Los grandes modelos de lenguaje: una oportunidad para la profesión bibliotecaria

Large language models: an opportunity for the library profession

Jorge Franganillo

Franganillo, Jorge (2023). "Los grandes modelos de lenguaje: una oportunidad para la profesión bibliotecaria". *Anuario ThinkEPI*, v. 17, e17a28.

https://doi.org/10.3145/thinkepi.2023.e17a28

Publicado en *IweTel* el 22 de septiembre de 2023

Jorge Franganillo
https://orcid.org/0000-0003-4128-6546

Universitat de Barcelona
Facultat d'Informació i Mitjans Audiovisuals
Centre de Recerca en Informació, Comunicació i Cultura (*CRICC*)
franganillo@ub.edu

Resumen: La Inteligencia Artificial (IA) generativa y los grandes modelos de lenguaje pueden cambiar la forma en que consultamos, procesamos y producimos información. Pero presentan desafíos técnicos y éticos, tales como inconsistencias, sesgos y falta de transparencia. El colectivo bibliotecario tiene aquí un papel clave, una oportunidad para apoyar el uso responsable de esta tecnología y promover la comprensión crítica de sus limitaciones. Las bibliotecas, por su parte, pueden ofrecer espacios y recursos para experimentar con la IA generativa y fomentar su uso en la investigación científica.

Palabras clave: Inteligencia Artificial; IA; Grandes modelos de lenguaje; Profesión bibliotecaria; Búsqueda generativa; Sesgo; Ética de la tecnología.

Abstract: Generative artificial intelligence (AI) and large language models can change the way we search, process and generate information. However, they also pose ethical and technical challenges such as inconsistencies, biases and lack of transparency. In this context, librarians play a key role, as they have the opportunity to support responsible use of this technology as well as to promote critical understanding of its limitations. Libraries, in turn, can offer spaces and resources to experiment with generative AI and encourage its use in scientific research.

Keywords: Artificial intelligence; AI; Large language models; Library profession; Generative search; Bias; Ethics of technology.

1. Los grandes modelos de lenguaje

La Inteligencia Artificial (IA) generativa es una rama de la informática que se ocupa de la creación automática de contenidos de diversa naturaleza, como textos, imágenes, voces clonadas y vídeos. Para ello, emplea técnicas de aprendizaje profundo y modelos entrenados con grandes volúmenes de datos con los que es capaz de producir contenido original y realista que a menudo resulta indistinguible del generado por humanos.

En estos últimos años, esta tecnología ha experimentado un auge sin precedentes, ha marcado tendencia y ha dominado los titulares. Al principio, nos maravillamos con la novedad y la diversión que ofrecía. Pero ahora comprobamos que es mucho más; tiene un impacto real y abre nuevas posibilidades en ámbitos tan diversos como el arte, la educación, la salud o el periodismo. Su potencial transformador es innegable.

Sin embargo, la IA generativa no está exenta de limitaciones y riesgos, tanto técnicos como éticos, sociales y legales. Por ello, conviene tener conciencia de las implicaciones de esta tecnología y de nuestras responsabilidades como usuarios, así como de los principios y las buenas prácticas que deben guiar su uso.

Una de las tecnologías más relevantes y populares, dentro del amplio campo de la IA, son los grandes modelos lingüísticos, esto es, sistemas que pueden procesar y generar lenguaje natural, que es el que usamos los humanos para comunicarnos. Se basan en redes neuronales artificiales, inspiradas en el funcionamiento del cerebro humano, que se entrenan con una ingente cantidad de texto procedente de fuentes como libros, periódicos o páginas web, y pueden producir textos coherentes y fluidos sobre cualquier tema.

Los modelos de lenguaje han progresado mucho gracias al desarrollo de arquitecturas complejas y potentes, tales como *GPT*, de *OpenAI*; *LLaMA*, de *Meta*; *PaLM*, de *Google*; o *Claude*, de *Anthropic*. Estos modelos destacan en la resolución de diversas tareas lingüísticas y han dado lugar a una gran variedad de aplicaciones. Como consecuencia, estamos asistiendo a una explosión cámbrica de tecnologías generativas (**Griffith**; **Metz**, 2023) que están causando un gran impacto social y un intenso debate.

Estas tecnologías ofrecen hoy una interfaz sencilla que se puede interrogar usando nuestro propio lenguaje natural, lo que las hace más accesibles para el público general. Los *chatbots* o asistentes de IA, como *Bard*, *ChatGPT* o *Claude*, utilizan técnicas de procesamiento de lenguaje natural para ofrecer respuestas pertinentes, en tiempo real, a casi cualquier consulta, simulando una conversación humana. Pero estas posibilidades también nos enfrentan a nuevos retos y responsabilidades. Por ello, es importante considerar sus consecuencias éticas, sociales y legales, y reflexionar sobre nuestro rol como usuarios.

En este contexto, el colectivo bibliotecario tiene un papel fundamental como agente educador en el uso de la IA generativa. Como profesionales de la información, los bibliotecarios pueden orientar y formar a los usuarios en la creación, evaluación y aprovechamiento de los contenidos generados por la IA. Y pueden contribuir también a la difusión y al desarrollo de la IA generativa, fomentando su uso responsable, crítico y creativo.

Sin embargo, interesa tener presente que los modelos de lenguaje no son infalibles. Pueden arrojar respuestas inexactas o erróneas, aunque suenen convincentes. Por eso, los profesionales de la información tienen que estar al tanto de las posibles imperfecciones de estos sistemas, han de ayudar a los usuarios a identificarlas y corregirlas, y deben concienciarlos sobre la necesidad de verificar el contenido artificial antes de confiar en él.

2. Desafíos de la búsqueda generativa

Los modelos generativos pueden crear contenido y así responder a consultas formuladas en lenguaje natural. Esta posibilidad implica un cambio radical en el acceso al conocimiento. Ello ha suscitado afirmaciones exageradas, incluso apocalípticas, sobre el futuro de los buscadores. A decir verdad, la IA se aplica desde hace tiempo en procesos internos de los buscadores para mejorar su rendimiento y eficacia. Uno de los primeros y más influyentes modelos de lenguaje, *BERT*, se diseñó precisamente para comprender mejor las búsquedas en *Google* (**Devlin** *et al.*, 2018).

En la actualidad, se prevé que la IA generativa gane protagonismo al sustituir, al menos en parte, los resultados basados en listas de referencias web o en extractos de las fuentes originales por respuestas directas en forma de texto artificial (**Codina**, 2023). Si se confirma esta tendencia, estaríamos ante un cambio de paradigma que podría transformar nuestra manera de buscar y obtener respuestas.

La búsqueda generativa es una tecnología emergente que busca mejorar la experiencia de búsqueda y facilitar el acceso a la información (**Lopezosa**, 2023a). Pero los grandes modelos lingüísticos también presentan desafíos y riesgos, como la veracidad, la calidad y la ética de las respuestas. Aplicados a la búsqueda, los modelos generativos pueden ofrecer resultados imprecisos, incompletos o indebidos, porque se basan en datos que pueden estar sesgados o anticuados, o que pueden incurrir en contradicciones.

La IA es válida entonces para situaciones que aceptan un margen de error, cierta superficialidad argumental, incluso algún disparate. Pero no lo es para cuestiones críticas como un trabajo científico, un consejo legal o financiero, o una consulta médica. Crea una engañosa ilusión de pensamiento racional porque imita capacidades humanas, pero no razona ni dispone de conocimiento fiable sobre el mundo (**Mitchell**; **Krakauer**, 2023). Es una "cotorra estocástica" que no entiende, en un sentido humano, nada de lo que dice (**Bender** *et al.*, 2021); produce lenguaje de apariencia humana, pero carente del significado que manejamos los humanos. Trata las palabras como objetos matemáticos.

Y puesto que los modelos generativos se han construido, sobre todo, para desarrollar conversaciones, no son necesariamente veraces. Las respuestas, aunque elocuentes e incluso persuasivas, a veces son incorrectas o absurdas, y se inventan hechos, personas, datos o fuentes. A menudo citan falsas fuentes de información o referencias fantasma (**Orduña-Malea**; **Cabezas-Clavijo**, 2023), que crean una falsa impresión de validez y fiabilidad.

A pesar de estos problemas, hay quienes aceptan todo lo que emana de estas herramientas. Y ello tiene explicación: hay propensión a atribuir a la IA unas habilidades cognitivas y de razonamiento que no tiene. Pero esta tendencia no es más que una ilusión de profundidad explicativa, un sesgo cognitivo que ha deformado la percepción pública de la IA hasta proyectar de ella una visión exagerada, antropomórfica y deificada.

Anthropic, la organización responsable del modelo *Claude*, acaba de publicar una investigación sobre cómo y por qué los asistentes de IA deciden dar las respuestas que dan. La cuestión es si los *chatbots* se basan en la "memorización" para generar resultados, o si existe una relación más profunda entre los datos de entrenamiento, el ajuste fino y lo que finalmente se emite. La pregunta todavía no tiene respuesta: los científicos ignoran por qué los modelos de IA dan las respuestas que generan (**Grosse** *et al.*, 2023). La cantidad de datos, patrones y pasos algorítmicos que manipula un modelo de lenguaje impide, hoy por hoy, contar con un método directo para rastrear el origen de un resultado.

3. La IA generativa, un intermediario problemático

Las numerosas limitaciones de los modelos generativos obligan a cuestionar su papel como fuente de información fidedigna. En realidad, la IA generativa no es una fuente en sí misma, sino un intermediario que procesa y transforma la información que está disponible en otros lugares. El proceso no es transparente ni tampoco infalible. Las respuestas pueden contener errores e imprecisiones, que se conocen en la jerga especializada como "alucinaciones".

Este término es problemático, ya que parece atribuir a la IA capacidades o comportamientos humanos que no tiene, como también hacen los términos "entrenamiento", "aprendizaje" o la mismísima expresión "inteligencia". Mientras que las alucinaciones humanas son fenómenos sensoriales causados por alteraciones químicas o cerebrales, las de la IA son fallos o inconsistencias en la generación de texto, que se deben a las limitaciones de los datos y de los algoritmos que utiliza. Emplear el mismo término para ambos casos genera una representación ilusoria de la IA y provoca confusión sobre su naturaleza y sus verdaderas capacidades.

Esta confusión también la produce el diseño de interacción de los *chatbots*. Al simular la conversación humana mediante rasgos humanos como las respuestas en primera persona y las expresiones de sentimientos, la interfaz puede inducir a algunos usuarios a creer con cierta ingenuidad que el sistema tiene emociones, personalidad, voluntad propia y otras cualidades humanas. Este fenómeno es el denominado "efecto Eliza", en referencia a la ilusión de estar hablando con una persona, un espejismo que experimentaron los usuarios de *Eliza*, el primer *chatbot* de la historia (**Tarnoff**, 2023).

En cualquier caso, las respuestas de la IA generativa no son información contrastada y fiable, sino meras construcciones sintéticas basadas en la probabilidad secuencial de las palabras, y no en la lógica. Estas construcciones dependen del corpus de texto con el que se ha entrenado el modelo, que es sólo una parte de lo accesible en Internet, y también son fruto del "intrigante" comportamiento de los algoritmos (**Grosse** *et al.*, 2023).

Descubrir cómo prevenir o corregir las respuestas erróneas se ha convertido en una obsesión para muchos investigadores. Es un problema que se menciona en docenas de artículos académicos (**De-Vynck**, 2023). A medida que la tecnología llega a más personas y se integra en campos críticos, como la medicina, el derecho o las finanzas, comprender las mencionadas "alucinaciones" y encontrar formas de mitigarlas es aún más crucial.

4. ¿Se pueden evitar los sesgos?

La IA generativa también plantea desafíos éticos relacionados con los sesgos que pueden afectar a los resultados. Los modelos lingüísticos son susceptibles de reproducir los prejuicios y los estereotipos dañinos que se encuentran en el corpus empleado para su entrenamiento. Para mitigar este problema se suele recurrir a la alineación de los sistemas, que consiste en ajustarlos y reentrenarlos para evitar que perpetúen sesgos perjudiciales.

No obstante, algunas voces sostienen que es imposible crear un modelo de lenguaje totalmente neutral y objetivo, dado que los sesgos se introducen en distintas fases del desarrollo del modelo, desde la selección de los datos hasta el entrenamiento y la evaluación. Además, los sesgos son problemas sociales complejos que no pueden abordarse solo con soluciones técnicas (**Heikkilä**, 2023).

Los sesgos en la IA generativa parecen hoy inevitables. Siendo así, hay que tener en cuenta que estos sesgos implican efectos negativos en dos sentidos opuestos:

- en un sentido, pueden producir resultados ofensivos, inexactos o injustos, como afirmaciones racistas o sexistas, o pueden ignorar o tergiversar las voces de las minorías;
- pero, en el otro sentido, los sesgos pueden hacer que la IA se vea poco auténtica, "políticamente correcta", incluso puritana, y así podrían limitarse la creatividad, la diversidad sociocultural y la libertad de expresión si la IA elude temas controvertidos o censura palabras consideradas urticantes o polémicas.

5. Nueva responsabilidad social

Los asistentes de IA pueden ofrecer respuestas rápidas y personalizadas a casi cualquier instrucción que se les dé. Por eso, saber formular instrucciones adecuadas —lo que se conoce como "ingeniería de peticiones" o *prompt engineering*— se ha convertido en una competencia muy valorada y parece evidente que los profesionales de la información deben dominarla, no sólo para explotarla en su práctica diaria, sino también para hacer buena pedagogía de ella.

Cabe insistir en que los modelos de lenguaje también presentan riesgos y desafíos, y que estos exigen una evaluación crítica. Las respuestas generadas por un asistente de IA son un material "en bruto" que puede contener errores y puede inducir a equívocos y a malas decisiones. Este peligro lo agrava, además, el hecho de que actualmente se produce y se difunde una cantidad enorme de contenido sintético, difícil de filtrar y carente de criterios que garanticen o prioricen la verdad.

Ante este panorama, el colectivo bibliotecario tiene una responsabilidad social y una oportunidad para formar a la población en el uso eficaz y responsable de los modelos generativos. Tiene la ocasión de ofrecer orientación tanto sobre el potencial de la IA como sobre sus riesgos. Los bibliotecarios, con amplia experiencia en la búsqueda, selección, evaluación y difusión de recursos informativos de calidad, pueden promover el desarrollo de competencias digitales e informacionales entre el público usuario. En este sentido, las bibliotecas asumen un papel aún más clave para la comunidad si apuestan por cultivar la alfabetización en IA entre el público en general (**Lo**, 2023).

Los profesionales de la información pueden ayudar a los usuarios a verificar las respuestas procedentes de un *chatbot*. Para ello, sería fácil enseñarles a aplicar criterios de calidad y fiabilidad, y a contrastar la información con otras fuentes, utilizando herramientas de búsqueda especializadas o bases de datos académicas. Asimismo, pueden ayudar a mitigar la ilusión de pensamiento racional que provocan los modelos lingüísticos, que tienden a simular cualidades humanas que no poseen e inducen a los usuarios a confiar demasiado en sus respuestas. Para ello, los bibliotecarios pueden explicar a los usuarios cómo funciona esta tecnología y qué limitaciones tiene, y fomentarán así una actitud crítica y reflexiva sobre el uso de estos sistemas.

El personal bibliotecario también tiene capacidad para crear conciencia sobre los sesgos presentes en los modelos de lenguaje. Para ello, será útil mostrar a los usuarios ejemplos y evidencias de la existencia y las consecuencias de estos sesgos. También pueden enseñarles a detectarlos, identificarlos como tales, y mitigarlos, en consecuencia, utilizando para ello técnicas como el cambio de perspectiva, la consulta de fuentes diversas o el análisis crítico del discurso.

Las bibliotecas, por su parte, pueden contribuir de varias maneras a la alfabetización en IA. Por ejemplo, organizando actividades formativas y divulgativas sobre esta tecnología, tales como talleres, cursos, charlas o exposiciones, que involucren a expertos, docentes, estudiantes y ciudadanos. También creando espacios y comunidades de aprendizaje colaborativo sobre IA, donde los usuarios puedan aprender, ensayar y crear con esta tecnología y compartir conocimientos, experiencias y proyectos. Al fin y al cabo, las bibliotecas fomentan el intercambio de conocimiento, habilidades e ideas. Uno de los entornos que lo hacen posible son los espacios de creación (*makerspaces*), talleres tecnológicos donde las personas materializan ideas y proyectos mediante herramientas informáticas.

A través de espacios propios, las bibliotecas podrían diseñar y ofrecer actividades que acerquen la IA generativa al público general para democratizar el acceso y el uso, y favorecer la alfabetización en esta tecnología. Ilustran esta posibilidad iniciativas como el proyecto *ExperimentAI*, impulsado por la *Universitat Autònoma de Barcelona* en la *Red de Bibliotecas Públicas de Barcelona* (figura 1), o la línea de experimentación *Espacios Read Maker*, de la Diputación de Badajoz. Además, las bibliotecas pueden facilitar el encuentro y la interacción entre usuarios interesados en la IA, creando redes de aprendizaje y colaboración que potencien el desarrollo personal y comunitario.

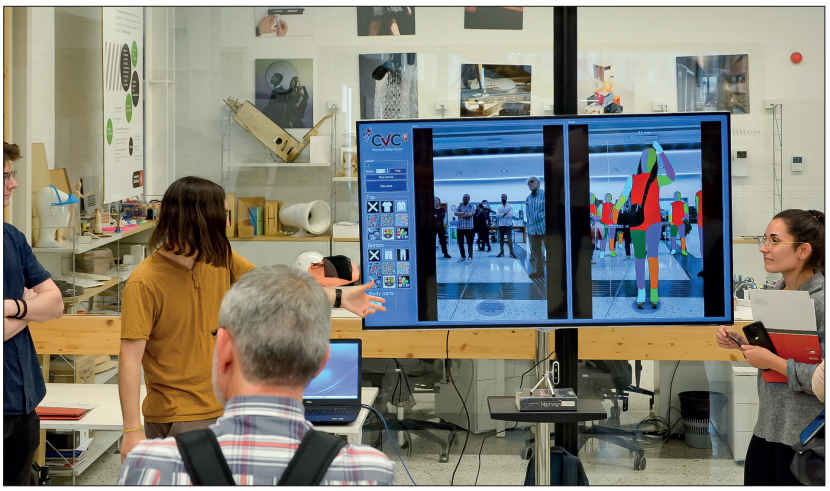

Figura 1. Sesión del ciclo *ExperimentAI* en la *Biblioteca Montserrat Abelló* (Barcelona)

Un ejemplo de actividad sería un taller de formulación de instrucciones, donde los usuarios puedan aprender a elaborar *prompts* precisos y eficaces, adecuados a propósitos específicos. Otro ejemplo sería un taller de arte generativo, donde los usuarios tengan la oportunidad de crear obras artísticas con modelos que generen textos, imágenes, música o vídeos. Y dado que está demostrado que la IA generativa puede dar un impulso significativo a la forma de hacer ciencia (**Lopezosa**, 2023b), desde las bibliotecas universitarias y especializadas se puede ofrecer asesoramiento a investigadores para mejorar el diseño de sus investigaciones, el método de recogida de datos y la escritura de artículos científicos.

Al promover el uso de los modelos generativos como una herramienta para el aprendizaje, la investigación y la creación, las bibliotecas asumirían un rol activo y transformador en esta nueva era del conocimiento. Los usuarios podrían aprovechar así las ventajas que ofrece esta tecnología, conscientes de sus riesgos y limitaciones. De esta manera, las bibliotecas tendrán la magnífica oportunidad de fomentar el aprendizaje, la investigación y la creación con la IA, y de contribuir a una sociedad más crítica e informada.

6. Referencias

Bender, **Emily M.**; **Gebru**, **Timnit**; **McMillan-Major**, **Angelina**; **Shmitchell**, **Shmargaret** (2021). "On the dangers of stochastic parrots: can language models be too big?". *FAccT '21: Proceedings of the 2021 ACM Conference on fairness, accountability and transparency*, p. 610–623.
https://doi.org/10.1145/3442188.3445922

Codina, **Lluís** (2023). "Buscadores alternativos a *Google* con IA generativa: análisis de *You.com*, *Perplexity AI* y *Bing Chat*". *Infonomy*, v. 1, e23002.
https://doi.org/10.3145/infonomy.23.002

De-Vynck, **Gerrit** (2023). "*ChatGPT* 'hallucinates': some researchers worry it isn't fixable". *The Washington Post*, May 30.
https://wapo.st/3KN7dRn

Devlin, **Jacob**; **Chang**, **Ming-Wei**; **Lee**, **Kenton**; **Toutanova**, **Kristina** (2018). "BERT: pre-training of deep bidirectional transformers for language understanding". *arXiv*.
https://doi.org/10.48550/arXiv.1810.04805

Griffith, **Erin**; **Metz**, **Cade** (2023). "'Let 1,000 flowers bloom': AI funding frenzy escalates". *New York Times*.
https://nytimes.com/2023/03/14/technology/ai-funding-boom.html

Grosse, **Roger**; **Bae**, **Juhan**; **Anil**, **Cem**; **Elhage**, **Nelson**; **Tamkin**, **Alex**; **Tajdini**, **Amirhossein**; **Steiner**, **Benoit**; **Li**, **Dustin**; **Durmus**, **Esin**; **Perez**, **Ethan**; **Hubinger**, **Evan**; **Lukošiūtė**, **Kamilė**; **Nguyen**, **Karina**; **Joseph**, **Nicholas**; **McCandlish**, **Sam**; **Kaplan**, **Jared**; **Bowman**, **Samuel R.** (2023). "Studying large language model generalization with influence functions". *arXiv*.
https://doi.org/10.48550/arXiv.2308.03296

Heikkilä, **Melissa** (2023). "Why it's impossible to build an unbiased AI language model". *MIT technology review*.
https://technologyreview.com/2023/08/08/1077403

Lo, Leo S. (2023). "AI policies across the globe: Implications and recommendations for libraries". *IFLA Journal*.
https://doi.org/10.1177/03400352231196172

Lopezosa, **Carlos** (2023a). "*Bing* chat: hacia una nueva forma de entender las búsquedas". *Anuario ThinkEPI*, v. 17, e17a04.
https://doi.org/10.3145/thinkepi.2023.e17a04

Lopezosa, **Carlos** (2023b). "*ChatGPT* y comunicación científica: hacia un uso de la inteligencia artificial que sea tan útil como responsable". *Hipertext.net*, n. 26, pp. 17–21.
https://doi.org/10.31009/hipertext.net.2023.i26.03

Mitchell, **Melanie**; **Krakauer**, **David C.** (2023). "The debate over understanding in AI's large language models". *PNAS*, v. 120, n. 13, e2215907120.
https://doi.org/10.1073/pnas.2215907120

Orduña-Malea, **Enrique**; **Cabezas-Clavijo**, **Álvaro** (2023). "*ChatGPT* and the potential growing of ghost bibliographic references". *Scientometrics*, n. 128, p. 5351–5355.
https://doi.org/10.1007/s11192-023-04804-4

Tarnoff, **Ben** (2023). "Weizenbaum's nightmares: how the inventor of the first chatbot turned against AI". *The Guardian*, 25 julio.
https://theguardian.com/technology/2023/jul/25/joseph-weizenbaum-inventor-eliza-chatbot-turned-against-artificial-intelligence-ai