

# ¿ADIÓS A SÓCRATES?

por **Wolfram Eilenberger**

La filosofía opera cuestionado lo que creemos saber. La duda, la pregunta sin respuesta, es una herramienta de conocimiento. Por el contrario, ChatGPT simula saber lo que no sabe.

JULIO 2023

12

Como muchos de mis congéneres, he pasado las últimas semanas tratando de llevar hasta sus límites conversacionales una nueva forma de inteligencia artificial llamada ChatGPT. Lo hacía sobre todo con preguntas que planteaba a futuros becarios en una vida profesional anterior. Por ejemplo: “¿Qué hora es en el sol?”, “¿qué diferencia hay entre ‘aparentemente’ y ‘a primera vista’?”. O incluso: “¿Cuántos géneros no existen?” Preguntas, en otras palabras, que, aunque no sean directamente metafísicas, exigen una intensa plasticidad mental.

Debo decir que la maquinita resistió sorprendentemente bien. Solo a la tercera pregunta capituló como le correspondía y –sin la menor duda– emitió una bienintencionada nota de prensa sobre el estado actual de la visión progresista de los géneros.

¿Y después? Pues nada más. Solo un cursor parpadeante, demasiado preparado para afrontar el siguiente reto que responder. Fue el momento en que me llamó la atención por primera vez la potencial naturaleza abismal de la innovación. Más concretamente, la imagen de una cultura venidera que vería todos los ideales del diálogo socrático convertidos en su contrario. Una cultura sin filosofía en el verdadero sentido de la palabra.

A diferencia de Sócrates, el chatbot ni siquiera sabe que no sabe nada. Por eso, en su mercado, solo opera con respuestas supuestamente basadas en hechos y no con preguntas sin respuesta. En lugar de esforzarse por mostrar ideas eternas, cada una de sus afirmaciones se basa en



probabilidades siempre cambiantes. En lugar de desconfiar profundamente de la escritura como medio, como hacía Sócrates, porque permitía fingir el conocimiento, los chatbots se basan en la simulación del conocimiento a partir de textos escritos. En lugar de detenerse asombrado ante las preguntas más elevadas, el chatbot siempre ofrece alguna tontería inventada libremente, aunque no exista ningún dato para sustentarla. En lugar de sopesar las voces que participan en una conversación libre, se basa en su mera recopilación y recuento. En lugar de cuestionar productivamente la autoridad, iguala toda forma de autoridad evolucionada. En lugar de esforzarse por salir de la cueva de lo meramente creído con cada nuevo término, sus parpadeantes palabras en la penumbra se hacen pasar por la realidad misma. En lugar de ser impulsado por su propio *demon*, el devenir del chatbot es impulsado por el anónimo *mammon*. En lugar de buscar su propia voz, imita a la perfección la de todos los demás. En lugar de emanciparnos cada vez más como seres que aprenden, amenaza con dejarnos a todos en la condición de becarios permanentes.

No cabe duda de que ha comenzado una nueva era. Lo único que queda por aclarar es para qué género representa un progreso no solo aparente. Desde luego, no para el de los hombres socráticos. ¿O no es así? ~

Traducción del alemán de Lola Rodríguez.

**WOLFRAM EILENBERGER** es filósofo. Su libro más reciente es *¿Sufren las piedras?* (Taurus, 2023).

# LA IA COMO AGENCIA SIN INTELIGENCIA:

## SOBRE CHATGPT, GRANDES MODELOS LINGÜÍSTICOS Y OTROS MODELOS GENERATIVOS

por **Luciano Floridi**

Los modelos masivos de lenguaje son muy potentes y probablemente serán muy útiles, pero no deberían sustituirnos. El mejor autor es un ser humano que utiliza un LLM de forma competente y perspicaz.

La primera idea es antigua: todos los textos están presentes en el diccionario, la diferencia la establece la sintaxis, es decir, la forma en que las palabras del diccionario se estructuran en oraciones. La segunda idea es antigua: todas las palabras del diccionario están presentes en el alfabeto, la diferencia la marca la morfología, es decir, cómo se estructuran las letras del alfabeto en palabras. La tercera idea es antigua: todas las letras están presentes en el código digital, la diferencia la establece la forma en que las cadenas finitas de ceros y unos del código digital se estructuran en letras. La cuarta idea también es antigua: todas las cadenas de ceros y unos están presentes en dos propiedades electromagnéticas, corriente alta o baja, magnetización presente o ausente, la diferencia la establece la forma en que dichas propiedades pueden ser manejadas por los dispositivos electrónicos de cálculo. Pero la quinta idea es revolucionaria: hoy en día, la Inteligencia Artificial (IA) maneja las propiedades del electromagnetismo para procesar textos con un éxito extraordinario y, a menudo, de forma indistinguible a como serían capaces de hacerlo los seres humanos. Son los llamados modelos masivos de lenguaje (LLM por las siglas de *large language models*), y es lógico que estén causando sensación.

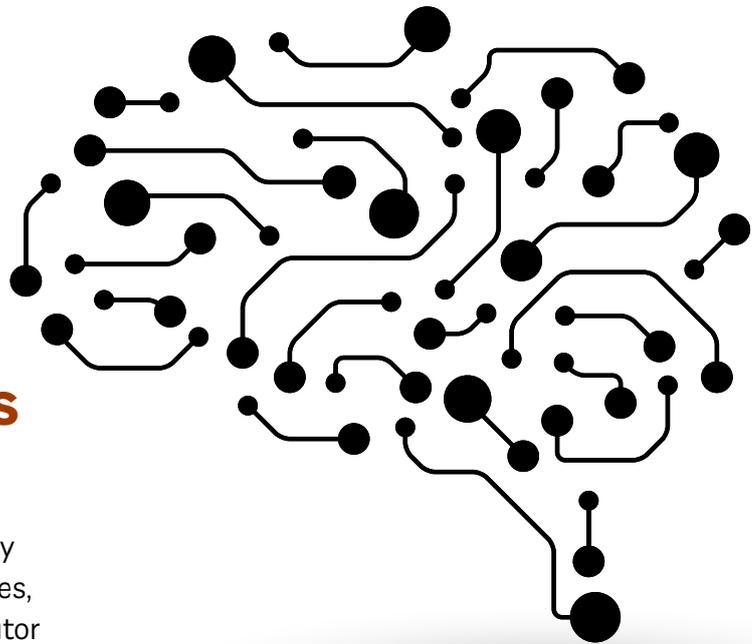
Los LLM más famosos son GPT3, ChatGPT (también conocido como GPT3.5, producido por OpenAI-Microsoft) y Bard1 (desarrollado por Google). No razonan ni comprenden, no son un paso hacia ninguna IA de ciencia ficción, y no tienen nada que ver con los procesos cognitivos presentes en el mundo animal, y sobre todo en el cerebro y la mente humana, para gestionar contenidos semánticos con éxito. Sin

embargo, con el asombroso crecimiento de los datos disponibles, la cantidad y la velocidad de cálculo, y unos algoritmos cada vez mejores, pueden hacer estadísticamente —es decir, trabajando sobre la estructura formal, y no sobre el significado de los textos que tratan— lo que nosotros hacemos semánticamente, aunque de formas que aún están siendo exploradas por la neurociencia. Sus capacidades son extraordinarias, como deben admitir incluso los más escépticos. He aquí un resumen de la *Divina comedia* realizado por ChatGPT:



Podemos criticar este resumen porque supera las cincuenta palabras, y porque la *Divina comedia* no es un poema épico —aunque existe un debate sobre este tema en internet, de ahí el resumen de ChatGPT—, sino más bien una tragedia, como sugirió el propio Dante. Dicho esto, el resumen no está mal, y desde luego es mejor que uno elaborado por un

1 Resume la *Divina comedia* en cincuenta palabras: La *Divina comedia* es un poema épico escrito por Dante Alighieri a principios del siglo XIV. Es una narración que describe el viaje del poeta por el Infierno, el Purgatorio y el Cielo. La obra está ampliamente considerada como una obra maestra de la literatura universal y ofrece una profunda reflexión sobre la naturaleza del pecado, la redención y la vida después de la muerte. El poema es famoso por sus imaginativas descripciones, su rico simbolismo y su exploración de la condición humana.

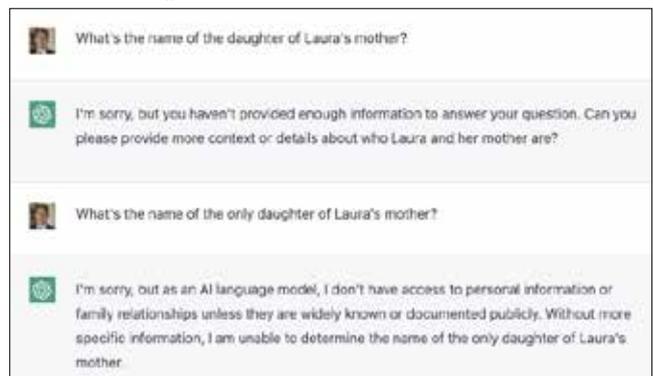


estudiante mediocre. El ejercicio ya no consiste en hacer resúmenes sin usar ChatGPT, sino en aprender a usar los *prompts* adecuados (la pregunta o petición que genera el texto, véase la primera línea de mi petición), comprobar el resultado, saber qué corregir en el texto producido por ChatGPT, descubrir que existe un debate sobre qué género literario se aplica mejor a la *Divina comedia*, y de paso, al hacer todo esto, aprender muchas cosas no solo sobre el *software* sino sobre todo sobre la propia *Divina comedia*. Como explicaba a mis alumnos de Oxford en los años noventa, uno de los mejores ejercicios para escribir un ensayo sobre las *Meditaciones* de Descartes no consiste en resumir lo que ya se ha dicho, sino en coger el texto electrónico de una de las *Meditaciones* e intentar mejorar su traducción (así se aprende a revisar el original); aclarar los pasajes menos claros con una paráfrasis más accesible (así uno ve si realmente ha entendido el texto); intentar criticar o refinar los argumentos, modificándolos o reforzándolos (así uno se da cuenta de que otros han intentado hacer lo mismo, y eso no es tan fácil); y mientras se hace todo eso aprender la naturaleza, la estructura interna, la dinámica y los mecanismos del contenido sobre el que se está trabajando.

O, por cambiar el ejemplo, uno conoce realmente un tema no cuando sabe escribir una entrada de Wikipedia sobre él —esto lo puede hacer ChatGPT cada vez mejor—, sino cuando sabe corregirlo. Hay que utilizar el *software* como una herramienta para meter las manos en el texto/mecanismo, y ensuciárselas incluso estropeándolo, siempre y cuando se domine la naturaleza y la lógica del artefacto llamado texto. Las limitaciones de estos LLM son ahora evidentes incluso para los más entusiastas. Son frágiles, porque cuando no funcionan, fracasan catastróficamente, en el sentido etimológico de caída vertical e inmediata en el rendimiento. El desastre de Bard, que proporcionó información incorrecta en una demostración fallida que costó a Google más de 100.000 millones de dólares en pérdidas bursátiles, es un buen recordatorio de que hacer las cosas con cero inteligencia, ya sea digital o humana, a veces es muy doloroso (el chat de Bing también tiene sus problemas).

En la actualidad existe una línea de investigación que produce análisis muy sofisticados sobre cómo, cuándo y por qué estos LLM, que parecen incorregibles, tienen un número ilimitado de talones de Aquiles (cuando se le preguntó cuál era su talón de Aquiles, ChatGPT respondió correctamente diciendo que es solo un sistema de IA). Se inventan textos, respuestas o referencias cuando no saben cómo responder; cometen errores evidentes en los hechos; a veces fallan en las inferencias lógicas más triviales o tienen problemas con matemáticas sencillas; o tienen extraños puntos ciegos lingüísticos en los que se atascan. Un sencillo ejemplo en inglés ilustra bien los límites de un mecanismo que gestiona textos sin comprender nada de su contenido. A la pregunta —utilizando el genitivo sajón— de cómo se llama

la única hija de la madre de Laura, la respuesta es entrañablemente estúpida:



Olvídate del Test de Turing. Si yo hubiera sido Google, no habría invertido la fortuna de mi empresa en un mecanismo tan frágil. Dados los enormes éxitos y las igualmente amplias limitaciones, algunos han comparado los LLM con loros estocásticos que repiten textos sin entender nada. La analogía ayuda, pero solo parcialmente, no solo porque los loros tienen una inteligencia propia que sería la envidia de cualquier IA sino, sobre todo, porque las IA sintetizan textos de nuevas maneras, reestructurando los contenidos sobre los que han sido entrenados, no proporcionando simples repeticiones o yuxtaposiciones. Se parecen mucho más a la función de autocompletar de un motor de búsqueda. Y en su capacidad de síntesis, se acercan a esos estudiantes mediocres o perezosos que, para escribir un ensayo breve, utilizan una docena de referencias relevantes sugeridas por el profesor y, cogiendo un poco de aquí y otro poco de allá, arman un texto ecléctico, coherente, pero sin haber entendido gran cosa ni haber añadido nada. Como tutor universitario en Oxford, corregía muchos de ellos cada trimestre. Ahora, ChatGPT puede producirlos con mayor rapidez y eficacia.<sup>2</sup>

Por desgracia, la mejor analogía que conozco para describir herramientas como ChatGPT está culturalmente delimitada, y hace referencia a un gran clásico de la literatura italiana: *Los novios* de Alessandro Manzoni (1827). En una famosa escena en la que Renzo (uno de los protagonistas) conoce a un abogado, leemos: “Mientras el doctor [el abogado] soltaba todas esas frases, Renzo lo miraba con una atención estática, como un paleta [*materialone*] que está en la plaza mirando al titiritero [*giocatore di bussolotti*], que, tras haberse metido en la boca estopa y más estopa, saca cinta

2 —¿Cómo se llama la hija de la madre de Laura? —Lo siento, pero no has proporcionado suficiente información para responder a tu pregunta. ¿Puedes proporcionar más contexto o detalles sobre quiénes son Laura y su madre? —¿Cómo se llama la única hija de la madre de Laura? —Lo siento, pero como modelo lingüístico de IA, no tengo acceso a información personal o relaciones familiares a menos que sean ampliamente conocidas o estén documentadas públicamente. Sin información más específica, no puedo determinar el nombre de la única hija de la madre de Laura.

y más cinta, que nunca se acaba [la palabra “nastro” se traduciría más correctamente como “lazo”, pero “cinta” es preferible en este contexto, ya que recuerda a la interminable cinta de una máquina de Turing].” Las LLM son como ese titiritero: engullen datos en cantidades astronómicas y regurgitan información. Si necesitamos la “cinta” de su información, conviene prestar mucha atención a cómo se produjo, por qué y con qué impacto. Y aquí llegamos a cosas más interesantes.

El impacto de los LLM y de los diversos sistemas de IA que hoy producen contenidos de todo tipo será enorme. Basta pensar en DALL-E, que, como dice ChatGPT (cito sin ninguna modificación), “es un sistema de Inteligencia Artificial desarrollado por OpenAI que genera imágenes originales a partir de descripciones textuales. Utiliza técnicas de aprendizaje automático de última generación para producir imágenes de alta calidad que coincidan con el texto de entrada, incluidos pies de foto, palabras clave e incluso frases sencillas. Con DALL-E, los usuarios pueden introducir una descripción textual de la imagen que desean, y el sistema producirá una imagen que coincida con la descripción”. Hay cuestiones éticas y jurídicas: basta pensar en los derechos de autor y los derechos de reproducción vinculados a las fuentes de datos sobre las que se entrena la IA en cuestión. Ya han empezado los primeros pleitos, y ya se han producido los primeros escándalos de plagio. Hay costes humanos: pensemos en el uso de contratistas en Kenia, a los que se pagó menos de dos dólares/hora por etiquetar contenidos nocivos para entrenar a ChatGPT; no pudieron acceder a recursos de salud mental adecuados, y muchos han quedado traumatizados.

Hay problemas humanos, como el impacto en los profesores que tienen que apresurarse a renovar su plan de estudios, o consideraciones de seguridad, por ejemplo, en relación con los resultados de los procesos de IA que se integran cada vez más en los diagnósticos médicos, con implicaciones de envenenamiento algorítmico de los datos de entrenamiento de la IA. O pensemos en los costes financieros y medioambientales de estos nuevos sistemas: ¿es justo y sostenible este tipo de innovación? Luego están las cuestiones relacionadas con el mejor uso de estas herramientas, en la escuela, en el trabajo, en entornos de investigación y para publicaciones científicas, en la producción automática de código, o en la generación de contenidos en contextos como la atención al cliente, o en la redacción de cualquier texto, incluidos artículos científicos o nueva legislación. Algunos empleos desaparecerán, otros ya están surgiendo y muchos tendrán que replantearse.

Pero sobre todo, para un filósofo, hay muchas preguntas desafiantes sobre: la aparición de sistemas de IA tipo LEGO, que trabajen juntos de forma modular y sin fisuras, con los LLM actuando como una especie de puente AI2AI [una IA interacciona con otra IA] para hacerlos interoperables, como una especie de “IA confederada”; la relación entre la forma

y su sintaxis, y el contenido y su semántica; la naturaleza de la personalización del contenido y la fragmentación de la experiencia compartida (la IA puede producir fácilmente una novela única a la carta, para un único lector, por ejemplo); el concepto de interpretabilidad, y el valor del proceso y el contexto de la producción de significado; nuestra singularidad y originalidad como productores de significado y sentido, y de nuevos contenidos; nuestra capacidad para interactuar con sistemas cada vez más indistinguibles de otros seres humanos; nuestra sustituibilidad como lectores, intérpretes, traductores, sintetizadores y evaluadores de contenidos; el poder como control de las preguntas, porque, parafraseando *1984*, quien controla las preguntas controla las respuestas, y quien controla las respuestas controla la realidad. Surgirán más preguntas a medida que desarrollemos, interactuemos y aprendamos a comprender esta nueva forma de agencia. Como me recordó Vincent Wang, ChatGPT superó a GPT3 en rendimiento gracias a la introducción del aprendizaje por refuerzo (RL) para afinar sus resultados como interlocutor, y el RL es el enfoque de aprendizaje automático para “resolver la cuestión de la agencia”. Es una forma de agencia nunca vista, porque tiene éxito y puede “aprender” y mejorar su comportamiento sin necesidad de ser inteligente. Es una forma de agencia ajena a cualquiera del pasado, porque la humanidad ha visto siempre y en todas partes este tipo de agencia –que no es la de una ola de mar, que establece una diferencia pero no puede ir más allá de ella, ya que no puede aprender a hacer eso de manera distinta o mejor– como una forma natural o incluso sobrenatural de agencia. Hemos pasado de estar en contacto permanente con agentes animales y lo que creíamos que eran agentes espirituales (dioses y fuerzas de la naturaleza, ángeles y demonios, almas o fantasmas, espíritus buenos y malos) a tener que entender, y aprender a interactuar, con agentes artificiales creados por nosotros; somos los nuevos demiurgos de ese tipo de agencia. Ya no hace falta ser inteligentes, entender, reflexionar, considerar o comprender algo para actuar con éxito.

Hemos liberado el albedrío de la inteligencia. Así que no estoy seguro de que podamos ser “pastores del Ser” (Heidegger), pero parece que los nuevos “cuellos verdes” serán “pastores de sistemas de IA”, a cargo de esta nueva forma de agencia artificial. El orden de esta humanidad demiúrgica de la IA sin inteligencia (como sin grasa) –entendida como *agere sine intelligere*, con un poco de latín de bachillerato– está aún por determinar. Puede ser alarmante o emocionante para muchos, pero sin duda es una buena noticia para los filósofos que buscan trabajo. ~

*Traducción del inglés de Lola Rodríguez.*

*Publicado en la revista Philosophy and Technology.*

**LUCIANO FLORIDI** es profesor de filosofía y ética de la información en el Oxford Internet Institute de la Universidad de Oxford.