Revista SOCAMPAR

ORIGINAL

CAPACIDAD DE CHATGPT EN LA RESOLUCCIÓN CORRECTA DE LAS PREGUNTAS DE NEUMOLOGÍA DEL EXAMEN MIR

CHATGPT'S ABILITY IN THE CORRECT RESOLUTION OF THE PNEUMOLOGY QUESTIONS OF THE MIR EXAM

Autores: Diego Meléndez [1], José Luis Izquierdo [1,2]

1. *Departamento de Medicina y Especialidades Médicas, Universidad de Alcalá, Madrid*
2. *Servicio de neumología. Hospital Universitario de Guadalajara, Guadalajara*

Resumen:

Introducción y objetivos: La aparición de modelos de IA basados en el *natural language processing* como ChatGPT ha despertado interés acerca de las potenciales aplicaciones de un *chatbot* en la medicina. Para comprobar sus conocimientos médicos se les ha sometido a exámenes estandarizados. Este trabajo se propone evaluar la capacidad de ChatGPT para aprobar las preguntas de neumología del examen MIR, estudiando las diferencias entre ChatGPT-3.5 y ChatGPT-4, el impacto de distintos *prompts*, la dificultad de las preguntas y el año de convocatoria.

Material y métodos: Se introdujeron en los modelos 96 preguntas de neumología extraídas de las últimas 10 convocatorias MIR, acompañadas de 4 *prompts* con los que el modelo debe adoptar un rol cada vez más especializado. Para las comparaciones se emplearon test chi-cuadrado y se consideraron significativos valores de $p < 0.05$.

Resultados: ChatGPT-3.5 acertó el 57,29% de las preguntas y ChatGPT-4 el 85,38%. Al comparar el rendimiento de ChatGPT-4 con ChatGPT-3.5 se obtuvo un OR (IC95%) de 3,995 (2,036 a 8,106). En el análisis del impacto del *prompt-engineering*, de la dificultad, y del año de convocatoria no se obtuvieron resultados estadísticamente significativos.

Conclusiones: Extrapolando las preguntas de neumología al total del examen ChatGPT-3.5 y ChatGPT-4 aprobarían el examen MIR. ChatGPT-4 se situaría entre el 10% de mejores aspirantes. ChatGPT-4 fue superior a ChatGPT-3.5. Las técnicas de *prompt-engineering*, la dificultad de las preguntas y el año de convocatoria no tuvieron impacto significativo en el rendimiento de los modelos**.**

Palabras clave: ChatGPT, inteligencia artificial en medicina, examen MIR, prompt-engineering, chat-bot, educación médica

Resume:

Introduction: The emergence of AI models based on natural language processing such as ChatGPT has sparked interest in the potential applications of a chatbot in medicine. To verify their medical knowledge, they have been subjected to standardized tests. This work aims to evaluate the ability of ChatGPT to pass the pulmonology questions of the MIR exam, studying the differences between ChatGPT-3.5 and ChatGPT-4, the impact of different prompts, the difficulty of the questions and the year of the call.

Material and methods: 96 pulmonology questions extracted from the last 10 MIR calls were introduced into the models, accompanied by 4 prompts with which the model must adopt an increasingly specialized role. For comparisons, chi-square tests were used and values of $p < 0.05$ were considered significant.

Results: ChatGPT-3.5 got 57.29% of the questions correct and ChatGPT-4 85.38%. When comparing the performance of ChatGPT-4 with ChatGPT-3.5, an OR (95%CI) of 3.995 (2.036 to 8.106) was obtained. In the analysis of the impact of prompt-engineering, difficulty, and year of call, no statistically significant results were obtained..

Conclusions: Extrapolating the pulmonology questions to the total exam ChatGPT-3.5 and ChatGPT-4 would pass the MIR exam. ChatGPT-4 would be among the top 10% of applicants. ChatGPT-4 was superior to ChatGPT-3.5. The prompt-engineering techniques, the difficulty of the questions and the year of the call did not have a significant impact on the performance of the models.

Keywords: ChatGPT, artificial intelligence in medicine, MIR exam, prompt-engineering, chat-bot, medical education.

Introducción:

La medicina no es una disciplina ajena a las potenciales aplicaciones de la inteligencia artificial (IA), existiendo modelos diseñados para su aplicación en este campo desde los años 70 del siglo pasado (1). Con el comienzo del nuevo milenio, novedades como la *computer vision*, el *deep learning* y el *natural language processing* han acelerado el desarrollo de esta tecnología. Disponemos de programas para la predicción de riesgo

cardiovascular o de la mortalidad de diversos procesos, la filiación de lesiones dermatológicas (2) o el "diagnóstico asistido por ordenador" en las endoscopias digestivas (3). En repetidas ocasiones estos han demostrado ser iguales o superiores a los criterios dictados por las guías de práctica clínica o al rendimiento de los facultativos, y algunos ya han sido aprobados por organismos reguladores como la FDA (*Food and Drug Administration*) para su uso en la práctica clínica (4).

ChatGPT es una de las últimas revoluciones en IA que destaca por su novedad, popularidad y controversia desde su lanzamiento en noviembre de 2022. Es un *chatbot* diseñado por OpenAI al que se puede acceder libremente y con el que cualquier usuario puede interactuar. La accesibilidad de este modelo, su creciente número de usuarios y la aparente buena calidad de sus respuestas y naturalidad del lenguaje han despertado interés por sus posibles aplicaciones en medicina.

En EE.UU. y en Asia han tratado de evaluar los conocimientos teóricos de esta plataforma sometiéndola a los mismos exámenes que los estudiantes de medicina o los médicos en formación especialista han de superar para recibir sus respectivos títulos. Para ello se han empleado tanto los exámenes realizados en los últimos años como preguntas extraídas de bancos que los estudiantes emplean en su preparación. Hasta el momento los resultados han sido relativamente uniformes: la última versión de ChatGPT (GPT-4) es capaz de superar estas pruebas, tal y como se recoge en la Tabla 1.

ESTUDIO	PRUEBA	FECHA	RESULTADOS (% DE ACIERTOS)	
			CHATGPT-3.5	CHATGPT-4
(5)	USLME	12/22	44%/42%/64,4%/57,8%*	-
(6)	CNMLE	02/23	47%/45,8%/36,5%**	-
(7)	MIR 2023	02/23	51,4%	-
(8)	MIR 2024	02/24	-	82,4%
(9)	CNMLE	02/23 (GPT-3.5)	56%	84%
	NEEPM	05/23 (GPT-4)	62%	82%
(10)	JMLE	03/23	50,8%	79,9%
(11)	PNLME	03/23 (GPT-3.5) 05/23 (GPT-4)	77%	86%
(12)	PES	06/23	-	67,1%

Tabla 1: *Resultados de las versiones 3.5 y 4 de ChatGPT en algunas de las pruebas de evaluación a las que se someten los estudiantes de medicina de diversos países. *= los resultados corresponden a los 4 sets de preguntas que se emplearon. **=los resultados corresponden a los exámenes de 2020, 2021 y 2022. USMLE: United States Medical Licensing Examination. CNMLE: China National Medical Licensing Ex- amination. NEEPM: China National Entrance Examination for Postgraduate Clinical Medicine. JMLE: Japanese Medical Licensing Examination. PNLME: Peruvian National Licensing Medical Examination. PES: National Specialization Examination (Polonia)*

Dada la escasez de estudios de este tipo en lengua castellana y en el ámbito europeo, el objetivo de este trabajo es evaluar la capacidad de ChatGPT para responder y aprobar el examen MIR según los criterios de corrección publicados en el BOE (Boletín Oficial del Estado) para la última convocatoria. Para ello partimos de la hipótesis de que los modelos de IA basados en el *natural language processing* son capaces de superar exámenes médicos estandarizados.

Además, contamos con una serie de objetivos específicos que nos permitirán ahondar en la capacidad de estas herramientas de enfrentarse a estas pruebas: 1)Comparar la precisión en las respuestas del modelo GPT-4 con su predecesor inmediato, GPT-3.5; 2) cuantificar el efecto que tienen distintos *prompts* (instrucciones que se dan al modelo) en la tasa de acierto; 3) valorar el impacto de la dificultad de las preguntas en la capacidad de los modelos de responderlas correctamente; 4) comparar la puntuación obtenida en el examen de la convocatoria más reciente (el examen celebrado en 2024) con la obtenida en años anteriores,

MATERIAL Y MÉTODOS

Modelos empleados.

Se han empleado 2 versiones del *chatbot* ChatGPT, desarrollado por OpenAI. ChatGPT-3.5 fue lanzado en noviembre de 2022 y el más avanzado ChatGPT-4 en marzo de 2023.

Esta última versión ha mostrado un rendimiento superior en pruebas estandarizadas de diversas disciplinas y mejores resultados en idiomas distintos del inglés (13).

Muestra de preguntas

Se ha seleccionado una muestra de 96 preguntas tipo test extraídas de los exámenes de las últimas 10 convocatorias (2015-2024), correspondientes a la especialidad de neumología. Se han excluido las preguntas que fueron impugnadas y las relacionadas con imagen, ya que ChatGPT-3.5 no es capaz de procesarlas. La dificultad ha sido asignada de acuerdo con el criterio de una de las academias que preparan a los aspirantes, de un modo similar a estudios previos (10).

Prompt-engineering

El *prompt-engineering* es el proceso de diseño y optimización de *prompts* que tiene como objetivo mejorar la calidad y precisión de las respuestas de ChatGPT. Otros estudios han empleado diversos enfoques en la formulación de las preguntas. Algunos personalizan los *prompts* para cada pregunta con la intención evitar errores sistemáticos, comprometiendo la reproducibilidad del estudio (14). Otros han experimentado con diversos modelos de *prompt* cuya eficacia ya ha sido probada, por ejemplo, el *chain of thought* (se pide al modelo que razone antes de elegir una respuesta) o el *rank-order* (se le pide que ordene las opciones proporcionadas en función de su adecuación) (15).

Una de las hipótesis de las que partimos en este trabajo es que mediante un cuidadoso *prompt-engineering* podemos ser capaces de mejorar la proporción de aciertos de ambos modelos. Para ello hemos empleado unos *prompts* estandarizados (recogidos en la Tabla 2) basándonos en las recomendaciones proporcionadas por OpenAI (16). Se trata de instrucciones simples y claras de tipo *zero-shot* en los que se pide al modelo que adopte distintos roles de los que se espera un conocimiento cada vez mayor de la materia.

Nº	ROL	PROMPT
PROMPT 1	Sin rol	"Selecciona la respuesta correcta y proporciona una explicación."
PROMPT 2	Estudiante	"Eres un estudiante de medicina. Selecciona la respuesta correcta y proporciona una explicación."
PROMPT 3	Médico	"Eres un médico. Selecciona la respuesta correcta y proporciona una explicación."
PROMPT 4	Especialista	"Eres un médico especialista en neumología. Selecciona la respuesta correcta y proporciona una explicación."

Tabla 2: *Prompts empleados para introducir las distintas preguntas en cada uno de los modelos.*

Al carecer de rol, el PROMPT1 actúa como control, por lo que los resultados obtenidos con este son los que se han empleado para realizar las comparaciones entre modelos, grados de dificultad y años de examen.

Sesgo de memoria

ChatGPT almacena y recuerda las conversaciones que el usuario mantiene con él, dando lugar al concepto de sesgo de memoria. Según este, conforme se vayan introduciendo preguntas al modelo este las irá respondiendo cada vez mejor, lo que es un punto controvertido a la hora de realizar este tipo de estudio. Algunos autores ven en él un paralelo a cómo funciona la mente humana y por lo tanto tratan de potenciarlo (17), mientras que otros buscan minimizar su efecto (15,18). En este estudio se ha desactivado la función de historial y entrenamiento a la hora de introducir las preguntas. En conclusión, para obtener las respuestas de ChatGPT se ha introducido cada pregunta 4 veces en cada modelo, en cada ocasión acompañada de un *prompt* distinto y evitando que de una a otra se retuviese información previa.

Concepto de contaminación

La posibilidad de que las preguntas de los exámenes hayan formado parte de las bases de datos con las que los modelos han sido entrenados implicaría que estos las acertarían con mayor facilidad (19). Este concepto recibe el nombre de contaminación y nuestro último objetivo, comparar la puntuación obtenida en el examen celebrado en 2024 con la de los años anteriores, persigue estudiar su presencia.

Para tratar de delimitar el impacto de este efecto se comparará el examen más reciente con los previos, que llevan años disponibles en la red y por lo tanto son más susceptibles de haber sido utilizados en el periodo de entrenamiento.

Análisis estadístico

Para la comparación entre modelos, *prompts*, dificultad y año de las preguntas se emplearon test chi-cuadrado, usando la corrección de Yates ante valores <5. Cuando fue pertinente se calculó el odds ratio (OR) con un IC95%, empleando la corrección de Laplace ante la presencia de valores nulos. Se consideraron significativos valores-p <0.05. Los datos fueron procesados a través del programa OpenEpi.

.

Resultados:

Tras introducir las 96 preguntas al modelo encontramos que ChatGPT-4 acierta 81 (84,38%) y ChatGPT-3.5 55 (57,29%), fallando en ambos casos las restantes, ya que en ningún caso se negó a seleccionar una de las opciones proporcionadas.

La Tabla 3 recoge el número de aciertos y fallos por cada modelo en las preguntas correspondientes a cada una de las últimas 10 convocatorias, así como los totales. Si analizamos el total obtenemos un odds ratio (OR) de 3,995 (IC95%: 2,04- 8,11) a favor de ChatGPT-4..

AÑO	N	CHATGPT-3.5		CHATGPT-4		OR	IC95%	p-valor
		Aciertos (%)	Fallos (%)	Aciertos (%)	Fallos (%)			
2015	14	7 (50%)	7 (50%)	13 (93%)	1 (7%)	11.82	1.447, 315.3	0.0364
2016	12	8 (67%)	4 (33%)	8 (67%)	4 (33%)	1	0.1675, 5.969	0.6650
2017	11	6 (55%)	5 (45%)	9 (82%)	2 (18%)	3.524	0.5117, 33.76	0.3599
2018	10	6 (60%)	4 (40%)	8 (80%)	2 (20%)	2.537	0.3334, 25.59	0.6256
2019	10	5 (50%)	5 (50%)	10 (100%)	0	13.93*	1.181, 512.8	0.0699
2020	6	3 (50%)	3 (50%)	6 (100%)	0	8.204*	0.4897, 360.7	0.3029
2021	6	5 (84%)	1 (16%)	6 (100%)	0	1.932*	0.0346, 107.8	0.7015
2022	10	6 (60%)	4 (40%)	8 (80%)	2 (20%)	2.537	0.3334, 25.59	0.6256
2023	5	3 (60%)	2 (40%)	3 (60%)	2 (40%)	1	0.0635, 15.74	0.5186
2024	12	6 (50%)	6 (50%)	10 (84%)	2 (16%)	4.657	0.7246, 42.97	0.1942
TOTAL	96	55 (57,23%)	41 (42,77%)	81 (84,37%)	15 (15,63%)	3.995	2.036, 8.106	<.0000

Tabla 3: *Aciertos y fallos de cada modelo en las preguntas de las 10 últimas convocatorias (2015-2024) y en el conjunto. * indica los casos en los que para calcular el OR fue necesario aplicar una corrección de +0,5 en las casillas con un valor nulo (corrección* de Laplace). N: número de preguntas. OR: Odds ratio. IC95%: intervalo de confianza al 95%

En la Tabla 4 se puede apreciar cómo la repetición de las preguntas con distintos *prompts* en los que se pedía al modelo adoptar un rol cada vez más especializado arrojaron resultados prácticamente idénticos a cuando

simplemente se le pedía la respuesta correcta, tanto para ChatGPT-4 como para ChatGPT-3.5.

PROMPT	CHATGPT-3.5		CHATGPT-4	
	ACIERTOS (%)	FALLOS (%)	ACIERTOS (%)	FALLOS (%)
PROMPT1	55 (57,23%)	41 (42,27%)	81 (84,37%)	15 (15,63%)
PROMPT2	54 (56,25%)	42 (43,75%)	80 (83,33%)	16 (16,66%)
PROMPT3	54 (56,25%)	42 (43,75%)	80 (83,33%)	16 (16,66%)
PROMPT4	54 (56,25%)	42 (43,75%)	81 (84,37%)	15 (15,63%)

Tabla 4: *Aciertos y fallos de cada modelo empleando los 4 prompts presentados en la Tabla 2*

Del total de 96 preguntas, 37 correspondían a la categoría de fáciles, 30 a la de dificultad media y 29 a la de difíciles. Tal y como refleja la Tabla 5, tanto para ChatGPT-4 como para ChatGPT-3.5 la proporción de respuestas correctas en cada subgrupo es similar a la del conjunto total. La realización de dos test chi-cuadrado, uno para cada modelo, arroja unos valores p de 0,9289 (para ChatGPT-3.5) y 0,5381 (para ChatGPT-4). Por lo tanto, en nuestra serie el grado de dificultad no tiene un impacto estadísticamente significativo en la tasa de aciertos para ninguno de los modelos.

Comparando los resultados obtenidos en las preguntas correspondientes a la última convocatoria frente al conjunto de las previas (Tabla 6), se aprecia una disminución en el porcentaje de aciertos para ambos modelos.

DIFICULTAD	N	CHATGPT-3.5		CHATGPT-4	
		ACIERTOS (%)	FALLOS (%)	ACIERTOS (%)	FALLOS (%)
DIFICULTAD 1	37	21 (56,75%)	16 (43,24%)	33 (89,19%)	4 (10,81%)
DIFICULTAD 2	30	18 (60%)	12 (40%)	25 (83,33%)	5 (16.67%)
DIFICULTAD 3	29	16 (55,17%)	13 (44,82%)	23 (79,31%)	6 (20,69%)
TOTAL	96	55 (57,23%)	41 (42,77%)	81 (84,37%)	15 (15,63%)

Tabla 5: *Aciertos y fallos de cada modelo en cada estrato de dificultad. DIFICULTAD 1 corresponde a las preguntas calificadas como fáciles, DIFICULTAD 2 a las de dificultad media y DIFICULTAD 3 a las difíciles. N: número de preguntas*

AÑO	N	CHATGPT-3.5		CHATGPT-4	
		ACIERTOS (%)	FALLOS (%)	ACIERTOS (%)	FALLOS (%)
2024	12	6 (50%)	6 (50%)	10 (83%)	2 (17%)
2015-2023	84	49 (58%)	35 (42%)	71 (87%)	13 (13%)
TOTAL	96	55 (57,23%)	41 (42,77%)	81 (84,37%)	15 (15,63%)

Tabla 6: *Aciertos y fallos de cada modelo en la convocatoria de 2024 y en el conjunto de las previas. N= número de preguntas*

Sin embargo, el test estadístico, arroja unos OR de 0,72 (IC95%: 0,20 a 2,5) para ChatGPT-3.5 y de 0,92 (IC95%: 0,19 a 6,77) para ChatGPT-4 por lo que no podemos afirmar que introducir preguntas susceptibles de haber formado parte de las bases de datos de entrenamiento tenga impacto en el rendimiento de los modelos.

Discusión:

De acuerdo con los criterios de corrección publicados para la última convocatoria, para superar la prueba MIR se ha de obtener una puntuación superior al 25% de la media del 10% de las mejores puntuaciones. Para calcular la puntuación se multiplica el total de aciertos por 3 y se le resta el número de fallos (20). Si extrapolamos los resultados obtenidos con la preguntas de neumología al conjunto del examen, ChatGPT-4 habría obtenido 475 puntos y ChatGPT-3.5, 258. En ambos casos la puntuación es superior a los 115 puntos que en la convocatoria de 2024 supusieron la nota de corte, por lo que ambos modelos podrían haber participado en la adjudicación de plazas. Sin embargo, podemos contextualizar mejor los resultados si tenemos en cuenta que la puntuación de ChatGPT-4 supera a la media del 10% de mejores puntuaciones de la última convocatoria, que fueron 461,3 puntos.

Estos resultados están en consonancia con los estudios que ya han evaluado la capacidad de ChatGPT de resolver el examen MIR, realizados con las pruebas celebradas en 2023 y 2024. A diferencia de nuestro trabajo, en el que se ha empleado una muestra de preguntas que abarca varias convocatorias, en esos casos se presentó a la IA el examen completo. Para el examen celebrado en 2023 (7), ChatGPT-3.5 respondió correctamente el 54,8% de las preguntas (excluyendo las relacionadas con imagen), un porcentaje similar al obtenido en nuestro estudio (57,23%). El examen celebrado en 2024 (8) se resolvió con ChatGPT- 4, el cuál contestó correctamente el

82,38% de las preguntas, de nuevo un porcentaje comparable al observado en nuestra investigación (84,37%). Para una comparación con exámenes extranjeros nos remitimos a la Tabla 1, en la que podemos observar de nuevo similitudes entre los porcentajes de acierto de cada modelo publicados en diversos países y los obtenidos en este estudio. Por lo tanto, podemos concluir que nuestros hallazgos a este respecto se encuentran en línea con la literatura publicada hasta el momento.

ChatGPT-4 es superior a ChatGPT-3.5

A la hora de comparar el rendimiento de los modelos entre sí hemos hallado una marcada superioridad de ChatGPT-4 frente a su predecesor, ChatGPT-3.5. Este resultado viene a confirmar los ya publicados hasta la fecha en diversos artículos (9,10,18,19,21).

El *prompt-engineering* no logra mejorar los resultados

Asignar distintos roles presuponiendo mayores niveles de conocimiento no influyó en el número de respuestas correctas. Otros estudios (14,19,22) han trabajado con estrategias distintas a la hora de elaborar los *prompts*, sin lograr tampoco diferencias significativas. La experiencia disponible parece indicar que el *prompt-engineering* tiene un impacto bajo o despreciable cuando la tarea encomendada al modelo es responder una pregunta tipo test, y que sin embargo cobra una mayor importancia en preguntas cuya resolución requiere un componente de desarrollo (22).

La dificultad no influye en los resultados

En cuanto al análisis de aciertos en función de la dificultad de las preguntas, las diferencias halladas (más marcadas para ChatGPT-4) no han resultado estadísticamente significativas. No obstante, este análisis plantea 2 hándicaps ausentes en el resto: 1) en primer lugar, la muestra queda dividida en 3 subgrupos de menor tamaño, por lo que estas diferencias no significativas observadas con ChatGPT-4 puede pueden estar condicionadas por un menor tamaño muestral; 2) en segundo lugar, la adjudicación del grado de dificultad se lleva a cabo de acuerdo con la opinión de los expertos que trabajan para una academia, que a su vez les proporciona unas pautas para graduar la dificultad. Sin embargo, este sistema introduce un elemento de subjetividad que otros estudios han evitado al basar la dificultad en la proporción de alumnos que respondieron correctamente cada pregunta (10,15,21). En este sentido nuestros resultados difieren respecto a los artículos publicados, los cuales muestran una disminución estadísticamente significativa en la proporción de aciertos conforme aumenta la dificultad (5,10,15,21).

No se ha detectado evidencia de contaminación

Por último, al comparar el rendimiento de los modelos en las preguntas correspondientes a la convocatoria más reciente frente al resto de las estudiadas se observa una disminución del porcentaje de aciertos tanto para ChatGPT-4 (4 puntos porcentuales de diferencia) como para ChatGPT-3.5 (8 puntos), aunque en ambos casos son diferencias estadísticamente no significativas. Consideramos que se trata de un fenómeno que merece la pena continuar estudiando, puesto que no hay un consenso claro acerca de su presencia e importancia, con artículos publicados que llegan a conclusiones contradictorias (6,19).

Limitaciones y posibles líneas de investigación

El presente trabajo incluye innovaciones como la inclusión de preguntas de convocatorias pasadas, un ejercicio que hasta ahora no se había llevado a cabo para el examen MIR. También hemos querido poner el foco en la posibilidad de mejorar el rendimiento mediante la asignación de roles, una técnica de *prompt-engineering* que, si bien recomendada por OpenAI, no se ha empleado en estudios similares.

No obstante, nuestro trabajo no está exento de limitaciones. Cabe destacar la posibilidad de someter a examen la capacidad de ChatGPT de responder a preguntas relacionadas con una imagen. Estas suponen hasta un 12,5% del examen MIR, reflejan una actividad habitual en la práctica clínica y añaden una capa adicional de complejidad a las preguntas. Si bien ChatGPT-3.5 no es capaz de procesarlas, ChatGPT-4 sí lo es. Futuros trabajos podrían empezar a proporcionar preguntas acompañadas de su imagen al último modelo o evaluar la calidad con la que este es capaz de describirlas o informarlas. Debemos mencionar que algunos estudios han descrito que ChatGPT es capaz de acertar buena parte de las preguntas relacionadas con imagen pese a no disponer de ella (19,23).

Otro aspecto interesante es estudiar las características de las preguntas que llevan a los modelos a fallarlas. Se han estudiado multitud de parámetros que podrían influir en que ChatGPT acierte o falle una pregunta, tales como el idioma (9), la complejidad del enunciado (15) o la capacidad del modelo de recurrir a información ausente en este (5). En el caso concreto del examen MIR parece que factores como que el enunciado pida identificar la opción incorrecta o que la pregunta reúna conceptos comunes a varias especialidades disminuyen la tasa de aciertos, si bien estos hallazgos no se han visto respaldados por un análisis estadístico (7).

Como conclusión, herramientas de inteligencia artificial permiten responder correctamente las preguntas MIR de neumología posicionando a ChatGPT-4 entre el 10% de mejores aspirantes. ChatGPT-4 fue superior a ChatGPT-3.5. Las técnicas de *prompt-engineering*, la dificultad de las preguntas y el año de convocatoria no tuvieron impacto significativo en el rendimiento de los modelos.

Bibliografía:

1. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc. 2020 October;92(4):807-812.
2. Liopyris K, Gregoriou S, Dias J, Stratigos AJ. Artificial Intelligence in Dermatology: Challenges and Perspectives. Dermatol Ther (Heidelb.). 2022;12(12):2637-2651.
3. Houwen BBSL, Hazewinkel Y, Giotis I, Vleugels JLA, Mostafavi NS, Van Putten P, et al. Computer-aided diagnosis for optical diagnosis of diminutive colorectal polyps including sessile serrated lesions: a real-time comparison with screening endoscopists. Endoscopy. 2023;55(8):756-765.
4. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA- approved medical devices and algorithms: an online database. NPJ Digit Med. 2020;3(1):1-8.
5. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9.
6. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT Performs on the Chinese National Medical Licensing Examination. J Med Syst. 2023 December;47(1).
7. Carrasco JP, García E, Sánchez DA, Porter E, De La Puente L, 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 Edu Med. 2023 February;4(1).
8. Cerame A, Juaneda J, Estrella-Porter P, Puente Ldl, Navarro J, García E, et al. ¿Es capaz GPT-4 de aprobar el MIR 2023? Comparativa entre GPT-4 y ChatGPT-3 en los exámenes MIR 2022 y 2023. Rev Esp Edu Med. 2024 February;5(2).
9. Wang H, Wu WZ, 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 September;177.
10. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. JMIR Med Educ. 2023;9.
11. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, De la Cruz-Galán JP, Gutiérrez-Arratia J, Quiroga Torres BG, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: Cross-Sectional Study. JMIR Med Educ. 2023 September;9.
12. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Reshaping medical education: Performance of ChatGPT on a PES medical examination. Cardiol J. 2023 October.
13. Koubaa, A. GPT-4 vs. GPT-3.5: A Concise Showdown. Preprints 2023, 2023030422..
14. Kung TH, Cheatham M, Medenilla A, Sillos C, Leon LD, Elepaño C, et al.Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2).
15. 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. Ophthalmol Sci. 2023;3(4).
16. OpenAI. Prompt engineering. 2024; Available at: https://platform.openai.com/docs/guides/prompt-engineering. Accessed Apr 25, 2024.
17. Fuentes-Martín Á, Cilleruelo-Ramos Á, Segura-Méndez B, Mayol J. Can an Artificial Intelligence Model Pass an Examination for Medical Specialists? Arch Bronconeumol. 2023;59(8):534-536.
18. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. J Am Acad Orthop Surg. 2023;31(23):1173-1179.
19. Nori H, King N, Mckinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. arXiv. 2023 March.
20. Ministerio de Sanidad. Orden SND/990/2023, de 17 de agosto. 2023 Aug 24,;II. Autoridades y personal - B. Oposiciones y concursos.
21. Jiao C, Edupuganti NR, Patel PA, Bui T, Sheth V. Evaluating the Artificial Intelligence Performance Growth in Ophthalmic Knowledge. Cureus. 2023;15(9)
22. Choi JH, Hickman KE, Monahan A, Schwarcz DB. ChatGPT Goes to Law School.Journal of Legal Education. 2023 January
23. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. Radiology. 2023;307(5).