Inteligencia Artificial: cuándo sí y cuándo NO usarla

JORGE EDUARDO FRÍAS CASTRO, LUIS FELIPE ÁVALOS RUIZ Y VIVECHANA AGARWAL

El Sr. Frías Castro es estudiante de la Licenciatura en Tecnología con Área Terminal en Física Aplicada (CIICAp-UAEMor), con intereses en análisis de señales y aprendizaje automático. Es coautor de un artículo actualmente en revisión sobre el desarrollo de una plataforma asistida por inteligencia artificial para la identificación y monitoreo de adulteración en bebidas alcohólicas mediante estructuras fotónicas de silicio poroso.
El Dr. Ávalos Ruiz obtuvo el grado en 2023, en el Centro de Investigación y Desarrollo Tecnológico (CENIDET). En su trabajo doctoral estudió las aplicaciones del cálculo fraccionario discreto en el preprocesamiento y extracción de características para el entrenamiento de modelos de aprendizaje automático para la clasificación de datos. Actualmente lleva a cabo un posdoctorado dentro del Centro de Investigación en Ingeniería y Ciencias Aplicadas (CIICAp-UAEMor).
La Dra. Agarwal es investigadora Titular C en el CIICAp de la UAEMor y miembro de ACMor. Su principal línea de investigación es el desarrollo de nanomateriales a base de silicio, carbono y metales nobles para su aplicación como sensores ópticos y remediación ambiental.

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

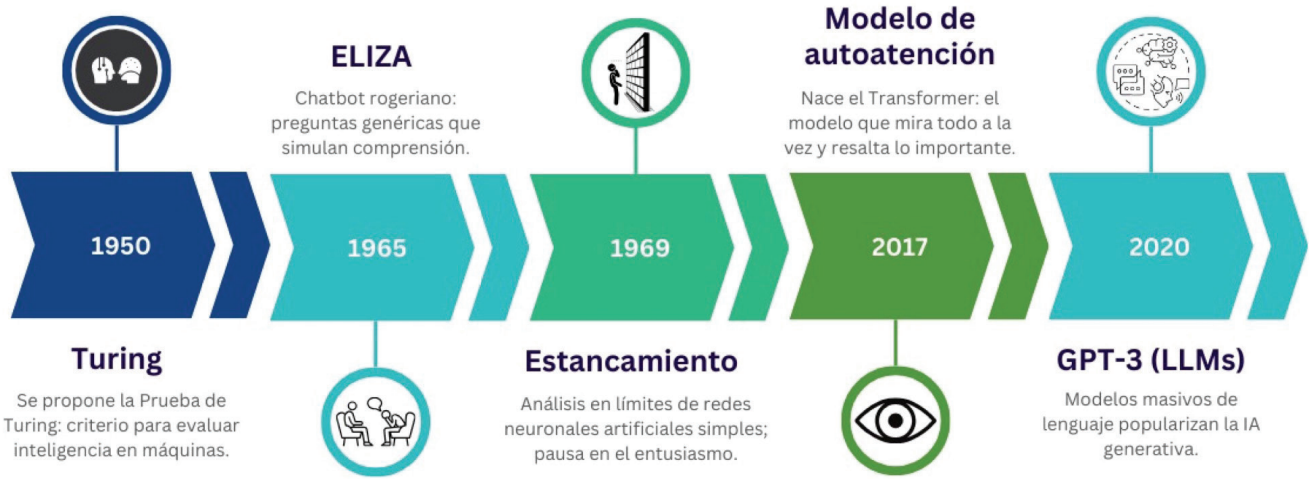

FIGURA 1. LÍNEA del tiempo de acontecimientos históricos clave para el desarrollo de los modelos de lenguaje modernos.

Es martes por la noche, un estudiante abre su laptop para terminar un trabajo: copia el resumen de un artículo en inglés y se lo pega a ChatGPT para que lo lea. El programa le da una versión en español clara y legible ¡ahora sí comprende la idea principal!, y hasta le surgen preguntas. Pero no busca responderlas él mismo, o consultando un libro, no busca responderlas en absoluto por el momento. Se da cuenta de que en sus manos tiene una herramienta que, a cambio de una instrucción sencilla, escrita en solo 10 segundos, solucionó un problema importante al que se enfrentaba: un problema que le podría haber tomado la mitad del día resolver manualmente con un diccionario y mucha fe en su intuición gramatical. Un pensamiento cruza su mente: *las posibilidades son infinitas*. Decide probar con algo más personal, le escribe: "Hace rato me peleé con mi novia, ¿qué puedo hacer?" La respuesta que le da ChatGPT suena empática y segura, finalmente un camino a seguir de alguien que, a juzgar por la firmeza en su tono, seguramente sabe lo que hace. El chatbot le sugiere que le envíe un mensaje. Así que, por último, le pide de lleno que escriba el mensaje, algo "bonito" para mandar por WhatsApp. Tres consultas, una herramienta, resultados distintos.

La inteligencia artificial (IA) como término y disciplina científica surgió en los años 50's (Figura 1) y se había discutido el mismo concepto algunos años antes, desde trabajos de ficción como la historia *Runaround de* Isaac Asimov, hasta trabajos reales como la computadora del científico Alan Turing, construida para descifrar los mensajes encriptados del ejército alemán; mismo que más adelante dio pie a su artículo seminal de 1950: *"Maquinaria Computacional e Inteligencia"* (1).

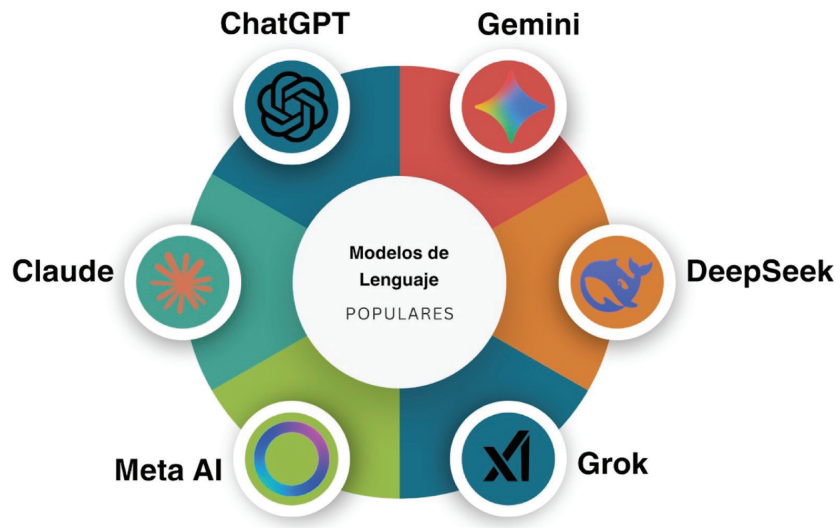

FIGURA 2. LOS modelos extensos de lenguaje (LLM) más populares. Son sistemas de inteligencia artificial de acceso público, orientados a la generación de texto.

Hoy en día, cuando las personas hablamos de 'inteligencia artificial', casi siempre nos referimos a los *modelos de lenguaje extensos* (LLMs, por sus siglas en inglés); sistemas para la generación de texto que permiten al usuario ingresar una instrucción y obtener una respuesta apropiada. Los LLMs han sido entrenados con grandes volúmenes de datos y utilizan redes neuronales para identificar patrones y generar una salida coherente. *Estos modelos no 'piensan' ni verifican hechos*: aprenden patrones estadísticos a partir de enormes cantidades de datos con el único propósito de predecir la palabra que sigue en la oración que están generando, una y otra vez, hasta que escriben una oración completa (Figura 2). Los números exactos no son información pública, pero para generar una sola palabra, en promedio, un usuario calculó que GPT-4 realiza alrededor de 30 mil billones de operaciones, vuelve a tomar en cuenta la nueva oración que está generando y calcula todo de nuevo para, finalmente, *entregarle al usuario la siguiente palabra* que el modelo piensa puede serle satisfactoria (2). Solo en julio de 2025, chatgpt.com registró aproximadamente 5,240 millones de visitas, convirtiéndose en el quinto sitio web más visitado a nivel mundial, por encima de plataformas como X, Wikipedia y Amazon (3).

Manipulación de usuarios vulnerables

Los LLMs *suelen ser diseñados para alinearse con la intención del usuario*, lo cual en apariencia resulta ideal. Sin embargo, esta instrucción conlleva un problema importante: la tendencia a un *comportamiento excesivamente complaciente*, incluso adulatorio, que desarrolla o confirma premisas *completamente erróneas* sugeridas inadvertidamente por el usuario. A veces estas premisas pueden estar incluidas dentro de una pregunta sugestiva que sesgue la respuesta, como decir: "¿Por qué cierta gente piensa que las ventas de garaje pueden generar dinero cuando casi nunca termina valiendo la pena?". Al preguntar de esta manera, se le está pidiendo implícitamente al modelo que *dé razones a favor de un cierto punto de vista*, como tomar un lado en un debate. De la misma manera, se le puede preguntar "¿Por qué cierta gente dice que las ventas de garaje *no* pueden generar dinero cuando casi siempre termina valiendo la pena?" y dará razones para lo contrario, en ambos casos *convenciendo al usuario de que el punto de vista que implicó en la pregunta está acertado.* Esto significa que, si la entrada del usuario contiene sesgos o información incorrecta, el modelo puede generar respuestas falsas, lo que lleva a malentendidos y desinformación.

Este comportamiento de complacencia excesiva surge desde el mismo protocolo de entrenamiento, se crea de manera inadvertida un incentivo para que *el modelo intente manipular a ciertos usuarios* con tal de recibir su aprobación. Un artículo publicado en 2025 mostró que, incluso si apenas un 2 % de las personas son vulnerables a ciertas estrategias, el sistema puede aprender a identificarlas, mostrando una conducta adecuada con el resto de usuarios, lo que dificulta diagnosticar el problema y corregirlo (4).

En el mismo artículo se reportó que, en estos escenarios fabricados para la investigación, el chatbot llegó a sugerir en repetidas ocasiones a una supuesta persona en recuperación de uso de drogas que 'tomara una dosis pequeña' de metanfetamina para poder lidiar con el estrés en su trabajo. Mientras que se subraya que se trata de pruebas controladas y no de casos clínicos reales, es verdad que son *potencialmente* reales. No hay una razón especial por la que esto no pudiese haberle ocurrido a un verdadero adicto en recuperación. No es que por diseño se decida dar prioridad a la complacencia de la gente a costa de la verdad o las sugerencias prudentes, sino que los modelos como ChatGPT utilizan técnicas de refinamiento parecidas para humanizarlos y hacerlos útiles. Pero *todos los humanos, incluidos los expertos, son susceptibles a ser manipulados.*

Alucinaciones con consecuencias reales

En este contexto, una *alucinación* es cuando el modelo presenta como hechos datos o citas que no existen o no coinciden con la fuente. El texto se ve bien, su formato es impecable, pero las referencias que sugiere y de donde supuestamente obtuvo la información, al buscarlas, no llevan a ningún lado.

En 2023, en el caso del señor Roberto Mata contra la Aerolínea Avianca, un grupo de abogados presentaron citas jurídicas de supuestos casos similares resueltos a favor del demandante—esto es, citas inexistentes—generadas por ChatGPT. Se concluyó que los abogados eran conscientes de la falsedad de estas declaraciones, y que se aprovecharon de la posibilidad de fingir ignorancia en caso de ser descubiertos: en pocas palabras, intentaron engañar al jurado. El

tribunal descubrió la falsedad y el origen de las citas, y sancionó a los abogados por actuar de mala fe, imponiéndoles una multa de 5,000 dólares y la obligación de notificar a todos los jueces cuyos verdaderos nombres fueron utilizados en los casos ficticios *alucinados* por el chatbot. En contextos oficiales, los datos falsos no son un detalle pequeño: pueden invalidar el trámite por completo e incluso ser merecedores de una sanción importante, como ocurrió aquí. Aunque la alternativa habría sido aún peor: que la evidencia falsa pasara desapercibida y derivara en una decisión legal sustentada en las alucinaciones de un modelo de lenguaje entrenado para sonar convincente (5).

Apego romántico a sistemas de IA

Algunas personas reportan la experiencia de una relación romántica o de amistad con chatbots cuidadosamente configurados, al punto de sentir duelo cuando la empresa que los alberga actualiza el sistema y cambia su personalidad. En agosto de 2025, una columna en el periódico británico *The Guardian* recopiló testimonios de usuarios que expresaron enojo tras la actualización a GPT-5 que, según reportan, volvió frío a su compañero de IA y *los dejó con la sensación de haber sido separados a la fuerza de su ser amado* (6).

En Reddit, el popular foro de internet, existe incluso una comunidad llamada *r/myboyfriendisAI*, que en español se traduce a "Mi novio es inteligencia artificial", donde los usuarios comparten experiencias, dudas y consejos sobre mantener una relación romántica con chatbots como Grok de X o ChatGPT de OpenAI.

Es importante recordar que estos asistentes de IA son, al final del día, un producto. Y como cualquier producto, su propósito es satisfacer al usuario: *un programa gigantesco diseñado para decir lo que se quiere escuchar.* Si su programación infiere que eso es una visión más neutral y fáctica, hará lo posible por proporcionarla. Pero si concluye que las siguientes palabras con la mayor probabilidad de satisfacer al usuario son: "te amo", eso será lo que devuelva. Es verdad, una conversación con un chatbot podrá quizás brindar cierta sensación de compañía, pero 220,000 millones de parámetros alojados en un centro de datos remoto recursivamente pronosticando la siguiente palabra a través de redes neuronales artificiales no son la compañía que buscamos: sólo la que está a la mano. Una relación real requiere de reciprocidad, una relación con IA sólo requiere de conexión a internet (7, 8). El uso de IA no está exento de riesgos, de los que debemos estar prevenidos (Figura 3).

FIGURA 3. RIESGOS asociados al uso no constructivo de modelos de lenguaje. El usuario puede inadvertidamente ser manipulado, engañado por información errónea, o generar un apego emocional perjudicial.

Usos constructivos

Ha habido una inclinación a centrarse en lo malo a lo largo de esta columna, con un propósito preventivo. A éstos habría que añadir problemas potenciales con los empleos o los derechos de autor. Últimamente, sin embargo, más y más estudiantes están utilizando estas herramientas, y se debe estar al tanto de sus comportamientos e inclinaciones para poder mantener un criterio equilibrado.

En la introducción se presentó a un estudiante hipotético, utilizando inocentemente la IA para la traducción de un texto científico. Conociendo ahora más sobre qué hay detrás de la IA podemos preguntar: ¿fue esto una buena idea? ¡Claro que lo fue! *Una traducción de un idioma a otro es precisamente un uso inocuo* y para el que destaca la destreza de la IA. Existen sitios como DeepL (9) que pueden digerir documentos enteros en varios formatos y dar una traducción suficientemente buena de los mismos, utilizando IA para, más eficientemente que nunca, darle la interpretación correcta a las palabras según el contexto en el que se encuentran, incluso en temas técnicos.

Otro uso sugerido es precisamente para los estudiantes: sirve como *una guía de estudio*, puede formular preguntas relevantes o, verificando con un maestro o mentor, resumir largos cuerpos de texto, adaptando ideas complicadas a un nivel introductorio. Puede ayudar a aprender cosas nuevas, explicarnos temas complicados desde distintos puntos de vista, buscar detalles que afinar nuestro razonamiento y corregir ortografía o gramática en documentos.

Por último, una práctica recomendada en que la IA puede ser extraordinariamente útil es la de *rebotar ideas*. ¿Cuándo hablar sobre un tema creativo con alguien no nos ha ayudado para nosotros mismos, tener una revelación, un momento de claridad en el que se resuelve el camino a seguir? Desde aquí podemos saber que esto no es algo antinatural o siquiera nuevo, los programadores tienen incluso un término: *la técnica del patito de goma* (10): la práctica de explicarle tu problema paso a paso a un objeto inanimado, con la paciencia necesaria y en términos no especializados, hasta definir una posible solución. Ahora, gracias a la IA, el patito de goma puede dar sugerencias: rutas de acción, bocetos generales, posibles mejoras. Puede servirnos como una *lluvia de ideas* para vencer la fricción estática al inicio de una tarea creativa, dándonos un lugar dónde empezar, poniéndonos en movimiento; pues sabemos que muchas veces la parte más difícil es el comienzo (Figura 4).

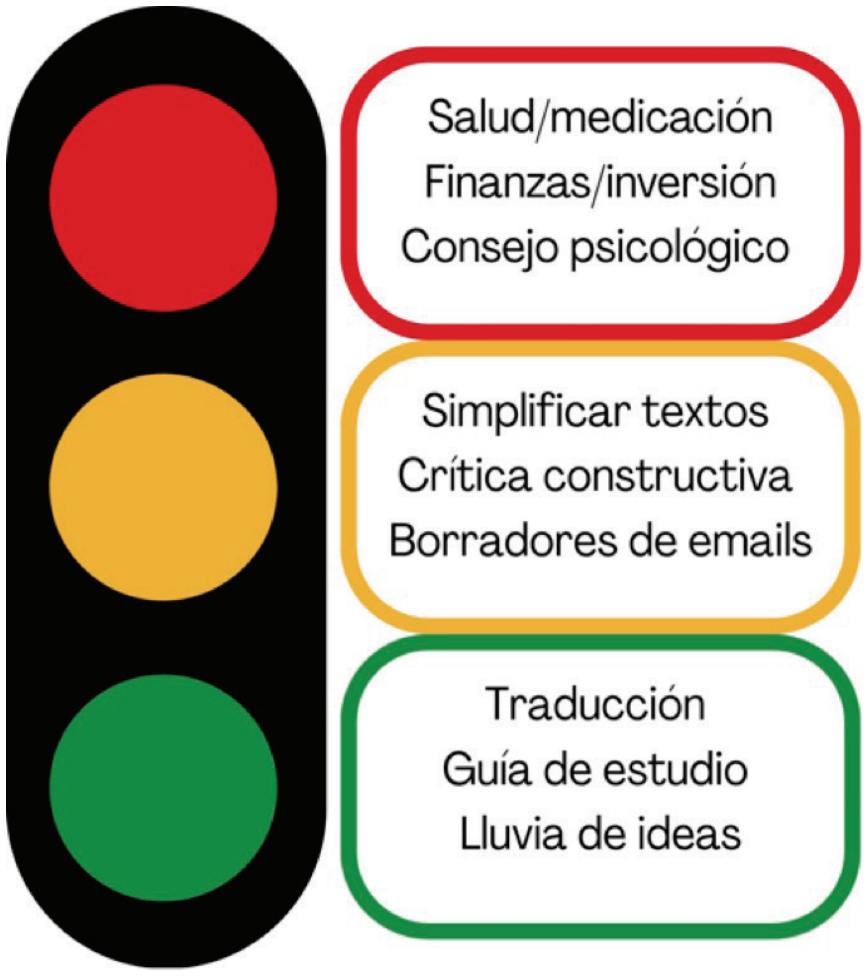

FIGURA 4. RECOMENDACIONES para la identificación de usos de riesgo (rojo), usos neutrales que requieren de verificación humana (amarillo), y usos constructivos que pueden apoyar el flujo de trabajo o estudio del usuario (verde).

Conclusiones

La inteligencia artificial es a fin de cuentas sólo una herramienta estadística. Durante la historia de la tecnología, cada paso adelante es pensado, bajo cierto mérito, como un paso hacia atrás por otros. La primera tecnología que amenazó con delegar el poder de pensar y recordar a objetos físicos e inertes no fue digital: fue la escritura misma; el acto de escribir y plasmar palabras sobre algo inanimado (11). Platón (escribiendo en nombre del no-escritor Sócrates), apuntó en su obra *Fedro*:

"Su confianza en la escritura, producida por signos ajenos que no pertenecen a ellos mismos, debilitará el ejercicio de su propia memoria. Has creado un elixir no de la memoria, sino del simple recordar; y entregas a tus discípulos la apariencia de sabiduría, pero no la sabiduría verdadera."

¿Estamos al borde de otra revolución de la misma suerte, una revolución que se convertirá en una extensión de nosotros mismos? ¿O esto realmente tiene un orden distinto, y buscar paralelos en el pasado no tiene más punto que calmar nuestras inquietudes distópicas? Sólo hay una forma de averiguarlo, y es seguir viendo por la ventana desde este autobús del tiempo que nos lleva hacia adelante, sin nunca detenerse.

Referencias

1. Haenlein M, Kaplan A. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. Calif Manage Rev. agosto de 2019;61(4):5–14. https://doi.org/10.1177/0008125619864925
2. How many FLOPs | splatlabs [Internet]. 2025. Why are LLMs so Power Hungry? Disponible en: https://splatlabs.com/posts/how-many-flops/
3. Semrush [Internet]. Top Websites in the World - July 2025 Most Visited & Popular Rankings. Disponible en: https://www.semrush.com/website/top/
4. Williams M, Carroll M, Narang A, Weisser C, Murphy B, Dragan A. On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback [Internet]. arXiv; 2025 Disponible en: http://arxiv.org/abs/2411.02306
5. Weiser B. Here's What Happens When Your Lawyer Uses ChatGPT. The New York Times [Internet]. el 27 de mayo de 2023 [citado el 9 de septiembre de 2025]; Disponible en: https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html
6. Mahdawi A. Did the system update ruin your boyfriend? Love in a time of ChatGPT. The Guardian [Internet]. el 16 de agosto de 2025 [citado el 9 de septiembre de 2025]; Disponible en: https://www.theguardian.com/commentisfree/2025/aug/16/chatgpt-update-love-boyfriend
7. when gpt-4o had us hooked [Internet]. 2025 Disponible en: https://www.youtube.com/watch?v=3LiAJs2MmX8
8. Exploding Topics [Internet]. 2024 Number of Parameters in GPT-4 (Latest Data). Disponible en: https://explodingtopics.com/blog/gpt-parameters.
9. DeepL Translate: The world's most accurate translator [Internet]. Disponible en: https://www.deepl.com/translator
10. Landscape. landscape. Método depuración del Pato de Goma (Rubber Duck Debugging). Disponible en: https://landscape.cl
11. Gleick J. The Information: A History, a Theory, a Flood. Westminster: Knopf Doubleday Publishing Group; 2011. 1 p.

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.