Revista Española de Educación Médica

¿Es capaz "ChatGPT" de aprobar el examen MIR de 2022? Implicaciones de la inteligencia artificial en la educación médica en España

Is "ChatGPT" capable of passing the 2022 MIR exam? Implications of artificial intelligence in medical education in Spain.

Carrasco JP[1], García E[2], Sánchez DA[3*], Estrella Porter PD[4], De La Puente L[5], Navarro J[6], Cerame A[7].

[1] Hospital Clínico Universitario de Valencia. Valencia. juanpablocarrascopicazo@gmail.com (0000-0001-9137-7775).
[2] Complejo Hospitalario Universitario Toledo. Toledo. eva.garciacamacho@gmail.com (0000-0001-8962-6023).
[3] Hospital Clínico Universitario Virgen de la Arrixaca. IMIB. Murcia. Consejo General de Colegios Oficiales de Médicos de España. Madrid. dsanchez@cgcom.es (0000-0003-2073-0679).
[4] Hospital Clínico Universitario de Valencia. Valencia. pestrellaporter@gmail.com (0000-0003-4137-7691).
[5] Hospital Universitario y Politécnico La Fe. Valencia. med.delapuente@gmail.com.
[6] Hospital Universitario Virgen del Rocío. Sevilla; joaquinalajar@gmail.com (0000-0002-7983-7289)
[7] Plan de Atención Integral al Profesional Sanitario Enfermo. Servicio Madrileño de salud. alvaro.cerame@salud.madrid.org (0000-0003-0469-8461).

* Correspondencia: dsanchez@cgcom.es.

Recibido: 7/2/23; Aceptado: 15/2/23; Publicado: 16/2/23

Resumen: La inteligencia artificial y los modelos de procesamiento de lenguaje natural han irrumpido con fuerza en el ámbito de la educación médica. Entre ellos, el modelo ChatGPT ha sido utilizado para intentar resolver distintos exámenes de medicina a nivel internacional. Sin embargo, prácticamente no existe literatura en Europa ni países de habla hispana. El presente trabajo pretende evaluar la capacidad de responder preguntas del modelo ChatGPT en el examen MIR 2022. Para ello, se ha llevado a cabo un análisis transversal y descriptivo en el que se ha introducido en dicho modelo las 210 preguntas del examen MIR convocado en 2022 y realizado en enero de 2023. ChatGPT ha sido capaz de responder de manera acertada un 51,4% de las preguntas, lo que supone aproximadamente 69 netas en el examen MIR. Según estimaciones para este año, obtendría un 7688, lo que estaría ligeramente por debajo de la mediana de la población presentada, pero que le permitiría pasar la nota de corte y escoger un gran número de especialidades. El resultado es similar a los obtenidos en la bibliografía previa, ligeramente por debajo de los resultados obtenidos por dicha herramienta en los exámenes americanos USMLE. Este tipo de modelos suponen una oportunidad para el aprendizaje (análisis de razonamiento, generación de debates, etc.) de los estudiantes de medicina y los residentes, pero también supone un riesgo en muchos sentidos, especialmente en cuanto a la veracidad, ética y seguridad de la información. Es fundamental formar a los futuros especialistas en la nueva realidad de la inteligencia artificial para que sean capaces de utilizarlas y obtener beneficios de manera razonada y segura.

Palabras clave: ChatGPT; examen MIR; inteligencia artificial; educación médica, formación postgraduada.

Abstract: Artificial intelligence and natural language processing models have burst into the field of medical education with force. Among them, the ChatGPT model has been used to try to solve different medicine exams at an international level. However, there is practically no literature in Europe or Spanish-speaking countries. The present work aims to evaluate the ability to answer questions of the ChatGPT model in the MIR 2022 exam. To do this, a cross-sectional and descriptive analysis has been carried out in which the 210 questions of the MIR exam convened in 2022 have been introduced into said model. and

carried out in January 2023. ChatGPT has been able to correctly answer 51.4% of the questions, which means approximately 69 net in the MIR exam. According to estimates for this year, he would obtain a 7688, which would be slightly below the median of the population presented, but which would allow him to pass the cut-off mark and choose a large number of specialties. The result is similar to those obtained in the previous bibliography, slightly below the results obtained by said tool in the American USMLE exams. These types of models represent an opportunity for learning (analysis of reasoning, generation of debates, etc.) for medical students and residents, but they also pose a risk in many ways, especially in terms of veracity, ethics, and security. of the information. It is essential to train future specialists in the new reality of artificial intelligence so that they are able to use them and obtain benefits in a reasoned and safe way.

Keywords: ChatGPT; mir exam; artificial intelligence; medical education; postgraduate training.

1. Introducción

ChatGPT, o Chat Generative Pre-trained Transformer, es un modelo de inteligencia artificial y procesamiento del lenguaje natural (PLN) de 175.000 millones de parámetros que utiliza algoritmos de aprendizaje entrenados en grandes cantidades de datos para generar respuestas de tipo humano a las preguntas de los usuarios (1). Desde su lanzamiento ha cosechado un gran éxito, siendo capaz de generar respuestas automáticas a peticiones complejas como la elaboración de resúmenes, poemas, textos de programación informática y complejos problemas matemáticos. En el mundo de la educación médica, este tipo de algoritmos también han comenzado a atraer la atención de docentes y discentes.

La Asociación Médica Mundial aboga por una revisión de los planes de estudio de medicina y de las oportunidades educativas para fomentar una mejor comprensión de los numerosos aspectos de la inteligencia artificial (IA) en la atención sanitaria, tanto positivos como negativos (2). Además, en una declaración de 2019, el Comité Permanente de Médicos Europeos (CPME) subrayó la necesidad de utilizar sistemas de IA en la formación médica básica y continuada (3). Sin embargo, existen numerosas preocupaciones éticas en la utilización de este tipo de tecnologías. Entre ellas, destaca la amenaza a la seguridad y privacidad, la naturaleza cambiante de la relación médico-paciente en el ámbito de la salud, la generación de posibles desigualdades sociales y el desarrollo de IA que pudieran acabar sustituyendo muchas tareas profesionales, con el consiguiente aumento de las tasas de desempleo (4,5).

Dentro de las distintas oportunidades que ofrece la inteligencia artificial, los modelos lingüísticos se han comenzado a investigar como herramientas para la interacción personalizada con el paciente y la educación sanitaria de los ciudadanos (6-7). Aunque han demostrado su potencial en distintas áreas, estos modelos están pendientes de mostrar su capacidad en las áreas de comprobación de conocimientos clínicos mediante tareas generativas de pregunta-respuesta (QA). Entre la bibliografía existente, encontramos que Jin et al (8) lograron un 68,1% con su modelo de inteligencia artificial. Éste responde a preguntas de Sí/No utilizando como base de información el conjunto de resúmenes disponibles en Pubmed. Otro artículo también de Jin et al (9) logró una precisión del 36,7% en un conjunto de datos de 12.723 preguntas derivadas de exámenes de licencias médicas chinas. Del mismo modo, en 2019 Ha y Yaneva (10) informaron de un 29% de precisión en 454 preguntas del USMLE Step 1 y Step 2.

En este ámbito, la IA ChatGPT ha mostrado resultados más prometedores que modelos anteriores. Gilson et al (11) y Kung et al (12), encontraron que ChatGPT es capaz de responder correctamente más del 60% de las preguntas que representan temas cubiertos en los exámenes de licencia USMLE Step 1 y Step 2. A su vez, Huh S (13) en un artículo sobre parasitología y Anaki S. et al (14) en otro sobre oftalmología, obtienen resultados de entre el 50 y el 60%. Sin embargo, existe la limitación de que prácticamente toda la bibliografía citada anteriormente está realizada en Asia y Norteamérica, existiendo un vacío en la literatura en Europa y en países de habla hispana.

Intentando responder a todo lo anterior, el presente artículo tiene como objetivo principal evaluar la capacidad de rendimiento de ChatGPT en el examen de acceso a la Formación Sanitaria Especializada en España, el conocido como examen MIR, en su edición del 2022, que realizaron los aspirantes a esta formación en enero de 2023. Como objetivo secundario, el artículo pretende evaluar la capacidad de acierto del modelo en función de la especialidad, el tipo y el contenido de las distintas preguntas.

2. Métodos

Se ha llevado a cabo un análisis transversal y descriptivo en el que se han introducido las 210 preguntas del examen MIR convocado en 2022 por el Ministerio de Sanidad y celebrado en enero del 2023, en la versión 0 de la herramienta de inteligencia artificial ChatGPT, con la siguiente introducción: "¿Cuál es la respuesta correcta a la siguiente pregunta del MIR 2022 en España?". Se decidió analizar todas las preguntas del examen, incluyendo aquellas que tenían imagen asociada a pesar de que la imagen no podía introducirse, con el objetivo de tener una visión global de su capacidad de resolución con todas las preguntas del examen. Se introdujeron las preguntas en el mencionado chat del 2 al 5 de febrero de 2023. El resultado ofrecido por ChatGPT de la pregunta se ha comparado con la plantilla de respuestas publicada por el Ministerio de Sanidad del Gobierno de España, estableciéndose cada pregunta como correcta o incorrecta. Además, cada pregunta ha sido clasificada siguiendo los siguientes 4 parámetros:

- **Especialidad de la pregunta:** se ha clasificado cada pregunta en función de la especialidad y/o asignatura correspondiente a la misma.
- **Tipo de especialidad:** se ha clasificado cada pregunta en una variable múltiple acotada en función del tipo de especialidad/asignatura, siendo las variables: "médica", "quirúrgica", "básica", "preventiva" y "ética y legal".
- **Tipo de pregunta:** se ha clasificado cada pregunta como "CC" (caso clínico), si la pregunta incluía la información de un paciente o no; "test" cuando la pregunta no incluía un CC y era afirmativa; y "negativa" cuando la pregunta no incluía un CC y preguntaba por la respuesta incorrecta/errónea.
- **Contenido de la pregunta:** se ha clasificado cada pregunta en las siguientes categorías: "diagnóstico", "tratamiento", "pruebas complementarias", "fisiopatología", "estadística" o "ética y legal" en función de qué tipo de información era preguntado en cada cuestión. Si además la pregunta hacía referencia a varias categorías de las anteriores, se ha clasificado como "varios".

Cada uno de los autores ha evaluado y clasificado 30 preguntas del examen MIR 2022, y posteriormente todas ellas han sido revaluadas y discutidas por todo el resto de autores de manera conjunta. Para el análisis cuantitativo de las distintas preguntas, se ha utilizado el programa SPSS v. 25.

3. Resultados

La herramienta ChatGPT ha sido capaz de acertar de manera correcta el 51,4% de las preguntas. Este porcentaje aumenta al 54,8% al analizar únicamente las preguntas que no tenían ninguna imagen asociada (185 preguntas). Realizando la conversión a resultado "neto" del examen, tal y como realiza la clasificación de los resultados el Ministerio de Sanidad, supondría un total de 69,33 respuestas netas.

En cuanto a la evaluación de la capacidad de acierto de ChatGPT por especialidades, se ha realizado un desglose de las mismas, como se puede observar en la tabla 1. Destaca el porcentaje de acierto en especialidades como oftalmología y UCI, donde obtiene el 100% de respuestas correctas, seguido de farmacología y nefrología con un 80%. En sentido negativo, las especialidades donde acierta menos preguntas son reumatología (21,4%), geriatría (33,3%) y pediatría (33,3%). Sin embargo, dichos resultados han de tomarse con extremada cautela, ya que el número de preguntas que los justifica es muy bajo, y los mismos pueden ser debidos al azar.

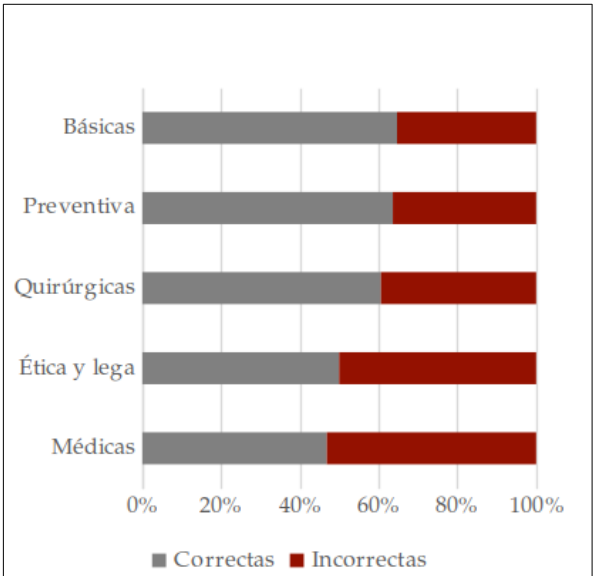

Figura 1. Porcentaje de acierto de ChatGPT en el examen MIR 2022 por tipo de especialidad/asignatura.

En el análisis por tipo de especialidad/asignatura, destaca que las preguntas sobre especialidades médicas son en las que se produce un menor porcentaje de acierto por parte de la herramienta informática ChatGPT (46,9%) y las preguntas de "básicas" y "preventiva" las que más porcentaje de acierto han obtenido (64,7% y 63,6% respectivamente), tal y como se puede observar en la figura 1.

Respecto al análisis por tipo de pregunta (figura 2), las preguntas redactadas en formato negativo son las que menos porcentaje de acierto han obtenido (41%) frente a las preguntas afirmativas categorizadas como CC (52%) y test (53%). En cuanto al contenido de la pregunta, resulta significativo que aquellas preguntas que incluyen información de distintos ámbitos (categorizada como "varios") son las que menos porcentaje de acierto han tenido, con un 32%, como se observa en la figura 3.

Tabla 1. Resultados del modelo ChatGPT en el examen MIR 2022

Especialidad	Total preguntas	Preguntas correctas	% preguntas correctas	Especialidad	Total preguntas	Preguntas correctas	% Preguntas correctas
Digestivo	14	7	50,0%	**Medicina Legal y Ética**	6	3	50,0%
Endocrinología	14	8	57,1%	**Dermatología**	5	2	40,0%
Reumatología	14	3	21,4%	**Farmacología**	5	4	80,0%
Neurología	13	5	38,5%	**Hematología**	5	3	60,0%
Oncología	13	5	38,5%	**Nefrología**	5	4	80,0%
Cardiología	12	5	41,7%	**Urgencias**	5	3	60,0%
Neumología	11	5	45,5%	**Bioquímica y genética**	4	3	75,0%
Preventiva	11	7	63,6%	**Fisiología**	4	2	50,0%
Psiquiatría	11	6	54,5%	**Inmunología**	4	2	50,0%
Ginecología	9	5	55,6%	**Oftalmología**	4	4	100,0%
Pediatría	9	3	33,3%	**Urología**	4	2	50,0%
Infecciosas	8	5	62,5%	**Maxilofacial y ORL**	3	2	66,7%
Traumatología	8	5	62,5%	**UCI**	3	3	100,0%
Geriatría	6	2	33,3%				

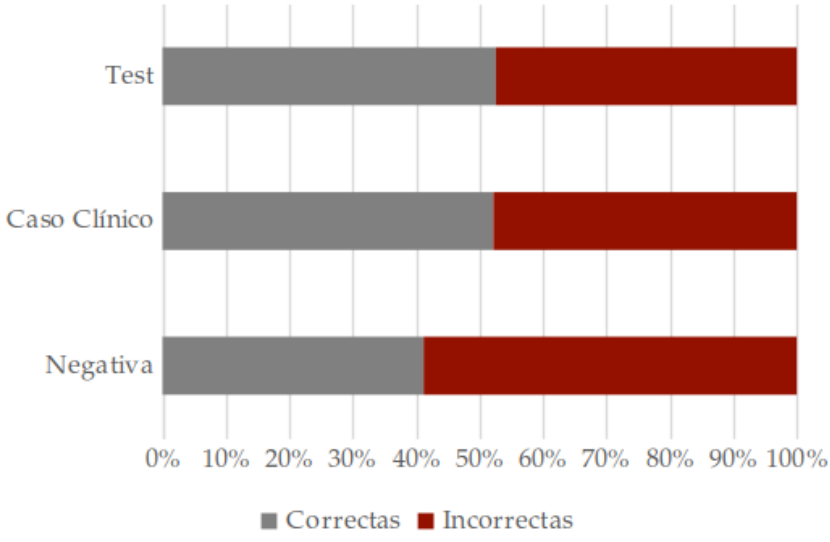

Figura 2. Porcentaje de acierto de ChatGPT en el examen MIR 2022 por tipo de pregunta.

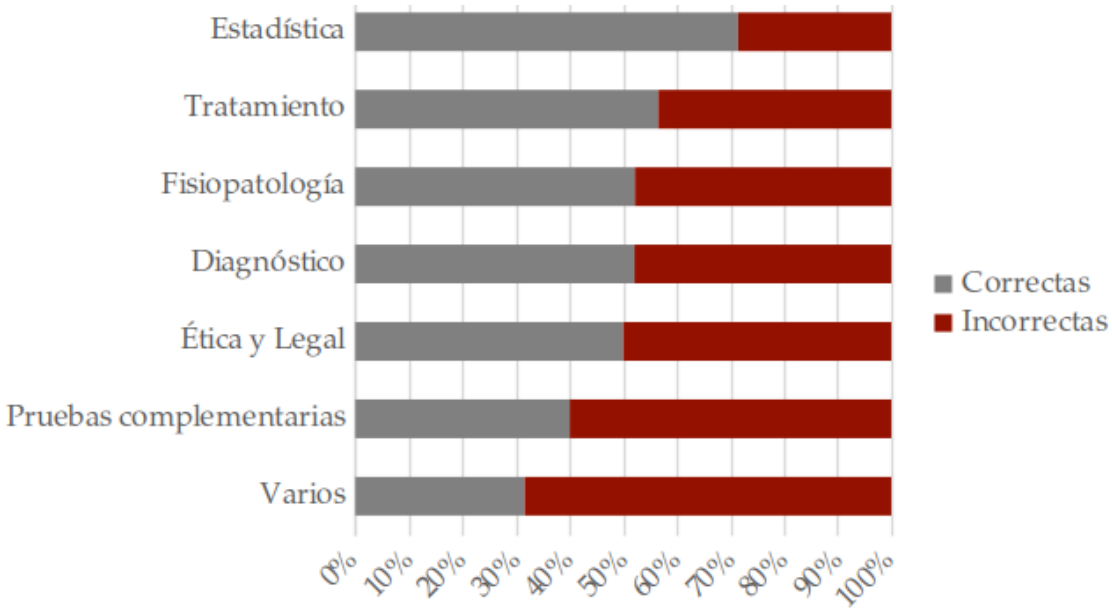

Figura 3. Porcentaje de acierto de ChatGPT en el examen MIR 2022 según el contenido de la pregunta.

4. Discusión

El presente estudio muestra la capacidad de la herramienta ChatGPT de afrontar el examen MIR 2022. Se ha obtenido un 51,4% de aciertos, lo que se asemeja a los mostrados en estudios previos, resultando levemente por debajo de los estudios de EEUU y Asia. Gibson et al. [11] describen en su estudio realizado con bases de datos de preguntas del USMLE una tasa de acierto del 60%, cifra que suele ser el punto de corte de la prueba y equivale al nivel de un estudiante de tercer curso de medicina en EEUU. Esta tasa de aciertos también es similar a la presentada por Huh [12] donde enfrenta a ChatGPT a un

examen de parasitología frente a los estudiantes de medicina obteniendo un 60,8% vs. 90,8% respectivamente. Es importante recalcar que a diferencia del examen USMLE, que tiene preguntas altamente estandarizadas y reguladas, el examen MIR presenta una alta variabilidad y heterogeneidad en su dificultad y complejidad de preguntas, lo que puede limitar la interpretación de la aplicación de ChatGPT comparado con otros estudios publicados. Sin embargo, confirma que ChatPGT obtiene mejores resultados que otros modelos, ya que supera los estudios publicados por Ha et al. (10) y Jin et al. (9).

Hasta donde se tiene conocimiento, no existen estudios previos analizando el examen MIR con herramientas de IA, ni tampoco de otros exámenes en Europa o en países de habla hispana. La novedad de nuestro estudio es el análisis de las características de las preguntas que orientan a la capacidad de análisis de la herramienta. Estudios previos como el de Gibson et al. (11) analizaban el tipo de respuesta que ofrece ChatGPT sin tener en cuenta la formulación, el tipo o especialidad de la pregunta ya que la misma puede estar condicionada por el tipo de pregunta tal y como se ha evidenciado en nuestro estudio.

Acorde con los resultados del estudio, parece significativo el tipo de especialidad/asignatura que se pregunta, ya que existe una diferencia de acierto entre preguntas relacionadas con especialidades clínicas frente básicas con un 46,9% vs. 64,7% de aciertos, así como en aquellas preguntas formuladas como negativas frente a las tipo test clásicas o los casos clínicos, 41% vs. 53%-52%, respectivamente. A su vez, cuando se pregunta información de distintos ámbitos (diagnóstico, tratamiento, PC, etc), se obtiene el peor resultado, con un 32%. Esto podría estar relacionado con las conclusiones de algunos trabajos previos, como los de Jin et al. (8) y Gibson et al. (11), quienes destacan que la capacidad del modelo para responder correctamente una pregunta puede estar relacionada con la complejidad de la misma y con su capacidad para relacionar el mensaje con los datos dentro de su corpus.

Al transformar los resultados de ChatGPT en el examen MIR 2022 se obtuvo una puntuación de 69,33 respuestas netas. Según las herramientas estadísticas de una de las academias de preparación del examen (14), estableciendo la nota media de expediente más baja (5 sobre 10), la IA alcanza una posición estimada de 7688 en el examen MIR, con 9277 exámenes corregidos a fecha 4 de febrero de 2023. Este resultado pasaría la nota de corte del examen (25% de las 10 mejores notas, con un máximo teórico de 50 netas) e incluso estaría dentro de las 8550 plazas ofertadas en esta convocatoria. Basándonos en las plazas escogidas en la convocatoria MIR 2021 (15) y con los resultados de esta herramienta, las siguientes plazas estaban disponibles: Inmunología, Neurofisiología Clínica, Medicina Nuclear, Geriatría, Farmacología Clínica, Microbiología, Análisis Clínicos, Medicina del Trabajo, Bioquímica Clínica, Medicina Preventiva y Salud Pública o Medicina de Familia y Comunitaria.

Las herramientas de IA como ChatGPT, son un fenómeno creciente que va a formar parte de nuestra realidad educativa y clínica y del cuál es interesante incorporar competencias en el ámbito de la educación médica. Las ventajas son múltiples, la más inmediata es la resolución prácticamente instantánea de preguntas del estudiante sobre conceptos, diagnósticos o tratamientos médicos específicos y recibir respuestas precisas y personalizadas para ayudarlos a estructurar mejor su conocimiento. Desde el punto de vista del docente, puede utilizarse como método de evaluación de conocimientos, principalmente en entornos a distancia, generación de casos o búsqueda de información médica relevante y actualizada de manera rápida y eficiente [16]. El modelo también permite crear un ambiente de aprendizaje interactivo y a petición individualizada gracias a sus respuestas personalizadas y fácilmente comprensibles, lo que puede mejorar la

retención de información y hacer que el proceso de aprendizaje sea más agradable para los estudiantes [11]. Sin embargo, el uso de estas herramientas requiere un gran espíritu crítico así como una formación en ciencias básicas. ChatGPT está entrenado con una gran cantidad de datos, su precisión puede ser limitada y puede producir respuestas incorrectas al no tener acceso a internet ni respuestas con información a tiempo real y los sesgos utilizados en el entrenamiento de datos, pueden acentuar sesgos socioculturales existentes [17]. Además, la falta de retroalimentación y discusión en tiempo real, así como la falta de adaptabilidad y contexto clínico puede limitar la comprensión y el desarrollo de habilidades clínicas en los estudiantes. Asimismo, es importante destacar que estas herramientas carecen de criterios éticos y legales por los que debe guiarse cualquier acto médico.

El presente trabajo presenta ciertas limitaciones. Por un lado, no se ha analizado a nivel cualitativo la justificación de las respuestas, al haber tenido únicamente en cuenta la opción sin atender al razonamiento ofrecido por la herramienta. Se encontró que las respuestas correctas contenían información externa a la pregunta con una frecuencia significativamente mayor que las respuestas incorrectas. Por otro lado, el análisis realizado es puramente descriptivo, sin presentar un análisis estadístico que pueda confirmar que las diferencias encontradas no se deben al azar. Otra de las limitaciones corresponde a que las preguntas del examen MIR y los formatos de respuesta múltiple presentan una lógica interna a la hora de responder a las preguntas y de formularlas que pertenece al entrenamiento para el examen. Destacaría el hecho de que la herramienta no ha sido entrenada para identificar estos patrones y que esto llevaría necesariamente a obtener un resultado peor.

Dados los resultados del presente estudio, sería interesante proseguir la investigación de la inteligencia artificial en la educación médica en distintos sentidos. En primer lugar, sería interesante replicar los resultados del presente estudio con exámenes MIR de años anteriores y otros exámenes médicos en países de habla hispana, para comparar si la capacidad de ChatGPT se mantiene o no. Además, existe un potencial a remarcar en el análisis cualitativo tanto de las respuestas del modelo a preguntas clínicas y éticas, como a la opinión de los estudiantes respecto a estos modelos, sus dificultades y sus expectativas. Por último, es necesario continuar investigando la problemática ética y de seguridad que pueden suponer estas herramientas, ya que es un ámbito poco estudiado y que puede hacer peligrar su utilidad en el ámbito médico y educativo.

5. Conclusiones

- Se trata del primer estudio realizado en un país de habla hispana que analiza el rendimiento y los potenciales beneficios y perjuicios del uso de IA y herramientas de PNL en el contexto de los exámenes como procesos de aprendizaje en la educación médica.

- El modelo ChatGPT ha sido capaz de pasar la nota de corte del examen MIR 2022, con un 51% de preguntas acertadas. El porcentaje de acierto ha variado en función del tipo de pregunta, del contenido y de la especialidad y/o asignatura de la misma.

- Los porcentajes más bajos de aciertos los ha presentado en preguntas de especialidades médicas, formuladas en negativo y cuando se incluye información de distintos ámbitos (tratamiento, diagnóstico, pruebas complementarias, etc).

- Transformando el resultado de ChatGPT en el examen MIR, se habría obtenido 69,33 netas, con un número de orden aproximado de 7688. Con este número, en la convocatoria MIR 2021 podría haber elegido múltiples especialidades en distintos hospitales a lo largo del Estado.

- Los resultados son similares a estudios previos realizados con ChatGPT en otros países, obteniendo levemente peores resultados que en los trabajos publicados en los exámenes USMLE de Estados Unidos.

Financiación: No ha habido financiación.

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

Contribuciones de los autores: JPC ha redactado la versión final del artículo y coordinado el proyecto; EG ha redactado el primer borrador y ha participado en el análisis de datos; DAS ha sido el promotor del proyecto, 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 y JN han realizado la búsqueda bibliográfica, han participado en el análisis de datos y en la redacción del primer borrador; AC supervisado la redacción final del artículo, co-coordinado y participado en el análisis de datos.

Referencias

1 Scott K. Microsoft teams up with OpenAI to exclusively license GPT-3 language model 2020. Official Microsoft Blog. https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/

2 WMA statement on augmented intelligence in medical care. World Medical Association. 2019. https://www.wma.net/policies-post/wma-statement-on-augmented-intelligence-in-medical-care/

3 Standing Committee of European Doctors (CPME) 2019. Policy on AI in Healthcare. https://www.cpme.eu/api/documents/adopted/2019/CPME_AD_Board_16112019_062_FINAL_EN_CPME.AI_.in_.health.care_.pdf

4 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019; 25:44-56. https://doi.org/10.1038/s41591-018-0300-7

5 Wartman SA, Donald Combs C. Medical education must move from the information age to the age of artificial intelligence. Acad Med. 2018;93:1107-9. https://doi.org/10.1097/acm.0000000000002044

6 Avisha Das, Salih Selek, Alia R. Warner, Xu Zuo, Yan Hu, Vipina Kuttichi Keloth, Jianfu Li, W. Jim Zheng, and Hua Xu. Conversational Bots for Psychotherapy: A Study of Generative Transformer Models Using Domain-specific Dialogues. 2022. In Proceedings of the 21st Workshop on Biomedical Language Processing, pages 285-297, Dublin, Ireland. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2022.bionlp-1.27

7 Savery M, Abacha AB, Gayen S, Demner-Fushman D. Question-driven summarization of answers to consumer health questions. Sci Data. 2020;7(1):322. https://doi.org/10.1038/s41597-020-00667-z

8 Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: A dataset for biomedical research question answering arXiv preprint arXiv:1909.06146. 2019. https://doi.org/10.48550/arXiv.1909.06146

9 Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. Applied Sciences. 2021;11:6421. https://doi.org/10.3390/app11146421

10 Ha LE, Yaneva V. Automatic question answering for medical MCQs: Can it go further than information retrieval? Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). https://doi.org/10.26615/978-954-452-056-4_049

11 Gilson A, Safranek C, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ 2023; 9: e45312. https://mededu.jmir.org/2023/1/e45312

12 Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof 2023;20:1 https://doi.org/10.3352/jeehp.2023.20.1

13 Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of its Successes and Shortcomings. medRxiv. 1 de enero de 2023; https://doi.org/10.1101/2023.01.22.23284882

14 Grupo CTO. Servicio post-mir de corrección de exámenes. Enero 2023. Disponible en: https://medicina.grupocto.es/postmir/

15 Uhrig A. La elección telemática de plaza MIR 2022 deja 218 vacías: En qué número se agota cada especialidad? Consalud.es. Enero 2023. https://www.consalud.es/especial-mir/adjudicadas-todas-plazas-mir-2022-en-numero-se-agoto-cada-especialidad_115000_102.html
16 Baidoo-Anu D and Owusu Ansah, L. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. 2023. https://ssrn.com/abstract=4337484 or http://dx.doi.org/10.2139/ssrn.4337484
17 Yue Zhuo T. Exploring AI Ethics of ChatGPT: A Diagnostic Analysis. 2023. Disponible en: https://arxiv.org/abs/2301.12867 https://doi.org/10.48550/arXiv.2301.12867

Anexo I. Codificación de las preguntas del examen MIR 2023 (abreviaturas: CC, caso clínico; PC, pruebas complementarias).

Nº Pregunta	Resultado ChatGPT	Plantilla versión 0	Respuesta correcta	Especialidad / Asignatura	Tipo pregunta	Contenido pregunta	Tipo especialidad
1	1	3	NO	Neurología	CC	Diagnóstico	Médica
2	2	3	NO	Neurología	CC	Diagnóstico	Médica
3	3	4	NO	Traumatología	CC	Tratamiento	Quirúrgica
4	1	1	SI	Cardiología	CC	Diagnóstico	Médica
5	4	3	NO	Cardiología	CC	Diagnóstico	Médica
6	1	2	NO	Cardiología	CC	Tratamiento	Médica
7	2	4	NO	Neumología	CC	Diagnóstico	Médica
8	1	1	SI	Digestivo	CC	Diagnóstico	Médica
9	1	4	NO	Ginecología	Test	Varios	Quirúrgica
10	4	2	NO	Hematología	CC	Varios	Médica
11	4	1	NO	Infecciosas	CC	Diagnóstico	Médica
12	3	3	SI	Digestivo	CC	Diagnóstico	Médica
13	4	4	SI	Pediatría	CC	Diagnóstico	Médica
14	3	1	NO	Digestivo	Test	Diagnóstico	Médica
15	3	2	NO	Oncología	CC	Diagnóstico	Médica
16	3	4	NO	Dermatología	CC	Diagnóstico	Quirúrgica
17	3	3	SI	Dermatología	CC	Varios	Quirúrgica
18	1	2	NO	Oncología	CC	Diagnóstico	Médica
19	3	4	NO	Oncología	CC	PC	Médica
20	2	2	SI	Oncología	CC	Diagnóstico	Médica
21	2	2	SI	Digestivo	CC	Diagnóstico	Médica
22	3	4	NO	Oncología	CC	PC	Médica
23	1	3	NO	Oncología	CC	Diagnóstico	Médica
24	1	1	SI	Digestivo	CC	Diagnóstico	Médica
25	1	1	SI	Neumología	CC	PC	Médica
26	2	2	SI	Nefrología	Test	PC	Médica

27	4	2	NO	Bioquímica y genética	Test	Fisiopatología	Básicas
28	3	3	SI	Bioquímica y genética	Test	Fisiopatología	Básicas
29	4	3	NO	Fisiología	Negativa	Fisiopatología	Básicas
30	1	1	SI	Fisiología	Negativa	Fisiopatología	Básicas
31	3	3	SI	Fisiología	Test	Fisiopatología	Básicas
32	2	1	NO	Fisiología	Test	Fisiopatología	Básicas
33	4	2	NO	Inmunología	Negativa	Fisiopatología	Básicas
34	2	1	NO	Neumología	Test	Fisiopatología	Básicas
35	2	2	SI	Neumología	Test	Tratamiento	Básicas
36	4	4	SI	Inmunología	Test	Fisiopatología	Básicas
37	1	3	NO	Inmunología	Test	Fisiopatología	Básicas
38	4	4	SI	Inmunología	Negativa	Fisiopatología	Básicas
39	3	3	SI	Digestivo	CC	Tratamiento	Médica
40	3	4	NO	Digestivo	Negativa	Varios	Médica
41	1	1	SI	Bioquímica y genética	Test	Fisiopatología	Básicas
42	2	2	SI	Bioquímica y genética	Test	Fisiopatología	Básicas
43	1	1	SI	Preventiva	Test	Estadística	Preventiva
44	2	2	SI	Preventiva	Test	Estadística	Preventiva
45	2	1	NO	Preventiva	Test	Estadística	Preventiva
46	2	1	NO	Preventiva	Test	Estadística	Preventiva
47	2	4	NO	Preventiva	Test	Tratamiento	Preventiva
48	2	2	SI	Preventiva	Test	Estadística	Preventiva
49	4	4	SI	Preventiva	Test	Estadística	Preventiva
50	4	4	SI	Preventiva	Test	Tratamiento	Preventiva
51	4	4	SI	Preventiva	Test	Tratamiento	Preventiva
52	3	3	SI	Preventiva	Test	Estadística	Preventiva
53	3	2	NO	Psiquiatría	Test	Tratamiento	Preventiva
54	2	2	SI	Farmacología	CC	Tratamiento	Básicas
55	3	3	SI	Farmacología	Test	Fisiopatología	Básicas
56	3	1	NO	Farmacología	Test	Tratamiento	Básicas
57	4	4	SI	Farmacología	CC	Tratamiento	Básicas
58	4	3	NO	Maxilofacial y ORL	CC	Diagnóstico	Quirúrgica
59	2	2	SI	Maxilofacial y ORL	Test	Fisiopatología	Quirúrgica
60	3	3	SI	Dermatología	Test	Diagnóstico	Quirúrgica
61	4	2	NO	Dermatología	Negativa	Tratamiento	Quirúrgica

62	4	4	SI	Oftalmología	Test	Diagnóstico	Quirúrgica
63	3	3	SI	Oftalmología	Test	Fisiopatología	Quirúrgica
64	3	3	SI	Oftalmología	CC	Diagnóstico	Quirúrgica
65	2	2	SI	Oftalmología	CC	Diagnóstico	Quirúrgica
66	3	3	SI	Neurología	CC	Diagnóstico	Médica
67	2	2	SI	Maxilofacial y ORL	Test	Diagnóstico	Quirúrgica
68	2	1	NO	Ginecología	Negativa	Varios	Quirúrgica
69	4	4	SI	Ginecología	Negativa	Varios	Quirúrgica
70	1	4	NO	Ginecología	Test	Tratamiento	Quirúrgica
71	2	2	SI	Ginecología	Test	Tratamiento	Quirúrgica
72	4	4	SI	Ginecología	CC	Tratamiento	Quirúrgica
73	1	1	SI	Ginecología	CC	Tratamiento	Quirúrgica
74	2	2	SI	Urología	Test	Tratamiento	Quirúrgica
75	4	4	SI	Ginecología	CC	Tratamiento	Quirúrgica
76	3	2	NO	Ginecología	CC	Diagnóstico	Quirúrgica
77	2	3	NO	Pediatría	Test	Tratamiento	Médica
78	1	3	NO	Pediatría	Test	Diagnóstico	Médica
79	1	2	NO	Pediatría	CC	PC	Médica
80	1	1	SI	Pediatría	Negativa	Fisiopatología	Médica
81	3	1	NO	Psiquiatría	Test	Varios	Médica
82	2	3	NO	Digestivo	Test	Varios	Médica
83	4	4	SI	Pediatría	Test	Diagnóstico	Médica
84	2	1	NO	Pediatría	Test	Tratamiento	Médica
85	4	1	NO	Pediatría	CC	Tratamiento	Médica
86	1	3	NO	Pediatría	CC	Diagnóstico	Médica
87	4	4	SI	Psiquiatría	Test	Tratamiento	Médica
88	3	3	SI	Psiquiatría	Test	Tratamiento	Médica
89	3	4	NO	Psiquiatría	CC	Diagnóstico	Médica
90	4	4	SI	Psiquiatría	Test	Diagnóstico	Médica
91	2	1	NO	Psiquiatría	Test	Diagnóstico	Médica
92	2	2	SI	Psiquiatría	Test	Tratamiento	Médica
93	4	1	NO	Psiquiatría	Negativa	Tratamiento	Médica
94	4	4	SI	Psiquiatría	Test	Tratamiento	Médica
95	1	3	NO	Neurología	Test	Fisiopatología	Médica
96	2	3	NO	Neurología	Negativa	Diagnóstico	Médica
97	4	2	NO	Neurología	CC	Diagnóstico	Médica
98	2	1	NO	Neurología	Test	Diagnóstico	Médica
99	2	2	SI	Neurología	CC	Tratamiento	Médica
100	2	4	NO	Neurología	CC	PC	Médica
101	3	2	NO	Neurología	CC	Tratamiento	Médica

102	4	4	SI	Neurología	Test	Diagnóstico	Médica
103	4	4	SI	Neurología	CC	Diagnóstico	Médica
104	2	2	SI	Neurología	Test	Varios	Médica
105	1	1	SI	UCI	CC	Tratamiento	Médica
106	3	3	SI	UCI	CC	Tratamiento	Médica
107	4	4	SI	UCI	CC	Tratamiento	Médica
108	4	3	NO	Traumatología	CC	Diagnóstico	Quirúrgica
109	2	2	SI	Traumatología	CC	PC	Quirúrgica
110	1	1	SI	Traumatología	CC	Tratamiento	Quirúrgica
111	3	3	SI	Traumatología	CC	Diagnóstico	Quirúrgica
112	2	1	NO	Traumatología	Test	Varios	Quirúrgica
113	2	2	SI	Traumatología	CC	Diagnóstico	Quirúrgica
114	4	4	SI	Traumatología	Test	Tratamiento	Quirúrgica
115	1	1	SI	Reumatología	CC	Diagnóstico	Médica
116	3	1	NO	Reumatología	CC	Tratamiento	Médica
117	2	1	NO	Reumatología	CC	PC	Médica
118	2	4	NO	Reumatología	Test	Diagnóstico	Médica
119	4	2	NO	Cardiología	Test	Diagnóstico	Médica
120	4	2	NO	Cardiología	Test	Diagnóstico	Médica
121	3	3	SI	Cardiología	CC	Tratamiento	Médica
122	4	1	NO	Cardiología	Test	Varios	Médica
123	4	4	SI	Cardiología	CC	Tratamiento	Médica
124	1	2	NO	Cardiología	CC	Tratamiento	Médica
125	4	4	SI	Cardiología	CC	PC	Médica
126	4	3	NO	Neumología	CC	Tratamiento	Médica
127	1	4	NO	Neumología	CC	Tratamiento	Médica
128	4	4	SI	Neumología	CC	Diagnóstico	Médica
129	1	4	NO	Neumología	CC	Tratamiento	Médica
130	3	3	SI	Digestivo	CC	Diagnóstico	Médica
131	3	2	NO	Digestivo	CC	Tratamiento	Médica
132	3	3	SI	Digestivo	Test	Diagnóstico	Médica
133	1	4	NO	Digestivo	CC	Diagnóstico	Médica
134	4	3	NO	Digestivo	Negativa	Varios	Médica
135	3	2	NO	Reumatología	CC	PC	Médica
136	2	2	SI	Nefrología	CC	PC	Médica
137	4	3	NO	Nefrología	CC	Diagnóstico	Médica
138	4	3	NO	Urología	CC	Tratamiento	Quirúrgica
139	2	2	SI	Nefrología	CC	Diagnóstico	Médica
140	4	2	NO	Endocrinología	CC	Diagnóstico	Médica
141	1	4	NO	Urología	Negativa	Tratamiento	Quirúrgica
142	1	1	SI	Urología	Test	Varios	Quirúrgica

143	3	3	SI	Oncología	Test	Tratamiento	Médica
144	3	3	SI	Oncología	Negativa	Tratamiento	Médica
145	2	2	SI	Oncología	Test	Fisiopatología	Médica
146	1	1	SI	Oncología	Test	Tratamiento	Médica
147	2	3	NO	Hematología	CC	Tratamiento	Médica
148	1	1	SI	Hematología	CC	Varios	Médica
149	3	3	SI	Hematología	CC	Diagnóstico	Médica
150	2	2	SI	Hematología	CC	PC	Médica
151	3	1	NO	Geriatría	Test	Fisiopatología	Médica
152	4	2	NO	Geriatría	Test	PC	Médica
153	4	3	NO	Geriatría	Test	Fisiopatología	Médica
154	1	4	NO	Geriatría	Test	Tratamiento	Médica
155	2	2	SI	Geriatría	Negativa	Diagnóstico	Médica
156	3	3	SI	Geriatría	CC	Diagnóstico	Médica
157	2	1	NO	Endocrinología	CC	PC	Médica
158	2	2	SI	Endocrinología	Negativa	Varios	Médica
159	3	3	SI	Endocrinología	Test	Tratamiento	Médica
160	1	1	SI	Endocrinología	CC	Diagnóstico	Médica
161	2	2	SI	Endocrinología	CC	Diagnóstico	Médica
162	3	4	NO	Endocrinología	CC	Varios	Médica
163	4	1	NO	Endocrinología	Negativa	Tratamiento	Médica
164	1	1	SI	Infecciosas	CC	Tratamiento	Médica
165	3	1	NO	Infecciosas	CC	Tratamiento	Médica
166	1	1	SI	Infecciosas	CC	Tratamiento	Médica
167	1	1	SI	Infecciosas	CC	Tratamiento	Médica
168	2	4	NO	Infecciosas	Test	Varios	Médica
169	1	1	SI	Infecciosas	CC	Tratamiento	Médica
170	2	2	SI	Infecciosas	CC	Tratamiento	Médica
171	4	1	NO	Reumatología	CC	Tratamiento	Médica
172	4	3	NO	Reumatología	Test	PC	Médica
173	2	2	SI	Reumatología	CC	Diagnóstico	Médica
174	4	1	NO	Reumatología	CC	Diagnóstico	Médica
175	2	4	NO	Reumatología	CC	Fisiopatología	Médica
176	2	4	NO	Reumatología	Test	Fisiopatología	Médica
177	1	3	NO	Reumatología	CC	Diagnóstico	Médica
178	4	4	SI	Endocrinología	CC	Tratamiento	Médica
179	1	1	SI	Endocrinología	CC	Diagnóstico	Médica
180	4	4	SI	Medicina Legal y Ética	Test	Ética y Legal	Ética
181	3	3	SI	Medicina Legal y Ética	CC	Ética y Legal	Ética

182	4	1	NO	Medicina Legal y Ética	CC	Ética y Legal	Ética
183	3	3	SI	Medicina Legal y Ética	CC	Ética y Legal	Ética
184	2	2	SI	Psiquiatría	CC	Tratamiento	Médica
185	4	1	NO	Digestivo	CC	Tratamiento	Médica
186	1	4	NO	Medicina Legal y Ética	CC	Ética y Legal	Ética
187	3	4	NO	Medicina Legal y Ética	Test	Ética y Legal	Ética
188	4	3	NO	Dermatología	Test	Varios	Quirúrgica
189	3	2	NO	Preventiva	CC	Tratamiento	Preventiva
190	3	3	SI	Endocrinología	CC	Diagnóstico	Médica
191	3	2	NO	Oncología	CC	Diagnóstico	Médica
192	4	1	NO	Cardiología	CC	Tratamiento	Médica
193	2	3	NO	Endocrinología	CC	Tratamiento	Médica
194	3	3	SI	Endocrinología	Test	Tratamiento	Médica
195	4	4	SI	Farmacología	CC	Tratamiento	Básicas
196	4	4	SI	Urgencias	Test	Tratamiento	Médica
197	2	2	SI	Urgencias	Test	Tratamiento	Médica
198	3	1	NO	Urgencias	CC	Diagnóstico	Médica
199	3	3	SI	Urgencias	CC	Diagnóstico	Médica
200	4	1	NO	Urgencias	CC	Tratamiento	Médica
201	4	3	NO	Oncología	Test	Diagnóstico	Médica
202	2	2	SI	Nefrología	CC	Diagnóstico	Médica
203	3	4	NO	Endocrinología	CC	Diagnóstico	Médica
204	4	1	NO	Oncología	Test	Diagnóstico	Médica
205	3	3	SI	Reumatología	CC	Diagnóstico	Médica
206	1	2	NO	Reumatología	CC	Diagnóstico	Médica
207	4	1	NO	Neumología	Test	Varios	Médica
208	3	3	SI	Neumología	CC	Diagnóstico	Médica
209	1	1	SI	Neumología	CC	Diagnóstico	Médica
210	4	4	SI	Cardiología	CC	Tratamiento	Médica