto palabra por palabra. Allí comparamos estos modelos con un "superautocompletador", capaz de sostener una conversación tras leer millones de páginas. Sin embargo, muchas preguntas quedaron sin responder. ¿Qué ocurre realmente dentro del modelo cuando escribimos una pregunta y recibimos una respuesta? ¿Cómo logra que las palabras, que para una computadora no significan nada, terminen conectándose en ideas coherentes? Y, sobre todo, ¿por qué a veces se equivoca con tanta seguridad? ¿Por qué refleja prejuicios humanos? ¿Y qué podemos hacer para usar estas herramientas de manera responsable? En esta segunda entrega abrimos la caja del modelo para mirar sus engranajes principales: cómo transforma el lenguaje en números, cómo aprende a prestar atención a las palabras relevantes, y cuáles son los límites y riesgos que debemos tener siempre presentes. No pretendemos convertir al lector en un ingeniero de inteligencia artificial, pero sí ofrecerle una mirada más informada que le permita usar estas herramientas con mayor claridad y espíritu crítico. Para facilitar la comprensión, hemos resumido los aspectos que analizaremos en la Figura 1, a la que pueden referirse para cada concepto.

### Del lenguaje a los números: tokens y vectores

Una computadora no entiende palabras; solo procesa números. El primer desafío para cualquier modelo de lenguaje es, por tanto, traducir el texto a cantidades con las que pueda operar. Este proceso inicia con la *tokenización*. Un *token* es una unidad de texto que el modelo reconoce como tal. En muchos casos coincide con una palabra completa ("casa", "correr"), pero con frecuencia los tokens son fragmentos más pequeños: prefijos, sufijos, raíces o incluso signos de puntuación. La palabra "inteligencia", por ejemplo, podría dividirse en tokens como "intel", "igen" y "cia". Esta estrategia permite al modelo manejar palabras que nunca había visto, combinando fragmentos que sí conoce [2].

Una vez que el texto se ha dividido en tokens, cada token se convierte en un vector numérico llamado *embedding* (o incrustación, en español). Podemos imaginar el embedding como un punto en un mapa de muchísimas dimensiones, donde palabras de significado similar quedan cerca entre sí. En ese mapa, los vectores de "perro" y "gato" estarán próximos porque ambos se refieren a mascotas; los de "rey" y "reina" tendrán una relación geométrica similar a la que existe entre "hombre" y "mujer". Lo sorprendente es que el modelo no aprende estas relaciones porque alguien se las explique, sino porque las deduce del uso que las palabras tienen en los textos con los que fue entrenado. Al leer millones de oraciones, el sistema descubre que ciertos términos aparecen en contextos similares y los ubica en regiones cercanas en ese mapa.

De esta manera, el lenguaje humano, con toda su riqueza y ambigüedad, se transforma en una geografía matemática que el modelo puede recorrer. Cada oración deja de ser una secuencia de símbolos para convertirse en una trayectoria por ese mapa, y las operaciones matemáticas sobre los vectores permiten medir similitudes, analogías y relaciones que antes sólo percibíamos intuitivamente los humanos.

### La arquitectura *Transformer*: aprender a prestar atención

Hasta 2017, los modelos de lenguaje procesaban el texto palabra por palabra, en orden, como si lo leyeran con un dedo marcando la línea. Esto hacía muy difícil recordar la información mencionada al inicio de un párrafo largo. Ese año, un grupo de investigadores de Google publicó un artículo con un título que se volvió célebre: *Attention is All You Need* ("La atención es todo lo que necesitas"). En él presentaron una nueva arquitectura llamada *Transformer*, que cambió el rumbo de la inteligencia artificial [3]. Todos los modelos modernos, desde ChatGPT hasta Claude, Gemini o DeepSeek, se basan en variantes de esta arquitectura.

La idea central del Transformer es un mecanismo llamado *atención*. En lugar de procesar las palabras en orden estricto, el modelo puede analizar simultáneamente todas las palabras del texto y determinar cuáles son las más relevantes para entender cada una. Consideremos la frase: "El perro que perseguía al gato cruzó la calle". Para entender quién cruzó la calle, el modelo necesita conectar "cruzó" con "perro", aunque en la oración estén separados por varias palabras. El mecanismo de atención le permite establecer esa conexión de forma directa, asignando mayor peso a "perro" y menor a las demás palabras al procesar el verbo "cruzó".

Esto es lo que hacemos los humanos cuando leemos un párrafo complicado: no memorizamos cada palabra por igual, sino que nuestra mente destaca automáticamente los términos clave y los conecta entre sí. El mecanismo de atención imita esta selección, pero lo hace de manera *paralela* para todas las palabras al mismo tiempo y con una sofisticación matemática que le permite considerar múltiples relaciones simultáneas. Un Transformer moderno realiza este cálculo de atención en docenas de "cabezas" distintas, cada una especializada en capturar un tipo diferente de relación: gramatical, semántica, temporal o de otra naturaleza.

Este mecanismo permite que un LLM sostenga el hilo de una conversación, recuerde detalles mencionados párrafos atrás y mantenga la coherencia en textos largos. Es lo que explica el enorme poder de cómputo y consumo de energía asociado a estos modelos: calcular la atención entre todas las palabras de un texto y hacerlo muchas veces por cada capa del modelo demanda billones de operaciones matemáticas por segundo.

### Parámetros y ventana de contexto: las dos escalas que importan

Cuando se habla de un modelo, suelen mencionarse dos cifras: el número de parámetros y el tamaño de la ventana de contexto. Los *parámetros* son los valores numéricos que el modelo ajusta durante el entrenamiento; podemos imaginarlos como las perillas internas que determinan cómo procesa la información. GPT-3 tenía alrededor de 175 mil millones de parámetros; los modelos más recientes superan el billón. Esta cifra se asocia con la capacidad general del modelo: más parámetros suelen implicar una mayor capacidad para capturar patrones complejos, aunque también un mayor consumo de energía y de memoria.

La *ventana de contexto* es, por otra parte, la cantidad máxima de tokens que el modelo puede considerar simultáneamente al generar una respuesta. Podemos compararla con la memoria de trabajo de un lector: si la ventana es pequeña, el modelo solo recuerda las últimas frases; si es grande, puede tener presente un libro completo. Los modelos actuales manejan ventanas de decenas o cientos de miles de tokens, lo que equivale a cientos de páginas de texto. Esto tiene implicaciones prácticas importantes: un modelo con ventana amplia puede analizar un contrato legal extenso, una historia clínica completa o un código de programación de gran tamaño sin perder el hilo.

Conviene subrayar una idea clave: aunque la ventana de contexto puede ser amplia, no es infinita y lo que queda fuera de ella simplemente no existe para el modelo en ese momento. Además, una vez terminada la conversación, el modelo no conserva memoria de lo discutido, a menos que una aplicación específica implemente ese recuerdo de forma externa. Por ello, cuando alguien dice que su asistente "recuerda" conversaciones pasadas, en realidad, alguna pieza de software está reinyectando ese historial en la ventana de contexto cada vez que escribimos.

### Cuando el modelo se equivoca: *alucinaciones* y por qué ocurren

En la entrega anterior mencionamos, de paso, el fenómeno de las *alucinaciones*: cuando el modelo genera información que suena plausible, pero es falsa. Ahora conviene entender por qué ocurren. Un LLM no consulta una base de datos al responder; lo que hace es producir la continuación estadísticamente más probable a partir de la pregunta recibida. Si la información correcta estaba presente en sus datos de entrenamiento y fue reforzada muchas veces, el modelo tenderá a generarla con precisión. Pero si el tema es oscuro, o si las fuentes entrenadas son contradictorias, o si la pregunta se formula de manera poco habitual, el modelo puede construir una respuesta que suena convincente, pero carece de sustento.

Un caso frecuente es la invención de referencias bibliográficas. Si se le pide a un LLM que cite artículos científicos sobre un tema muy específico, puede generar nombres de autores, títulos y revistas que simplemente no existen, mezclando fragmentos de obras reales de manera verosímil. Esto ocurre porque el modelo aprendió el patrón general de cómo se ve una cita académica y lo reproduce, sin tener acceso real a la base de datos de publicaciones científicas. La frase fabricada cumple con el patrón, pero no con la realidad.

La alucinación no es un error ocasional sino una consecuencia estructural de cómo funcionan estos modelos. Un LLM no sabe cuándo no sabe. No existe, dentro del modelo base, un mecanismo que distinga entre lo que efectivamente aprendió y lo que está extrapolando. Por eso, ante preguntas fuera de su zona de confianza, responde con la misma seguridad aparente que dentro de ella. Investigaciones recientes están desarrollando técnicas para detectar estas alucinaciones, ya sea midiendo la consistencia de múltiples respuestas del modelo a una misma pregunta o conectándolo con herramientas externas que verifiquen los datos en tiempo real.

### Sesgos heredados: cuando el espejo devuelve nuestros prejuicios

Los LLM aprenden del lenguaje humano registrado en libros, artículos, redes sociales, foros y sitios web. Ese corpus es un reflejo enorme y desigual de la producción cultural humana, con todos sus desequilibrios. Si en los textos disponibles las mujeres aparecen con mayor frecuencia asociadas a roles domésticos y los hombres a posiciones de liderazgo, el modelo aprenderá a replicar esa asociación. Si ciertos grupos étnicos o regiones del mundo están subrepresentados, el modelo tendrá menos capacidad para hablar con precisión sobre ellos. Los *sesgos* no son un defecto añadido por los programadores; son un eco de los patrones presentes en los textos.

Los estudios realizados en los últimos años han documentado numerosos casos concretos. Los modelos de traducción automática, al traducir del inglés a lenguas con género gramatical, tendían a asumir que "the doctor" era hombre y "the nurse" era mujer. Los modelos utilizados para evaluar currículums laborales penalizaban los nombres asociados a ciertos grupos minoritarios. Los sistemas de generación de imágenes producían representaciones estereotipadas al pedirles que ilustraran profesiones. Cada uno de estos casos se ha corregido parcialmente, pero el problema de fondo persiste: entrenar modelos con datos masivos obtenidos de internet reproduce los sesgos presentes en dichos datos.

Este fenómeno tiene implicaciones serias cuando se emplean los LLM en decisiones con consecuencias reales: recomendaciones médicas, evaluaciones crediticias, filtros de contratación o criterios en el sistema de justicia. Un sesgo sutil en el modelo puede traducirse en desventajas sistemáticas para grupos enteros de personas. Por eso, el diseño responsable de estas tecnologías incluye auditorías especializadas para medir sesgos, conjuntos de datos más representativos y mecanismos explícitos para detectar y mitigar respuestas discriminatorias.

### La alineación: enseñarle al modelo a ser útil y seguro

Entre el modelo recién preentrenado y el asistente conversacional que usamos hoy hay una enorme diferencia. Un modelo recién salido del preentrenamiento es, literalmente, un generador de texto que continúa cualquier secuencia de palabras según la probabilidad estadística aprendida. Si se le da el inicio de una receta, continuará con pasos plausibles de cocina; si se le da una teoría de conspiración, la ampliará con más contenido del mismo tipo. Transformar esa capacidad bruta en una herramienta útil y responsable requiere un proceso llamado *alineación* [4].

La técnica más difundida para lograrlo se llama *aprendizaje por refuerzo con retroalimentación humana* (conocida como RLHF, por sus siglas en inglés). El procedimiento, simplificado, es el siguiente. Personas capacitadas formulan preguntas al modelo, reciben varias respuestas posibles y las ordenan de mejor a peor según criterios de utilidad, veracidad y seguridad. Con esas comparaciones se entrena un segundo modelo, llamado *modelo de recompensa*, que aprende a imitar el juicio humano. Finalmente, el LLM original se ajusta para producir respuestas que el modelo de recompensa prefiere. El resultado es un asistente que no solo completa texto, sino que también intenta ser útil y claro y evitar contenido dañino.

Este proceso tiene limitaciones. Los anotadores humanos tienen, a su vez, sus propios criterios culturales, lo que influye en lo que el modelo considera una "buena respuesta". Un modelo alineado con los valores de una comunidad puede resultar inadecuado en otra. Además, la alineación no es absoluta: existen técnicas para *jailbreak* (liberar) modelos, es decir, formular preguntas de manera que eviten los filtros y que el sistema termine respondiendo cosas que, en condiciones normales, evitaría. La investigación sobre la alineación segura y robusta es hoy uno de los campos más activos de la inteligencia artificial.

### Recomendaciones para un uso informado y prudente

Entender cómo funcionan los LLM por dentro no es un ejercicio académico: tiene consecuencias directas en la forma en que conviene usarlos. A partir de todo lo anterior, pueden extraerse algunas recomendaciones prácticas.

Primero, conviene tratar las respuestas de un LLM como un primer borrador, no como una fuente definitiva. Es excelente para generar ideas, organizar información, redactar borradores y explorar conceptos, pero la verificación final debe realizarse con fuentes confiables, especialmente cuando hay decisiones importantes en juego. Segundo, cuando se les pidan hechos específicos (fechas, nombres, cifras, referencias), hay que cotejarlos siempre. La fluidez de la respuesta no garantiza su exactitud. Tercero, al plantear una pregunta, conviene ofrecer un contexto claro: un buen *prompt* aporta información relevante, especifica el tono deseado y, cuando es posible, incluye ejemplos del tipo de respuesta que se busca. Cuarto, en temas sensibles (salud, finanzas, aspectos legales, decisiones sobre otras personas), el LLM puede orientar, pero el juicio final siempre debe ser humano y estar respaldado por profesionales. Quinto, al notar un sesgo o un error, vale la pena señalarlo; muchas interfaces permiten retroalimentar al sistema y esa información se utiliza para mejorar futuras versiones.

### Conclusión

Los modelos de lenguaje grandes son, sin exagerar, una de las tecnologías más notables de nuestro tiempo. Pero la fascinación que producen no debe hacernos olvidar su naturaleza. Debajo del diálogo fluido y de las respuestas aparentemente inteligentes hay matemática, estadística, enormes cantidades de datos y decisiones de diseño que llevan implícitas ciertas visiones del mundo. Entender los engranajes básicos (*tokens*, *embeddings*, atención, parámetros, alineación) nos permite usar estas herramientas con mayor provecho y menos sobresaltos.

También nos permite participar informadamente en las conversaciones públicas que empiezan a definir cómo se regulará esta tecnología. Decisiones sobre qué datos pueden usarse para entrenar modelos, cómo auditar sus sesgos, qué aplicaciones deberían requerir supervisión humana obligatoria y qué derechos tenemos frente a sistemas que toman o sugieren decisiones sobre nosotros, son debates abiertos que nos incumben a todos.

La inteligencia artificial, bien comprendida y bien usada, puede ser una aliada extraordinaria para aprender, investigar, crear y resolver problemas. Mal comprendida, corre el riesgo de amplificar errores, sesgos y dependencias. La diferencia, como casi siempre ocurre con las tecnologías poderosas, está en la educación y el pensamiento crítico con los que la acompañemos.

### Referencias

[1] Maeda-Gutiérrez, V., y Oropeza-Valdez, J. J. (2025). Inteligencia artificial para todos: ¿Cómo funcionan los modelos que conversan contigo? *La Unión de Morelos*, 15 de septiembre de 2025. Disponible en: https://acmor.org/publicaciones/inteligencia-artificial-para-todos-c-mo-funcionan-los-modelos-que-conversan-contigo

[2] Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1715–1725. doi: 10.18653/v1/P16-1162

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., y Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. Disponible en https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[4] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., y col. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744. Disponible: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

*Esta columna se prepara y edita semana con semana, en conjunto con investigadores morelenses convencidos del valor del conocimiento científico para el desarrollo social y económico de Morelos.*

ACADEMIA DE CIENCIAS DE MORELOS, A.C.

**ESTA PUBLICACIÓN FUE REVISADA POR EL COMITÉ EDITORIAL DE LA ACADEMIA DE CIENCIAS DE MORELOS**

Para actividades recientes de la academia y artículos anteriores puede consultar: www.acmor.org
**¿Comentarios y sugerencias?, ¿Preguntas sobre temas científicos? CONTÁCTANOS: coord.comite.editorial.acmor@gmail.com**