

El impacto energético de la inteligencia artificial

El lanzamiento del chatbot ChatGPT a finales de 2022 fue un hito en la disponibilidad y utilidad de la inteligencia artificial (IA), y muchas otras empresas comenzaron a ofrecer sus propios servicios de IA [1, 2]. Aunque el usuario suele familiarizarse con el nombre de la capa de aplicación (por ejemplo, ChatGPT o Copilot), lo que realiza el trabajo de construir las respuestas es un modelo de lenguaje de gran tamaño (LLM, por las siglas en inglés). Actualmente existen decenas de LLM diferentes (sin contar las diferentes versiones del "mismo" LLM), y entre los más conocidos están Claude (Anthropic), DeepSeek-R1 (DeepSeek), Gemini (Google DeepMind), GPT (OpenAI), Granite (IBM), Grok (xAI), Llama (Meta), o4 (OpenAI), Qwen (Alibaba Cloud) y Stable LM (Stability AI) [3]. También se han desarrollado decenas de modelos que se encargan de generar imágenes o videos de unos pocos segundos a partir de entradas de texto.



A pesar del elevado impacto social de estas herramientas de IA, durante 2025 se ha generado cierta desilusión sobre ellas. Entre los motivos se encuentran los siguientes tres: (1) ChatGPT generó unas expectativas infladas sobre la IA; (2) los líderes de las mayores compañías de IA hicieron promesas sobre las capacidades de sus modelos que no podían cumplir; y (3) en general las empresas que han invertido en herramientas de IA no han logrado recuperar sus inversiones [1, 2]. Adicionalmente, aún no es claro cómo las grandes compañías desarrolladoras de IA recuperarán las increíbles cantidades de dinero que han gastado en construcción de centros de datos y en entrenamiento de modelos [1]. Las actualizaciones tecnológicas centrales (es decir, en los modelos) no son los grandes saltos de hace unos años, porque con el hardware y la estrategia preferida de modelado actuales se ha alcanzado el límite de parámetros que pueden incluirse en los LLM de manera que su entrenamiento tome un tiempo aceptable (esto es, unos pocos meses) [1, 4]. Y los LLM actuales son mejores que el humano promedio en muchas cosas diferentes, pero no son mejores que un humano experto, además de que tienden a ser mucho peores para generalizar que las personas [1].

Lo anterior no quiere decir que se haya alcanzado el límite superior operativo de la IA. El problema está en que los lanzamientos de años recientes dieron la impresión de que los avances se presentarían de manera espectacularmente veloz [1]. Pero en realidad, el desarrollo de las capacidades actuales de IA ha sido un trabajo gradual, con una teoría básica que comenzó a desarrollarse hace más de 60 años y un crecimiento progresivo de la producción de conocimiento sobre IA; para la muestra, el número de artículos científicos relacionados con IA se duplicó de 2010 a 2021 [2]. De modo que se espera que el uso de IA sea cada vez mayor, con un futuro en el que se utilicen ampliamente agentes de IA y otras aplicaciones de mayor complejidad que las actuales [5]. Y este inevitable uso creciente de la IA tendrá implicaciones sobre los sistemas energéticos y los recursos naturales, porque la IA requiere de recursos: los materiales que se usan en la construcción de los centros de datos y el hardware, agua para enfriamiento, y mucha energía. Vale la pena mencionar que no es una novedad que los desarrollos tecnológicos que permiten realizar procesos a mayor velocidad o de manera automatizada, o sustituir actividades humanas, requieran recursos para su fabricación y, sobre todo, energía para su operación.

Los modelos de IA consumen energía en sus dos etapas de existencia: durante su entrenamiento, y durante su uso (o "inferencia"). Si es un modelo que no se modifica automáticamente, la energía del entrenamiento es un gasto único, pero significativo. Por ejemplo, se estima que el entrenamiento de GPT-4 duró más de tres meses, costó más de 100 millones de dólares y consumió 50 GWh de energía [6], que es aproximadamente la cantidad de electricidad que consume la ciudad de Bucaramanga (más de 600 000 habitantes) durante 24 días [7]. Y durante la inferencia o uso, la energía que consuma dependerá de las veces que se utilice el modelo (que "se consulte"); se estima que actualmente más del 80 % del poder de cómputo para IA se usa para inferencia [5]. El consumo energético de una consulta depende de muchas variables, incluyendo el tamaño y tipo del modelo y el tipo de salida que se está generando. De manera asociada, las emisiones de gases de efecto invernadero de una consulta dependen, además del consumo energético, de la ubicación del centro de datos y la red de electricidad a la que está conectado, y de la hora del día. Según Google, una consulta promedio de texto, que no genere imágenes o videos, a Gemini consume 0.24 Wh (864 J) de electricidad, que es aproximadamente el consumo de un microondas en 1 s [8]. Y de acuerdo con la información presentada en ML.Energy Leaderboard por investigadores de la Universidad de Michigan, que calculan el consumo energético por parte de modelos abiertos, la creación de una imagen por parte de Stable Diffusion consume en promedio 3.56 kJ, y la creación de un video de 5 s a partir de la descripción de una escena por parte de Wan 2.1 requiere en promedio alrededor de 340 kJ [9]. Usando la analogía del microondas, la creación de la imagen equivaldría a usar el microondas unos 4 s, y generar el video de 5 s gastaría energía como el uso del microondas durante unos seis minutos y medio.

El crecimiento de nuevas tecnologías, incluyendo la IA, ha contribuido a nuevos récords en la demanda mundial de energía, y a pesar de que el despliegue de renovables haya sido creciente por vigesimotercer año consecutivo, también el consumo de energías fósiles y nuclear han tocado máximos históricos [10]. El consumo de electricidad para IA ocurre en los centros de datos que se destinan parcialmente para su entrenamiento y uso. Entre 2005 y 2017, el consumo eléctrico de los centros de datos se mantuvo más o menos constante, porque aunque se

construyeron muchos centros de datos, las mejoras en eficiencia impidieron que el consumo energético se incrementara [5]. Pero después se comenzaron a construir centros de datos con hardware para IA, que requieren mucha energía, y por esto para 2023 el consumo energético de los centros de datos en Estados Unidos se había duplicado; hoy en día los más de 3 000 centros de datos estadounidenses consumen el 4.4 % de la electricidad de ese país [5, 11, 12]. En Estados Unidos se teme que los tratos entre las empresas de electricidad y los centros de datos terminen ocasionando un incremento en el precio de la energía para los demás usuarios [12]. Adicionalmente, muchos centros de datos han recurrido a combustibles fósiles para satisfacer sus demandas inmediatas de energía [5, 12]. Todo esto contribuye a que se esté evidenciando una disminución en el impulso de los esfuerzos internacionales para reducir las emisiones de gases de efecto invernadero, a pesar de que 2024 fue el primer año en que la temperatura global promedio excedió 1.5 °C por encima de los niveles pre-industriales [10].

Referencias

- [1] Heaven, W. D. (2025). The great AI hype correction of 2025. Recurso web, disponible en <https://www.technologyreview.com/2025/12/15/1129174/the-great-ai-hype-correction-of-2025/>
- [2] Stanford Online. (s. f.). Getting beyond the hype: A guide to AI's potential. Recurso web, disponible en <https://online.stanford.edu/getting-beyond-hype-guide-ais-potential>
- [3] Caballar, R. D. y Stryker, C. (2025). Una lista de modelos de lenguaje de gran tamaño. Recurso web, disponible en <https://www.ibm.com/es-es/think/topics/large-language-models-list>
- [4] Patel, D. (2023). The AI brick wall – A practical limit for scaling dense transformer models, and how GPT 4 will break past it. Recurso web, disponible en <https://newsletter.semianalysis.com/p/the-ai-brick-wall-a-practical-limit>
- [5] O'Donnell, J. y Crownhart, C. (2025). We did the math on AI's energy footprint. Here's the story you haven't heard. Recurso web, disponible en <https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/>
- [6] Patel, D. y Wong, G. (2023). GPT-4 architecture, infrastructure, training dataset, costs, vision, MoE. Recurso web, disponible en <https://newsletter.semianalysis.com/p/the-ai-brick-wall-a-practical-limit>
- [7] Sistema Único de Información de Servicios Públicos Domiciliarios. (2025). Energía / Comercia l /Consolidado energía por empresa departamento y municipio. Recurso web, disponible en http://reportes.sui.gov.co/fabricaReportes/frameSet.jsp?idreporte=ele_com_096
- [8] Crownhart, C. (2025). In a first, Google has released data on how much energy an AI prompt uses. Recurso web, disponible en <https://www.technologyreview.com/2025/08/21/1122288/google-gemini-ai-energy/>
- [9] Chung, J.-W., Ma, J. J., Wu, R., Liu, J., Kweon, O. J., Xia, Y., Wu, Z., Chowdhury, M. (2025). The ML.ENERGY Leaderboard - How much time and energy do generative AI models consume? Recurso web, disponible en <https://ml.energy/leaderboard/>
- [10] International Energy Agency. (2025). World Energy Outlook 2025.
- [11] Ontiveros, J. E., Patel, D., Zhou, W., Kourabi, A. J. y Barkin, M. (2025). xAI's Colossus 2 - First gigawatt datacenter in the world, unique RL methodology, capital raise. Recurso web, disponible en <https://newsletter.semianalysis.com/p/xais-colossus-2-first-gigawatt-datacenter>
- [12] Samaras, C., Strubell, E. y Krishnan, R. (2025). AI's energy impact is still small—but how we handle it is huge. Recurso web, disponible en <https://www.technologyreview.com/2025/05/20/1116274/opinion-ai-energy-use-data-centers-electricity/>

CARLOS E GARCÍAS

CARLOS EDUARDO GARCIA SANCHEZ

Editor General

Revista Met&Flu