Revista Española de Educación Médica

¿Es capaz GPT-4 de aprobar el MIR 2023? Comparativa entre GPT-4 y ChatGPT-3 en los exámenes MIR 2022 y 2023

Is GPT-4 capable of passing MIR 2023? Comparison between GPT-4 and ChatGPT-3 in the MIR 2022 and 2023 exams

Álvaro Cerame*[1], Juan Juaneda*[2], Pablo Estrella-Porter[3], Lucía de la Puente[4], Joaquín Navarro[5], Eva García[6], Domingo A. Sánchez[7,8], Juan Pablo Carrasco[9]**

1 Plan de Atención Integral al Profesional Sanitario Enfermo, Servicio Madrileño de Salud, Madrid, España. ORCID: 0000-0001-9137-7775
2 Servicio de Medicina Preventiva y Salud Pública, Hospital Universitari i Politècnic La Fe, Valencia, España. ORCID: 0000-0002-6048-2457.
3 Servicio de Medicina Preventiva, Hospital Clínico Universitario de Valencia, Valencia, España. ORCID: 0000-0003-4137-7691
4 Departamento de Atención Primaria, Hospital Universitari i Politècnic La Fe, Valencia, España. ORCID: 0009-0007-3263-5691
5 Servicio de Cuidados Intensivos, Área de Gestión Sanitaria Norte de Huelva, Huelva, España. ORCID: 0000-0002-7983-7289.
6 Servicio de Cardiología, Complejo Hospitalario Universitario Toledo, Toledo, España. ORCID: 0000-0001-8962-6023
7 Consejo General Colegios Oficiales de Médicos de España, Madrid, España.
8 Servicio de Oncología Médica Hospital Universitario Morales Meseguer, Grupo de Oncología Clínica y Translacional IMIB-Arrixaca, Murcia, Región de Murcia, España. dsanchez@cgcom.es. ORCID: 0000-0003-2073-0679.
9 Servicio de Psiquiatría, Hospital Provincial de Castellón, Castellón, España. ORCID: 0000-0001-9137-7775
*Coautores principales del estudio.
**Autor de correspondencia.

Recibido: 8/2/24; Aceptado: 20/2/24; Publicado: 26/2/24

Resumen:

Introducción: La inteligencia artificial (IA) está generando nuevas controversias, oportunidades y riesgos en la educación médica. Este estudio evalúa la capacidad de las versiones de inteligencia artificial (IA) ChatGPT-3 y GPT-4 para responder a las preguntas del examen de acceso a la formación médica especializada MIR en España, comparando el rendimiento entre las convocatorias de 2022 y 2023.
Metodología: Se realizó un estudio descriptivo transversal, utilizando GPT-4 para responder a las 210 preguntas del examen MIR 2023, comparando los resultados con los de ChatGPT-3 en el examen MIR 2022. Se utilizó análisis estadístico para determinar el porcentaje de acierto en función de la especialidad, tipo de pregunta y contenido de la misma.
Resultados: GPT-4 consiguió 173 aciertos de un total de 210 preguntas, rendimiento superior al de ChatGPT-3, que obtuvo 108 aciertos en el examen de la convocatoria anterior. Se observó una mejora notable en especialidades como Reumatología, Pediatría, Geriatría y Oncología, aunque algunos campos como Neumología y Oftalmología mostraron menos progreso o incluso resultados inferiores.
Conclusión: GPT-4 demostró un mejor rendimiento en comparación con ChatGPT-3, indicando avances en el procesamiento y análisis de datos por parte de la IA, así como en su comprensión contextual y aplicación de conocimientos médicos. Sin embargo, se enfatiza la importancia de reconocer las limitaciones de la IA y la necesidad de un enfoque crítico en su uso en educación médica.

Palabras clave: Inteligencia Artificial, ChatGPT-3, GPT4, educación médica, MIR, médico residente

Abstract:

Introduction: Artificial intelligence (AI) is generating new controversies, opportunities and challenges in medical education. This study evaluates the ability of artificial intelligence (AI) versions ChatGPT-3 and GPT-4 to answer MIR exam questions of the entrance exam in the specialized training in Spain, comparing performance between the 2022 and 2023 exams.
Methodology: A descriptive cross-sectional study was conducted, using GPT-4 to answer the 210 questions of the MIR 2023 exam, comparing the results with those of ChatGPT-3 in the MIR 2022 exam. Statistical analysis was used to determine the percentage of correct answers according to speciality, type of question, and question content.
Results: GPT-4 achieved 173 correct answers out of 210 questions, a higher performance than ChatGPT-3, which obtained 108 correct answers in the previous exam. A marked improvement was observed in specialties such as Rheumatology, Paediatrics, Geriatrics and Oncology, although some fields such as Pneumology and Ophthalmology showed less progress or even lower results.
Conclusion: GPT-4 demonstrated better performance compared to ChatGPT-3, indicating advances in AI data processing and analysis, as well as in its contextual understanding and application of medical knowledge. However, the article emphasizes the importance of recognising the limitations of AI and the need for a critical approach in medical education.

Keywords: Artificial Intelligence, ChatGPT-3, GPT4, medical education, MIR, resident physician

---

1. Introducción

La rápida evolución de la inteligencia artificial (IA) en el siglo XXI ha supuesto una notable innovación en numerosos campos del conocimiento y de la práctica profesional, entre ellos, la educación médica (1). Los modelos lingüísticos multimodales de gran tamaño (multimodal large language model o LLMs en inglés), caracterizados por su capacidad de aprender, adaptarse y realizar tareas complejas, están transformando el panorama de la enseñanza y el aprendizaje en medicina. En este contexto, herramientas de procesamiento de lenguaje natural como ChatGPT (2), están en el centro del debate y de la innovación en el ámbito del aprendizaje y enseñanza en ciencias de la salud. Este modelo de IA tiene la capacidad de interactuar en lenguaje humano, proporcionar explicaciones detalladas, y potencialmente resolver preguntas que requieren de la integración de varios niveles de análisis, especialmente en el ámbito de las evaluaciones y exámenes del conocimiento médico (3).

Así pues, estas herramientas se han comenzado a utilizar para generar contenido educativo de gran interés. A pesar de su breve recorrido, ya existe literatura sobre su uso como complementos y asistentes para la docencia, el aprendizaje personalizado, el acceso rápido a diversas fuentes de información, la generación de casos clínicos y preguntas de examen, al tiempo que pueden realizar traducciones casi inmediatas a diferentes idiomas (4).

Al mismo tiempo, se han descrito múltiples riesgos asociados al uso de IA en educación médica. Por un lado, nos ofrecen información cuya veracidad, ética y profesionalidad no está comprobada ni revisada por un profesional, lo que podría suponer un riesgo para estudiantes y pacientes (5). En este sentido llama la atención que se han observado diversos ejemplos de errores fácticos, contenido inventado (conocido como alucinación en el campo de la IA), así como sesgos de género, raciales y políticos (6-7). Por otro lado, algunos autores alertan de que podría fomentar dinámicas en la que los

discentes tengan menos incentivos para elaborar e integrar procesos reflexivos propios; provocando una dependencia del uso de estas herramientas y en consecuencia mermando sus capacidades de aprendizaje (8). No obstante, un número considerable de publicaciones muestra una actitud optimista respecto al uso de esta tecnología como herramienta a implementar en los procesos de aprendizaje (9).

Una de las aplicaciones de herramientas como ChatGPT que ha despertado interés en la comunidad científica y educativa es su aplicación a la hora de responder a pruebas y exámenes de evaluación de profesionales médicos. Uno de estos ejemplos sería el examen de acceso a la Formación Sanitaria Especializada de España, conocido como examen MIR. En uno de los primeros estudios publicados en este ámbito (10), se determinó que ChatGPT-3 era capaz de aprobar el examen MIR (11), con aproximadamente un 51% de respuestas acertadas. Posteriormente, con GPT-4 se ha observado mayores porcentajes de acierto en exámenes realizados en España (12-13), alcanzando entre un 80-90% de respuestas correctas.

A pesar de esto, la cantidad de estudios e información publicada es limitada, y no se ha publicado hasta el momento ningún estudio con los resultados del examen MIR 2023, realizado en enero de 2024.

La bibliografía internacional publicada en otros países como Estados Unidos (14), Japón (15), China (16), Alemania (17) o Italia (18), también sugiere que la versión de GPT-4 ofrece tasas de acierto más elevadas, estando la mayoría de estudios entre un 70 y un 90% de aciertos con la versión GPT-4 y entre el 50 y el 70% con la versión ChatGPT-3. Además de esto se han hecho estudios que tratan de entender la competencia de esta herramienta en áreas concretas o en formatos de pregunta. Se han realizado estudios en especialidades como la traumatología (19), radiología (20) o anatomía (21), obteniendo porcentajes muy elevados de acierto. No obstante, al igual que en la situación descrita en la bibliografía española, los estudios han comenzado a aparecer apenas hace 1 año y del análisis de la literatura existente se concluye la necesidad de aumentar nuestro conocimiento en este ámbito.

Por todo lo anteriormente descrito, el objetivo principal del presente estudio es analizar la capacidad de GPT-4 de responder de manera correcta las preguntas de la convocatoria del examen MIR 2023, realizado en enero del año 2024. Como objetivos secundarios, se pretende en primer lugar realizar un análisis específico de la capacidad de respuesta de la herramienta en función de la especialidad, contenido y tipología de las distintas preguntas y, en segundo lugar, realizar una comparativa de la capacidad de acierto con la versión ChatGPT-3 en el examen MIR 2022, realizado en Enero de 2023.

2. Métodos

Se realizó un estudio descriptivo transversal que evaluó la capacidad de la herramienta basada en inteligencia artificial, GPT-4, para responder a las preguntas del examen MIR 2023, comparando su rendimiento con los resultados del estudio previo (examen MIR 2022) donde se usó el modelo ChatGPT (también conocido como GPT-3) (2). Para ello, se introdujeron de manera estandarizada en GPT-4 las 210 preguntas en bloques de 50 preguntas.

Se realizó posteriormente un análisis por separado de cada una de las preguntas con imagen para poder asignar la tarea de respuesta a la pregunta junto con el contenido visual correspondiente. Se crearon dos bases de datos, una con las respuestas de GPT-4 y otra con las respuestas oficiales publicadas por el Ministerio de Sanidad, clasificando como correcto cuando se producía una coincidencia entre ambas. Cada pregunta del examen MIR 2023 fue

clasificada con la misma variable utilizada en el estudio previo, usando tipo especialidad, tipo de pregunta y contenido de la pregunta. Para ello, se utilizó la metodología descrita en artículos anteriores (11). Se calculó el porcentaje de acierto de cada variable por medio de una comparativa entre los rendimientos de las dos versiones de ChatGPT de cada examen MIR. Se compararon los resultados de cada examen y cada herramienta usando gráficos radiales. El análisis estadístico fue realizado con R versión 4.3.0 y librerías especializadas.

3. Resultados

La herramienta GPT-4 fue capaz de acertar 173 preguntas de un total de 210 preguntas del examen MIR 2023, lo que supone 65 aciertos más que el obtenido en el examen MIR 2022 con ChatGPT-3 (108/210) (tabla 1). En la comparación por especialidades (figura 1), GPT-4 mostró mayor capacidad de acierto en la mayor parte de las especialidades, con especial diferencia en Reumatología, Dermatología, Pediatría y Neurología. Sin embargo, en algunas especialidades como Neumología, Maxilofacial y Otorrinolaringología (ORL), el incremento en aciertos fue menos pronunciado. Se observaron las mismas tasas de acierto en Nefrología, Medicina Legal y Ética y Unidad de Cuidados Intensivos (UCI)) y solo se observó un peor rendimiento en Oftalmología.

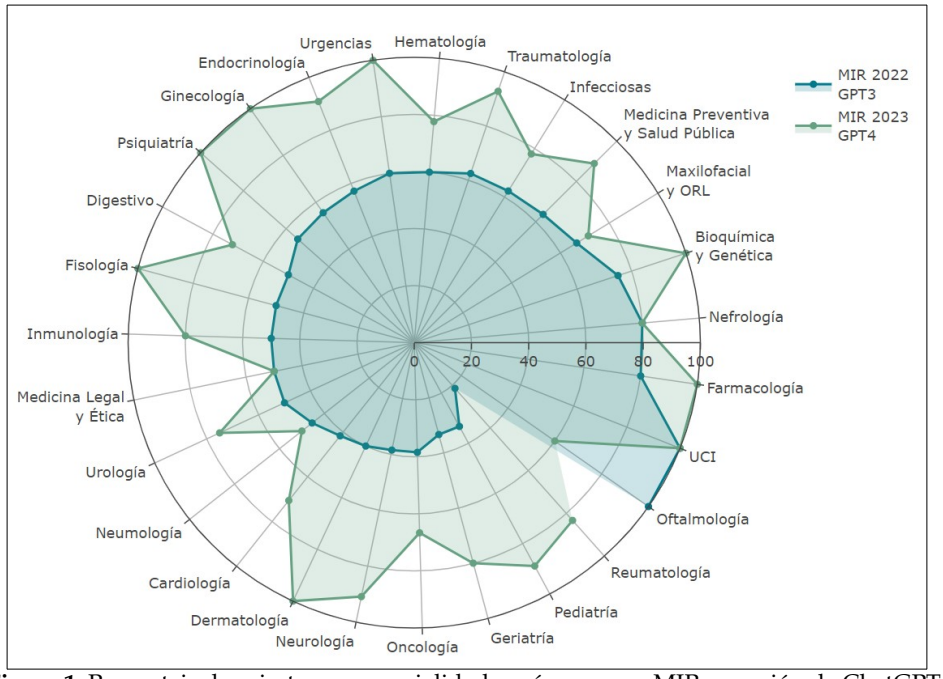

Figura 1. Porcentaje de acierto por especialidad según examen MIR y versión de ChatGPT.

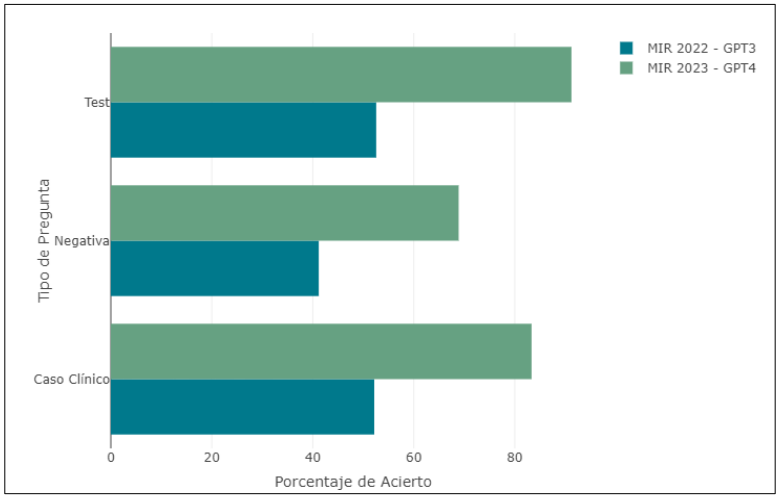

Figura 2. Porcentaje de acierto por tipo de pregunta según examen MIR y versión de ChatGPT.

El rendimiento por tipo de pregunta (figura 2) mostró una mejora sustancial con GPT-4 en todas las categorías de preguntas, siendo mayor en las preguntas tipo caso clínico y menor en las formuladas en formato negativo. Al realizar la comparación de los resultados entre las herramientas por el tipo del contenido de las preguntas (figura 3), GPT-4 superó a ChatGPT-3 especialmente en áreas como Fisiopatología, Tratamiento, Estadística y Diagnóstico.

Al trasponer el número de preguntas acertadas, a resultado neto de respuestas restando el porcentaje proporcional de preguntas falladas, GPT-4 obtuvo 153,3 aciertos netos. Al incluir la información en las distintas calculadoras de posición MIR, GPT-4 obtuvo una posición en comparación con los opositores del examen MIR 2023 de entre 1.100 y 1.300, lo que supondría un percentil entre 90 y 92.

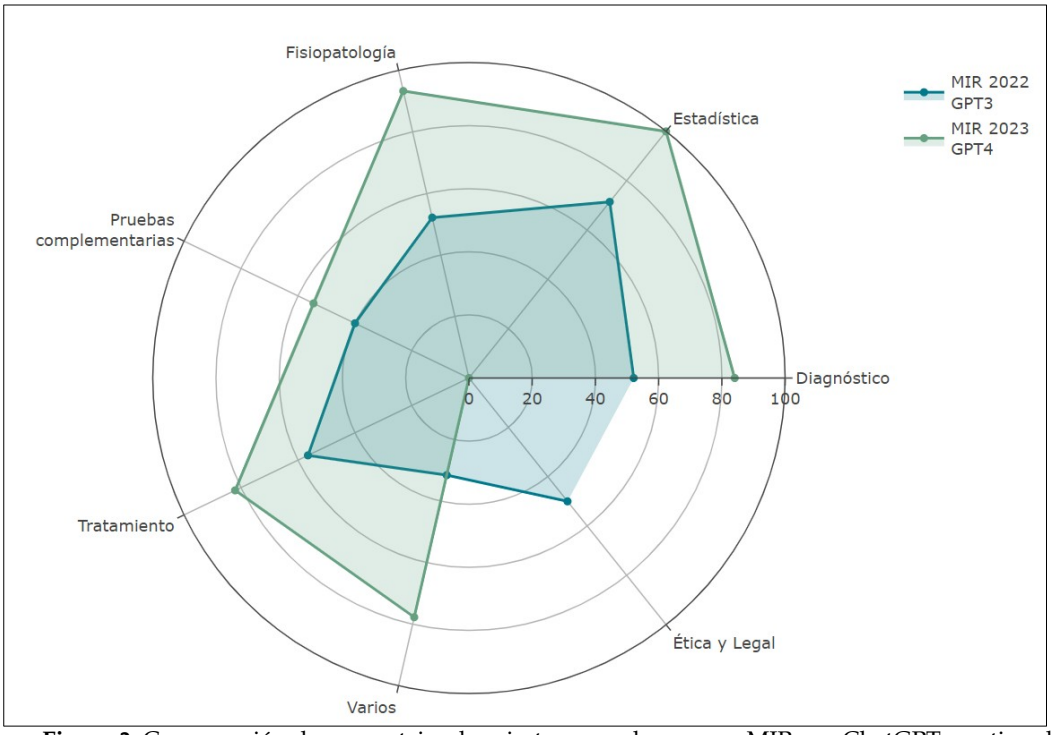

Figura 3. Comparación de porcentajes de aciertos en cada examen MIR por ChatGPT por tipo de contenido de pregunta.

Tabla 1. Respuestas totales y correctas por especialidad, tipo de pregunta y contenido según examen MIR y versión de ChatGPT.

	MIR 2022 GPT3		**MIR 2023 GPT4**	
	Total N = 210[1]	**Correctas** N = 108[1]	**Total** N = 210[1]	**Correctas** N = 173[1]
Especialidad				
Bioquímica y Genética	4 (1.9%)	3 (2.8%)	3 (1.4%)	3 (1.7%)
Cardiología	12 (5.7%)	5 (4.6%)	17 (8.1%)	12 (6.9%)
Dermatología	5 (2.4%)	2 (1.9%)	8 (3.8%)	8 (4.6%)
Digestivo	14 (6.7%)	7 (6.5%)	18 (8.6%)	13 (7.5%)
Endocrinología	14 (6.7%)	8 (7.4%)	11 (5.2%)	10 (5.8%)
Farmacología	5 (2.4%)	4 (3.7%)	6 (2.9%)	6 (3.5%)
Fisiología	4 (1.9%)	2 (1.9%)	5 (2.4%)	5 (2.9%)
Geriatría	6 (2.9%)	2 (1.9%)	5 (2.4%)	4 (2.3%)

Ginecología	9 (4.3%)	5 (4.6%)	10 (4.8%)	10 (5.8%)
Hematología	5 (2.4%)	3 (2.8%)	9 (4.3%)	7 (4.0%)
Infecciosas	8 (3.8%)	5 (4.6%)	9 (4.3%)	7 (4.0%)
Inmunología	4 (1.9%)	2 (1.9%)	5 (2.4%)	4 (2.3%)
Maxilofacial y ORL	3 (1.4%)	2 (1.9%)	7 (3.3%)	5 (2.9%)
Medicina Legal y Ética	6 (2.9%)	3 (2.8%)	2 (1.0%)	1 (0.6%)
Med. Preventiva y Salud Pública	11 (5.2%)	7 (6.5%)	9 (4.3%)	8 (4.6%)
Nefrología	5 (2.4%)	4 (3.7%)	5 (2.4%)	4 (2.3%)
Neumología	11 (5.2%)	5 (4.6%)	10 (4.8%)	5 (2.9%)
Neurología	13 (6.2%)	5 (4.6%)	11 (5.2%)	10 (5.8%)
Oftalmología	4 (1.9%)	4 (3.7%)	5 (2.4%)	3 (1.7%)
Oncología	13 (6.2%)	5 (4.6%)	9 (4.3%)	6 (3.5%)
Pediatría	9 (4.3%)	3 (2.8%)	9 (4.3%)	8 (4.6%)
Psiquiatría	11 (5.2%)	6 (5.6%)	7 (3.3%)	7 (4.0%)
Reumatología	14 (6.7%)	3 (2.8%)	6 (2.9%)	5 (2.9%)
Traumatología	8 (3.8%)	5 (4.6%)	14 (6.7%)	13 (7.5%)
UCI	3 (1.4%)	3 (2.8%)	2 (1.0%)	2 (1.2%)
Urgencias	5 (2.4%)	3 (2.8%)	4 (1.9%)	4 (2.3%)
Urología	4 (1.9%)	2 (1.9%)	4 (1.9%)	3 (1.7%)
Tipo de pregunta				
Caso Clínico	115 (55%)	60 (56%)	108 (51%)	90 (52%)
Negativa	17 (8.1%)	7 (6.5%)	45 (21%)	31 (18%)
Test	78 (37%)	41 (38%)	57 (27%)	52 (30%)
Contenido				
Diagnóstico	71 (34%)	37 (34%)	69 (33%)	58 (34%)
Estadística	7 (3.3%)	5 (4.6%)	5 (2.4%)	5 (2.9%)
Ética y Legal	6 (2.9%)	3 (2.8%)	1 (0.5%)	0 (0%)
Fisiopatología	23 (11%)	12 (11%)	30 (14%)	28 (16%)
Pruebas complementarias	15 (7.1%)	6 (5.6%)	11 (5.2%)	6 (3.5%)
Tratamiento	69 (33%)	39 (36%)	67 (32%)	55 (32%)
Varios	19 (9.0%)	6 (5.6%)	27 (13%)	21 (12%)

[1] n (%) (Porcentajes calculados por columnas)

4. Discusión

Los resultados obtenidos en el examen MIR 2023 por GPT-4 marcan un hito en la evolución de la inteligencia artificial en el ámbito de la educación médica. Este avance representa un incremento notable en el rendimiento, con 65 aciertos más en comparación con la versión anterior, ChatGPT-3, en el examen MIR 2022 (11). Este progreso es indicativo no solo de mejoras en la capacidad de procesamiento y análisis de datos por parte de la IA, sino también de un refinamiento en el procesamiento contextual y la aplicación de bases de datos y conocimientos médicos. La variabilidad en el rendimiento por especialidades médicas, con mejoras considerables en áreas como Reumatología, Pediatría, Geriatría y Oncología, refleja posiblemente una mayor capacidad de GPT-4 para integrar y aplicar conocimientos complejos en estas disciplinas. Estas mejoras pueden atribuirse a conjuntos de datos más amplios y diversificados, la inclusión de bancos de preguntas y fuentes de información relacionadas con evaluaciones como el examen MIR,

así como a algoritmos optimizados que permiten un análisis más profundo de las preguntas y una selección más precisa de las respuestas (9).

El porcentaje de acierto fue similar a la de artículos publicados en otros países con la versión GPT-4 (14-16). Esto confirma la tendencia de que las nuevas versiones de las IA tiene mejor capacidad de responder de manera correcta preguntas y desafíos médicos, superando en la gran mayoría de casos una tasa de acierto superior al 75%. Un hallazgo significativo del presente estudio radica en ilustrar qué tipo de preguntas presentan mayor dificultad para GPT-4, lo que no se ha descrito en estudios anteriores. En este sentido, destacan las preguntas negativas (formuladas desde una negación), sobre pruebas complementarias y las que integran contenidos de distintas categorías, las que resultan por segundo año consecutivo, como las preguntas con mayor dificultad para ser acertadas por una herramienta de IA (11).

Las implicaciones de estos avances para la educación médica son significativas. GPT-4 podría servir como una herramienta complementaria para la enseñanza y el aprendizaje, ofreciendo a los estudiantes de medicina una forma interactiva y adaptativa de reforzar sus conocimientos y habilidades clínicas (3). También emerge como una herramienta para los formadores, con la que generar preguntas, casos clínicos, herramientas y ejercicios de aprendizaje con los que enriquecer la formación de sus estudiantes (2).

Sin embargo, la fascinación generada por una IA que aprueba un examen MIR amerita una reflexión crítica. Es imprescindible reconocer que, aunque este logro destaca la evolución y el potencial de las LMMs, no debe interpretarse como una equiparación directa a la competencia clínica humana. Un examen, por su naturaleza, evalúa conocimientos bajo condiciones y formatos específicos, lo cual difiere sustancialmente de la complejidad y dinamismo e impredictibilidad de la práctica médica real. La capacidad de una IA para navegar por preguntas estructuradas y ofrecer respuestas basadas en datos contrasta con el juicio clínico, la empatía, y la toma de decisiones en contextos inciertos que definen la medicina clínica. Esta distinción subraya la importancia de no trasladar la capacidad de la IA en el procesamiento y reconocimiento de patrones de textos a contextos clínicos (8).

Una parte esencial del aprendizaje en la educación médica radica en la integración de conocimientos complejos, el reconocimiento de patrones, el establecimiento de una relación médico-paciente sólida y el análisis ético y contextual de las situaciones clínicas. El uso de herramientas tecnológicas, como IA, que simplifican la resolución de exámenes puede crear una ilusión de competencia, potencialmente desincentivando a los estudiantes de desarrollar estas habilidades críticas. Este enfoque podría desviar la formación médica de su objetivo fundamental: preparar profesionales con una comprensión profunda y multidimensional de la medicina, capaces de ejercer con empatía, juicio crítico y adaptabilidad en el complejo entorno clínico (4).

La integración de tecnologías emergentes, especialmente en el ámbito de la información y la comunicación, aún no ha permeado los currículos formativos en la enseñanza de Grado y la Formación Sanitaria Especializada. Es crucial que los profesionales de la salud, en general y los médicos en particular, adquieran conocimientos y competencias básicas sobre cómo operan estas tecnologías, sus beneficios, sus fallos y riesgos para poder desarrollar competencia mínima para realizar un análisis de riesgo-beneficio, entender la responsabilidad del profesional y los potenciales perjuicios para terceros al implementarlas. Esto se hace más necesario dado el actual debate de su uso como herramientas de apoyo en la toma de decisiones clínicas, una realidad que ya es

presente en algunas especialidades y que está suponiendo una enorme ayuda, pero que también sugiere un riesgo que ha de evaluarse de manera continua.

Limitaciones

Dada la trepidante rapidez con la que evolucionan este tipo de tecnologías, una limitación del presente estudio, es el corto periodo de tiempo en que los resultados de GPT-4 probablemente se vean superados por los de otras herramientas de IA o por nuevas versiones de la misma. En este sentido, los resultados del presente estudio es necesario contextualizarlos en el tiempo en el que fueron publicados. Por otro lado, dado que la herramienta ChatGPT se ha desarrollado principalmente en inglés, es posible que los resultados de nuestro estudio, al haber sido presentados a la herramienta en el idioma del examen MIR (castellano), infravaloren la capacidad de esta herramienta para acertar preguntas de contenido médico. Sin embargo, de manera contraria a esta posibilidad, los resultados coinciden de manera general con aquellos que han sido publicados en países de habla inglesa. Por último, la comparativa realizada entre ChatGPT-3 y GPT-4 corresponde a dos exámenes distintos respectivamente, el examen MIR 2022-2023 y el 2023-2024. Sin embargo, dada la representatividad que muestran dichos exámenes del conocimiento médico general y la marcada coincidencia con la bibliografía internacional en el porcentaje de aciertos de ambas herramientas, consideramos que la comparación continua siendo válida e interesante para el debate científico en este ámbito.

5. Conclusiones

- El presente estudio revela que GPT-4 no solo supera a su predecesor ChatGPT-3 en el examen MIR, sino que a su vez establece un nuevo estándar en la capacidad de las IA para procesar y analizar información médica de alta complejidad.
- Estos resultados subrayan la continua evolución y mejora de las herramientas de IA en el ámbito de la educación médica. Sin embargo, es crucial mantener un enfoque crítico y consciente de las limitaciones inherentes a la IA, especialmente en comparación con la competencia clínica humana y la toma de decisiones en situaciones médicas reales.
- Este avance en la tecnología de IA abre un camino prometedor hacia una integración más efectiva en la educación médica, potenciando el aprendizaje y la preparación de futuros médicos, aunque siempre complementando y no sustituyendo el juicio y la experiencia humana en los procesos de formación.

Financiación: No ha habido financiación.

Declaración de conflicto de interés: Los autores declaran no tener ningún conflicto de intereses.

Contribuciones de los autores: ACC ha redactado la versión final del artículo y coordinado el proyecto; JJ ha coordinado y realizado en el análisis de datos y en la redacción de la versión final del artículo; JPC ha redactado el primer borrador y ha participado en el análisis de datos; DAS ha participado en el análisis de datos y ha supervisado la redacción final del artículo; PE ha participado en la redacción de la versión final y en el análisis de datos; LDLP, EG y JCP han realizado la búsqueda bibliográfica, han participado en el análisis de datos y en la redacción del primer borrador.

Referencias

1. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: Is ChatGPT a blessing or blight in disguise? Med Educ Online. 2023 Dec; 28(1): 2181052. https://doi.org/10.1080/10872981.2023.2181052
2. OpenAI. Introducing ChatGPT. Disponible en: https://openai.com/blog/chatgpt/
3. Temsah O, Khan SA, Chaiah Y, Senjab A, Alhasan K, Jamal A, Aljamaan F, Malki KH, Halwani R, Al-Tawfiq JA, Temsah MH, Al-Eyadhy A. Overview of Early ChatGPT's Presence in Medical Literature:

Insights From a Hybrid Literature Review by ChatGPT and Human Experts. Cureus. 2023 Apr 8; 15(4): e37281. https://doi.org/10.7759/cureus.37281

4. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, Househ M. The Pros and Cons of Using ChatGPT in Medical Education: A Scoping Review. Stud Health Technol Inform. 2023 Jun 29; 305: 644-647. https://doi.org/10.3233/shti230580
5. Dergaa I, Chamari K, Zmijewski P, Ben Saad H. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. Biol Sport. 2023 Apr; 40(2): 615-622. https://doi.org/10.5114/biolsport.2023.125623
6. Davidson T, Bhattacharya D, Weber I. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In Proceedings of the Third Workshop on Abusive Language Online. 2019, pages 25–35, Florence, Italy. Association for Computational Linguistics. https://aclanthology.org/W19-3504/
7. Hadas K, Rikker D, David D. Gender bias and stereotypes in Large Language Models. In Proceedings of The ACM Collective Intelligence Conference (CI '23). Association for Computing Machinery, New York, NY, USA. 2023. 12–24. https://doi.org/10.1145/3582269.3615599
8. Jeyaraman M, Ramasubramanian S, Balaji S, Jeyaraman N, Nallakumarasamy A, Sharma S. ChatGPT in action: Harnessing artificial intelligence potential and addressing ethical challenges in medicine, education, and scientific research. World J Methodol. 2023 Sep 20; 13(4): 170-178. https://doi.org/10.5662/wjm.v13.i4.170
9. Sahu PK, Benjamin LA, Singh Aswal G, Williams-Persad A. ChatGPT in research and health professions education: challenges, opportunities, and future directions. Postgrad Med J. 2023 Dec 21; 100(1179): 50-55. https://doi.org/10.1093/postmj/qgad090
10. Scaioli G, Lo Moro G, Conrado F, Rosset L, Bert F, Siliquini R. Exploring the potential of ChatGPT for clinical reasoning and decision-making: a cross-sectional study on the Italian Medical Residency Exam. Ann Ist Super Sanita. 2023 Oct-Dec;59(4):267-270. https://doi.org/10.4415/ann_23_04_05
11. Carrasco JP, García E, Sánchez DA, Estrella-Porter P, Puente LDL, Navarro J, et al. ¿Es capaz "ChatGPT" de aprobar el examen MIR de 2022? Implicaciones de la inteligencia artificial en la educación médica en España. Rev Esp Educ Médica 2023 Feb 16; 4(1). https://revistas.um.es/edumed/article/view/556511/337361
12. Madrid-García, A., Rosales-Rosado, Z., Freites-Nuñez, D. et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. Sci Rep 13, 22129 (2023). https://doi.org/10.1038/s41598-023-49483-6
13. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, Alas-Brun R, Onambele L, Ortega W, Montejo R, Aguinaga-Ontoso E, Barach P, Aguinaga-Ontoso I. Evaluating the Efficacy of ChatGPT in Navigating the Spanish Medical Residency Entrance Examination (MIR): Promising Horizons for AI in Clinical Medicine. Clinics and Practice. 2023; 13(6):1460-1487. https://doi.org/10.3390/clinpract13060130
14. Mihalache A, Huang RS, Popovic MM, Muni RH. GPT-4: An assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach. 2023 Oct 15:1-7. https://doi.org/10.1080/0142159x.2023.2249588
15. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of ChatGPT.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. JMIR Med Educ 2023;9:e48002. https://doi.org/10.2196/48002
16. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI. Int J Med Inform. 2023 Sep; 177: 105173. https://doi.org/10.1016/j.ijmedinf.2023.105173
17. Jung LB, Gudera JA, Wiegand TLT, Allmendinger S, Dimitriadis K, Koerte IK. ChatGPT Passes German State Examination in Medicine With Picture Questions Omitted. Dtsch Arztebl Int. 2023 May 30;120(21):373-374. https://doi.org/10.3238/arztebl.m2023.0113
18. Alessandri Bonetti M, Giorgino R, Gallo Afflitto G, De Lorenzi F, Egro FM. How Does ChatGPT Perform on the Italian Residency Admission National Exam Compared to 15,869 Medical Graduates? Ann Biomed Eng. 2023 Jul 25. https://doi.org/10.1007/s10439-023-03318-7
19. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB. Evaluating ChatGPT Performance on the Orthopaedic In-Training Examination. JBJS Open Access 8(3): e23.00056, July-September 2023. https://doi.org/10.2106/jbjs.oa.23.00056

20. Kufel J, Paszkiewicz I, Bielówka M, Bartnikowska W, Janik M, Stencel M, Czogalik Ł, Gruszczyńska K, Mielcarska S. Will ChatGPT pass the Polish specialty exam in radiology and diagnostic imaging? Insights into strengths and limitations. Pol J Radiol. 2023 Sep 18;88:e430-e434. https://doi.org/10.5114%2Fpjr.2023.131215
21. Mehrabanian M, Zariat Y. ChatGPT passes anatomy exam. Br Dent J. 2023 Sep; 235(5): 295. https://doi.org/10.1038/s41415-023-6286-7