Atencion Primaria 57 (2025) 103223

Atención Primaria

www.elsevier.es/ap

EDITORIAL

La inteligencia artificial en atención primaria: ¿solución o problema?

Artificial intelligence in primary care: Friend or foe?

La inteligencia artificial, aunque parezca reciente, no es un concepto nuevo, especialmente en medicina[1]. En 1964 el software ELIZA[2], diseñado como un psicoterapeuta, mostró que con reglas simples era posible hacer creer a muchas personas que hablaban con un ser humano cuando en realidad interactuaban con un algoritmo. En los años 70 un médico internista creó el sistema MYCIN[3], capaz de seleccionar pautas antibióticas para septicemia con mayor acierto que muchos clínicos. MYCIN nunca llegó a implementarse de forma generalizada, evidenciando que el mayor desafío de la inteligencia artificial no es su desarrollo, sino su integración en la práctica clínica real.

Durante décadas el procesamiento del lenguaje natural aplicado a la medicina fue un área de interés principalmente académico[4]. Un avance crucial ocurrió en 2017, cuando investigadores de Google presentaron el *transformer*, una red neuronal diseñada para la traducción automática con una precisión sin precedentes[5]. A diferencia de los modelos tradicionales de aprendizaje automático, que requerían datos específicos para cada tarea, los grandes modelos de lenguaje (LLM por sus siglas en inglés) se entrenan mediante arquitecturas neuronales *transformer* a partir de textos ya existentes en Internet. Este entrenamiento se basa en una tarea aparentemente sencilla pero que permite, dado un volumen suficiente de textos, aprender las reglas del lenguaje humano, así como las relaciones entre conceptos presentes en esos textos: predecir la siguiente palabra dada una secuencia de palabras previas.

En 2021 investigadores de Google descubrieron de manera inesperada que estos LLM podían, además de traducir idiomas, resolver cualquier tarea que se les planteara sin necesidad de entrenamiento previo específico. Simplemente había que ofrecer al LLM una descripción en lenguaje natural sobre la tarea, algo que se denominó *zero-shot learning*[6]. Esta capacidad permite, por ejemplo, responder preguntas clínicas o resumir informes médicos sin que sea necesario un entrenamiento específico.

Sin embargo, los LLM ganaron verdadera popularidad en octubre de 2022, cuando OpenAI lanzó ChatGPT, haciendo accesible esta tecnología de manera gratuita. El alto coste de entrenar LLM, estimado en unos 40 millones de dólares para GPT-4[7], limita la competencia y favorece la consolidación de oligopolios. Se explica así el gran interés empresarial y, como resultado, mediático desde ese momento.

Otra consecuencia es la proliferación de «expertos» sobrevenidos, muchos de los cuales poco antes promovían tendencias como el metaverso y ahora se presentan como referentes de la «IA generativa». Este fenómeno recuerda al doctor Albert Abrams, quien hace un siglo, y aprovechando el auge de la entonces novedosa electricidad, atribuía propiedades curativas a sus máquinas de electroterapia, incluido el tratamiento del cáncer de estómago[8]. Antes de nada, es esencial cuestionar las afirmaciones de estos nuevos gurús y examinar críticamente su entusiasmo.

No hay duda de que los LLM representan un avance notable, reavivando debates sobre la posibilidad de que las máquinas razonen como humanos. Algunos expertos afirman que pueden realizar razonamientos clínicos complejos[9], mientras que otros los ven como simples «loros estocásticos»[10] sin capacidad de comprensión.

Si un sistema actúa «como si» razonara, podríamos considerarlo inteligente, dejando su naturaleza interna como una cuestión filosófica. Desgraciadamente, determinar si algo actúa inteligentemente exige definir qué significa inteligencia, lo cual no es sencillo. La historia muestra que este concepto evoluciona constantemente. Avances tecnológicos han cambiado nuestra perspectiva: las máquinas superaron a los humanos en tareas antes consideradas signos de inteligencia, como memorizar datos o jugar ajedrez, redirigiendo nuestro enfoque a la creatividad o el lenguaje. Con los LLM y su capacidad para procesar texto volvemos a cuestionar qué entendemos por inteligencia.

https://doi.org/10.1016/j.aprim.2025.103223

La paradoja de Moravec[11] ilustra un punto interesante: tareas motoras básicas, como limpiar un baño o alimentar a un bebé, aunque más antiguas desde una perspectiva evolutiva que habilidades como el lenguaje o la resolución de ecuaciones, son mucho más difíciles de replicar por máquinas, pese a no considerarse tradicionalmente «inteligencia».

En el ámbito clínico, los LLM nos hacen replantearnos la naturaleza del acto médico. Así como Internet redujo la asimetría de información entre médicos y pacientes, estas herramientas podrían transformar el rol del médico al ser capaces de sustituirnos en tareas cognitivas complejas, como explicar conceptos médicos de forma adaptada al nivel educativo del paciente[12].

Debemos ser cautos al evaluar el impacto de los LLM en medicina y en atención primaria. Aunque algunas investigaciones sugieren que superan a los médicos en ciertas tareas, a menudo utilizan métodos alejados de la realidad de la medicina de familia. Estos estudios se enfocan en diagnósticos diferenciales de casos infrecuentes[13] o en resolver preguntas tipo test[9], pero omiten aspectos esenciales, como el manejo de la incertidumbre o la interpretación de la narrativa del paciente[14]. Además, suelen utilizar a los médicos subespecialistas como estándar de referencia[15], reforzando estereotipos como que las enfermedades «fáciles» son tratables por cualquier médico, mientras que las «difíciles» requieren de subespecialistas. Otro tema controvertido es el hecho de que los pacientes valoren las respuestas de los LLM como más empáticas que las de los médicos[12]. Se plantean entonces dilemas éticos, como la reacción de un paciente al descubrir que un mensaje empático de su médico fue realmente redactado por una máquina.

Por último, los LLM enfrentan riesgos significativos, como sus «alucinaciones» o confabulaciones, respuestas incorrectas que parecen ciertas. Identificar estas confabulaciones requiere conocimientos especializados, lo que limitaría su utilidad para apoyar a médicos de familia, por ejemplo, en enfermedades raras. Irónicamente, los subespecialistas, mejor capacitados para detectar estos errores, no suelen ser considerados como usuarios iniciales de estos LLM, ya que se les percibe como el «estándar de referencia» y, por tanto, no necesitados de ayuda.

El auge de los LLM exige que los médicos de familia asuman un papel activo y crítico en su evaluación e implementación. Estas herramientas, aún en desarrollo, están influidas por dinámicas de poder preexistentes, lo que hace urgente defender las necesidades y prioridades de la atención primaria. El primer paso es entender su funcionamiento y aprender a usar los LLM. Un reciente estudio[16] no pudo encontrar mejoras en razonamiento clínico, ni en tiempos de resolución al comparar médicos que usaron buscadores tradicionales frente a otros que también usaron LLM, y los autores sugirieron que podría deberse a que los médicos aún desconocen cómo aprovechar su potencial.

Los médicos de familia deben además liderar investigaciones desde la atención primaria, detectando oportunidades y desafíos que suelen pasar desapercibidos en contextos hospitalarios. También es fundamental liderar el diálogo público para garantizar que estas tecnologías respondan a las necesidades reales de pacientes y profesionales sanitarios, y no solo a los intereses de grandes corporaciones tecnológicas.

La integración de los LLM debe hacerse con inteligencia, respetando la complejidad y particularidades de la atención primaria, reconociendo así las fortalezas y desafíos diferenciales de cada ámbito asistencial. Este es un momento clave, y el tiempo para actuar es ahora.

Bibliografía

1. Bonis J, Sancho JJ, Sanz F. Sistemas informáticos de soporte a la decisión clínica. Med Clin (Barc). 2004;122 Suppl 1:39–44.
2. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM. 1966;9:36–45, http://dx.doi.org/10.1145/365153.365168.
3. Shortliffe EH. Computer-based medical consultations: MYCI.N. New York: Elsevier; 1976. p. 286, http://dx.doi.org/10.1016/B978-0-444-00179-5.X5001-X.
4. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: A stimulus to scientific discovery. Artif Intell. 1997;91:183–203, http://dx.doi.org/10.1016/S0004-3702(97)00008-8.
5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, JonesL, Gomez AD, et al. Attention is all you need. ArXiv [Preprint]. 15 p. Disponible en: https://arxiv.org/abs/1706.03762 https://doi.org/10.48550/arXiv.1706.03762
6. Wei J, Bosma M, Zhao VY, Guu K, Wei Yu A, Lester B, et al. Finetuned language models are zero-shot learners. ArXiv [Preprint]. 2021:46, http://dx.doi.org/10.48550/arXiv.2109.01652. Disponible en https://arxiv.org/abs/2109.01652.
7. Cottier B, Rahman R, Fattorini L, Maslej N, Owen D. The rising costs of training frontier AI models. ArXiv [Preprint]. 2024:20, http://dx.doi.org/10.48550/arXiv.2405.21015. Disponible en https://arxiv.org/abs/2405.21015.
8. Abrams A. What the American Medical Association thinks of the electronic reactions of Abrams. Dent Regist. 1923;77:117–24. Disponible en https://pmc.ncbi.nlm.nih.gov/articles/PMC7872713/
9. Singhal K, Azizi S, Tu T, Mahdavi S, Wei J, Won H, et al. Large language models encode clinical knowledge. Nature. 2023;620:172–80, http://dx.doi.org/10.1038/s41586-023-06291-2.
10. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? En: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). New York, USA: Association for Computing Machinery. pp. 610-623. Disponible en: https://doi.org/10.1145/3442188.3445922.
11. Moravec HP. Mind children: The future of robot and human intelligence. Cambridge, Mass: Harvard University Press; 1988.
12. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183:589–96, http://dx.doi.org/10.1001/jamainternmed.2023.1838.
13. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. ArXiv [Preprint]. 2023:17, http://dx.doi.org/10.48550/arXiv.2312.00164. Disponible en: https://arxiv.org/abs/2312.00164
14. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30:2613–22, http://dx.doi.org/10.1038/s41591-024-03097-.

15. Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic AI. ArXiv [Preprint]. 2024:46, http://dx.doi.org/10.48550/arXiv.2401.05654. Disponible en: https://arxiv.org/abs/2401.05654
16. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: A randomized clinical trial. JAMA Netw Open. 2024;7, http://dx.doi.org/10.1001/jamanetworkopen.2024.40969, e2440969.

Julio Bonis Sanz[a,*] y Rafael Bravo Toledo[b,c]
[a] *Atención primaria, Investigador Independiente, Madrid, España*
[b] *Centro de Salud Linneo, SERMAS, Madrid, España*
[c] *GdT semFYC de Innovación Tecnológica y Sistemas de Información, Madrid, España*

[*] Autor para correspondencia.
Correo electrónico: drbonis@gmail.com (J. Bonis Sanz).