ChatGPT-4: su desempeño en un examen final de la carrera de médico especialista en oftalmología de la Universidad de Buenos Aires

Roberto Borrone

Cátedra de Oftalmología de la Facultad de Medicina de la Universidad de Buenos Aires (UBA).

Recibido: 18 de enero de 2024.
Aprobado: 9 de febrero de 2024.

Dirección del autor
Dr. Roberto Borrone
Coronel Díaz 2333, piso 2 B
(1425) Buenos Aires, Argentina
rborrone@intramed.net

Oftalmol Clin Exp (ISSNe 1851-2658)
2024; 17(1): e41-e45.

Agradecimientos
A Eugenia Piñeiro, ingeniera informática e investigadora en Inteligencia Artificial (ITBA); a Daniela Rivarola Meilán, analista de datos; y al Dr. Alan Grunberg.

Conflicto de interés
El autor declara no tener ningún conflicto de interés.

Resumen

Objetivos: Evaluar el desempeño del chatGPT-4 en un examen final de la Carrera de Médico Especialista Universitario en Oftalmología de la Universidad de Buenos Aires y compararlo con el desempeño de los alumnos ante el mismo examen y con el ChatGPT-3.5.
Material y métodos: Estudio observacional, retrospectivo y analítico. Se comparó el desempeño de 7 médicos en un examen final de posgrado de 50 preguntas con 4 opciones de respuesta rendido el 8 de septiembre de 2023 con el desempeño ante el mismo examen del ChatGPT versiones 3.5 y 4.
Resultados: La mediana de las respuestas correctas de los 7 alumnos fue 39 (rango 33-45) lo que representa una exactitud del 78%. El tiempo promedio para completar el examen fue de 75 minutos. El ChatGPT 3.5 respondió correctamente 31 preguntas (31/50) logrando una exactitud del 62%. El Chat GPT 4 respondió correctamente 40 preguntas (40/50) logrando una exactitud del 80% y completó el examen en 73.49 segundos.
Conclusiones: La versión ChatGPT-4 logró un desempeño superior a la mediana de los alumnos utilizando un tiempo 61 veces inferior. El ChatGPT-4 logró una exactitud superior a la versión 3.5. La calificación obtenida por las dos versiones del ChatGPT permite aprobar el examen dado que el umbral para lograrlo es de 30 respuestas correctas.
Palabras clave: inteligencia artificial, ChatGPT-4, educación médica, oftalmología.

ChatGPT-4: its performance in a final exam of the career of a medical specialist in ophthalmology at the University of Buenos Aires

Abstract

Objectives: To evaluate the performance of the GPTChat-4 in a final exam of the University Medical Specialist in Ophthalmology degree at the University of Buenos Aires and compare it with the performance of the students and the ChatGPT-3.5.
Materials and methods: Observational, retrospective and analytical study. The answers of a group of students were evaluated in a multiple choice exam of 50 questions with 4 answer options taken on September 8th, 2023. They were compared with the performance on the same exam of ChatGPT versions 3.5 and 4.
Results: Students (n = 7) correctly answered, on median, 39 questions (39/50), an accuracy of 78%, with a range of correct answers between 33 and 45. The average time to complete the exam was 75 minutes. ChatGPT-3.5 correctly answered 31 questions (31/50), an accuracy of 62%. Chat GPT-4 correctly answered 40 questions (40/50) an accuracy of 80% completing the exam in 73,49 seconds.
Conclusions: ChatGPT-4 achieved higher performance than the average student using 61 times less time. ChatGPT-4 achieved higher accuracy than ChatGPT-3.5. The grade obtained by the two versions of ChatGPT allows to pass the final exam of the University Medical Specialists in Ophthalmology Career at the University of Buenos Aires.

Keywords: artificial intelligence, ChatGPT-4, medical education, ophthalmology.

ChatGPT-4: seu desempenho no exame final dos estudos de médico especialista em oftalmologia da Universidade de Buenos Aires

Resumo

Objetivos: Avaliar o desempenho do chatGPT-4 em um exame final do curso universitário de médico com especialização em Oftalmologia da Universidade de Buenos Aires e compará-lo com o desempenho de dois alunos antes do mesmo exame e com o ChatGPT-3.5.
Material e métodos: Estudo observacional, retrospectivo e analítico. O desempenho de 7 médicos em exame final de pós-graduação de 50 questões com 4 opções de resposta realizado em 8 de setembro de 2023 foi comparado com seu desempenho no mesmo exame do ChatGPT versões 3.5 e 4.
Resultados: A média de acertos dos 7 alunos foi 39 (faixa 33-45), o que representa uma precisão de 78%. O tempo médio para conclusão do exame foi de 75 minutos. ChatGPT 3.5 respondeu corretamente 31 questões (31/50) alcançando uma precisão de 62%. O GPT Chat 4 acertou 40 questões (40/50) atingindo uma precisão de 80% e concluiu o exame em 73,49 segundos.
Conclusões: A versão ChatGPT-4 obteve um desempenho superior à mediana dos alunos utilizando 61 vezes menos tempo. ChatGPT-4 alcançou maior precisão que a versão 3.5. A pontuação obtida pelas duas versões do ChatGPT permite passar no exame, pois o limite para alcançá-lo é de 30 respostas corretas.

Palavras-chave: inteligência artificial, ChatGPT-4, educação médica, oftalmologia.

Introducción

ChatGPT (Open AI, San Francisco) es un chatbot de inteligencia artificial generativa. Es un transformador generativo preentrenado (*Generative Pre-trained Transformer*, GPT por su nombre y sigla en inglés) basado en una tecnología de procesamiento de lenguaje (*natural language processing*, NLP) definida como "grandes modelos de lenguaje" (LLM según sus sigla en inglés) que consiste en redes neuronales capaces de —entre otras posibilidades— leer, traducir y generar una respuesta ante una orden o instrucción (*prompt*). Estos sistemas han sido "alimentados" con los datos disponibles en los sitios abiertos de internet lo cual incluye textos médicos y artículos científicos.

La versión GPT4 está disponible desde marzo de 2023. La empresas Microsoft y Open AI han evaluado su uso en el área de la salud, entre otras aplicaciones.

Los estudios que evaluaron el desempeño del ChatGPT respondiendo preguntas de oftalmología han reportado una exactitud entre el 40% y el 80%[1].

Un área de interés es el potencial de esta tecnología para mejorar la educación médica. Respecto a su objetivo, este artículo es el primero en la literatura argentina que aborda el análisis del desempeño del ChatGPT-4 en un examen final de médico especialista universitario en oftalmología comparándolo con el desempeño de un grupo de alumnos.

Material y métodos

Estudio observacional, retrospectivo y analítico donde se evaluó el desempeño del Chat GPT en sus versiones 3.5 y 4 ante un examen final de la Carrera de Médico Especialista Universitario en Oftalmología de la Universidad de Buenos Aires (UBA), que rindió un grupo de siete alumnos de esa carrera el 8 de septiembre de 2023 en la Primera Cátedra de Oftalmología, Hospital de Clínicas "José de San Martín", Buenos Aires. El examen consistió en 50 preguntas de opción múltiple con 4 opciones de respuesta por pregunta abarcando todas las subespecialidades de la oftalmología. En cuanto al diseño del cuestionario: a) no hubo ninguna pregunta en la cual más de una opción de respuesta fuera correcta (ejemplo: "a y c son correctas"); b) no hubo ninguna pregunta con la opción "ninguna es correcta" y c) sólo una pregunta tenía como opción de respuesta "todas son correctas".

El análisis incluyó los siguientes ítems: a) el nivel de dificultad que plantearon las preguntas según el desempeño tanto de los alumnos como del GPT4 (preguntas "difíciles vs preguntas fáciles"); b) el desempeño de los alumnos ante las preguntas en las que GPT4 emitió respuestas incorrectas; c) el tipo de preguntas en las que GPT4 se equivocó y d) el tiempo utilizado para completar el examen (alumnos vs GPT4).

Resultados

La mediana de las respuestas correctas de los siete alumnos fue de 39 (39/50), lo que representa una exactitud del 78% con un rango de preguntas respondidas correctamente entre 33 y 45. Para completar el examen emplearon en promedio 75 minutos (con un rango entre 60 y 90 minutos). El tiempo límite disponible era de 90 minutos. El umbral de aprobación fue de 30 respuestas correctas (30/50).

Tabla 1. Cuadro comparativo de exactitud entre alumnos y GPT

	ALUMNOS	GPT 3.5	GPT 4
Exactitud	78 (66-90)	62%	80%
Aprobados	(7/7)	SI	SI

El ChatGPT-3.5 respondió correctamente 31 preguntas (31/50); esto representa una exactitud del 62%.

El Chat GPT-4 respondió correctamente 40 preguntas (40/50), lo que representa una exactitud del 80% (tabla 1).

El tiempo empleado por el ChatGPT-4 fue de 73,49 segundos. Esto representa un tiempo 61 veces inferior a los 4.500 segundos empleados en promedio por los alumnos (75 minutos).

Ningún alumno respondió bien la totalidad de las 10 preguntas en las que GPT-4 respondió erróneamente. En esas 10 preguntas, el porcentaje de respuestas incorrectas de los alumnos fue del 37,14% (26 respuestas incorrectas en las 70 respuestas de los alumnos para esas 10 preguntas).

Hubo 19 preguntas respondidas correctamente por todos los alumnos y también por GPT 4.

Dos preguntas fueron las que mostraron más respuestas incorrectas entre los alumnos (5 alumnos/7): preguntas 5 y 30, pero GPT4 las respondió correctamente.

Respecto de las respuestas coincidentes entre las de GPT4 y las respuestas de los alumnos, se subdividió el análisis entre las respuestas correctas de GPT4 y las respuestas incorrectas de GPT4.

Respecto de las coincidencias con las 40 respuestas correctas de GPT4, hubo 233 respuestas correctas de los alumnos coincidentes con GPT4 sobre 280 respuestas posibles (40 x 7). Esto representa un 83,92% de coincidencias de las respuestas de los alumnos comparadas con las respuestas correctas de GPT4.

Sobre las 10 respuestas incorrectas de GPT4, las respuestas de los alumnos a esas preguntas coincidieron en 17 casos con la respuesta que dio GPT 4 (17 coincidencias sobre 70 respuestas de los 7 alumnos a esas 10 preguntas). Esto representa un 24,28% de coincidencias de los alumnos con las respuestas equivocadas de GPT4.

GPT4 versus GPT3.5

De las respuestas equivocadas de GPT-4 en dos preguntas GPT-3.5 y GPT-4 coinciden en su respuesta y en 4 no coinciden.

La versión GPT-3.5 respondió correctamente 4 preguntas respondidas equivocadamente por GPT-4. Paralelamente, la versión GPT-4 respondió correctamente 13 preguntas en las que GPT3.5 respondió mal.

Discusión

En el transcurso de los últimos meses se ha reportado en la literatura científica el desempeño de Chatbots LLM en exámenes de estudiantes de medicina de grado y de posgrado y entre otras especialidades de la oftalmología.

En el examen del Colegio de Oftalmólogos del Reino Unido (UK), con la modalidad de examen de opción de respuestas múltiples (*multiple choice*) la exactitud en las respuestas para el ChatGPT-3.5 fue del 55,1% para la parte 1 del examen y del 49,6% para la parte 2, en tanto que el Chat GPT-4 logró una exactitud del 79,1% en la segunda parte[2].

Previamente se había reportado una exactitud del 67,6% en exámenes para médicos generalistas.

El desempeño de ChatGPT3.5 en el examen para la licencia médica en Estados Unidos (*United States Medical Licensing Exam*, USMLE) le permitió superar el umbral de exactitud del 60%[3].

Las respuestas erróneas de esta tecnología de inteligencia artificial configura en determinados casos lo que se denomina "alucinaciones" y se ha puesto el énfasis en el riesgo potencial de estos errores en un escenario médico dado que muchas veces esos errores son sutiles y están enmarcados en un texto coherente y convincente[4].

En otro estudio sobre el nivel de exactitud en preguntas oftalmológicas hubo una diferencia estadísticamente significativa en el desempeño entre GPT-3.5 y GPT-4. El Chat GPT-4 logró una exactitud del 73,2% vs 55,5% del GPT-3.5 ($p<0,001$)[5].

También se ha evaluado la competencia de los Chat GPT en lenguas diferentes del inglés. En uno de ellos se evaluó su desempeño en el examen japonés para la licencia médica (JMLE). Sobre un total de 254 preguntas GPT-4 superó los resultados de GPT-3.5[6].

La exactitud del GPT-3.5 en el Examen Nacional de Licencia Médica en China (*China National Medical Licensing Examination*, CNMLE) fue del 56% (56/100) en tanto que para GPT-4 fue del 84% (84/100)[7].

Como bien lo señala Eduardo de Vito, "el ChatGPT ha recibido críticas positivas y negativas. Estas últimas llegaron del ámbito de la educación y de la ciencia; por ejemplo, en la Universidad de Estrasburgo, Francia, descubrieron que 20 estudiantes realizaron un examen a distancia usando ChatGPT como asistente. El Departamento de Educación de la ciudad de Nueva York ha restringido el acceso a ChatGPT desde internet y los dispositivos en sus escuelas públicas de manera similar la Universidad Sciences Po de Paris"[8].

Más allá del debate respecto de la utilización de este tipo de instrumentos en la docencia médica, la utilidad práctica de la inteligencia artificial en el ejercicio de la medicina ha sido contundentemente demostrada en especialidades tan críticas como la oncología en la que —como lo expresa Luthy— "permitirá en un futuro cercano utilizar tratamientos personalizados para cada paciente con la pretensión de mejorar así su calidad de vida y su supervivencia"[9].

Respecto del análisis de los resultados obtenidos en la presente investigación, un aspecto interesante es concentrar la atención en el tipo de preguntas (y sus opciones de respuestas), en las que GPT-4 se equivocó. Se aplicaron dos abordajes en este análisis: 1) evaluar cuáles fueron los errores de GPT-4 independientemente del desempeño de los alumnos en esas preguntas, y 2) evaluar cómo respondieron los alumnos las respuestas en las que el GPT-4 se equivocó.

Respecto del primer abordaje nos preguntamos: a) si el enunciado de la pregunta había sido lo suficientemente claro; b) si alguna opción de respuesta construida como distractor no fue redactada adecuadamente; c) si las respuestas equivocadas tienen una implicancia seria desde la perspectiva clínica; y d) si las respuestas equivocadas ocurrieron a pesar de una opción de respuesta correcta muy evidente.

No hubo errores de GPT-4 adjudicables a los ítems a, b y c.

En cuanto al ítem "d" de este primer eje de análisis, GPT-4 se equivocó en una pregunta de bajo nivel de complejidad. La pregunta fue: "ante una leucocoria, la imagen en midriasis de procesos ciliares traccionados nos hace pensar en:

1) Vitreorretinopatía familiar exudativa
2) Retinopatía del prematuro
3) Retinoblastoma
4) Persistencia de vítreo primario hiperplásico.

GPT-4 optó por la "3".

No se detectó un "patrón" común en el tipo de preguntas en las que GPT4 se equivocó.

Respecto del desempeño de los alumnos en las 10 preguntas en las que se equivocó GPT-4, un aspecto para resaltar en el análisis es que: a) ningún alumno respondió correctamente la totalidad de esas 10 preguntas, b) los alumnos tuvieron en esas 10 preguntas un 37,14% de respuestas incorrectas (26 /70) y c) y en sólo 24,28%, la respuesta incorrecta de los alumnos y GPT4 coincidió. Los alumnos y GPT4 coincidieron en las preguntas "fáciles" (las que todos respondieron bien) pero en las dos preguntas "difíciles" para los alumnos, GPT4 no se equivocó.

Algo previsible pero que no deja de sorprender es la velocidad de respuesta de GPT-4 para completar el examen (73,49 segundos). Se trata de una cifra 61 veces inferior al promedio del tiempo utilizado por los alumnos.

Resultó claramente superior el desempeño de GPT-4 respecto de su antecesor 3.5.

Conclusiones

Chat GPT-4 demostró tener una exactitud para aprobar con amplia suficiencia el examen final teórico de la carrera de médico especialista universitario en oftalmología de la Facultad de Medicina de la Universidad de Buenos Aires (UBA) con sede Hospital de Clínicas, con un desempeño superior al de la mediana de los alumnos con los que fue comparado. Su velocidad de respuesta fue 61 veces superior a la del promedio de los alumnos y su exactitud fue significativamente mejor que la del ChatGPT-3.5.

Referencias

1. Ting DSJ, Tan TF, Ting DSW. ChatGPT in opthalmology: the dawn of a new era ? *Eye (Lond)* 2024; 38: 4-7.

2. Raimondi R; Tzoumas N; North East Trainee Research in Ophthalmology Network (NETRiON) *et al*. Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. *Eye (Lond)* 2023; 37: 3530-3533.

3. Kung TH, Cheatham M, Medenilla A *et al*. Performance of Chat GPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023; 2: e0000198.

4. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023; 388: 1233-1239.

5. Moshirfar M, Altaf AW, Stoakes IM *et al*. Artificial intelligence in opthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. *Cureus* 2023; 15: e40822.

6. 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: e48002.

7. Wang H, Wu W, Dou Z *et al*. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. *Int J Med Inform* 2023; 177: 105173.

8. De Vito E. Inteligencia artificial y chat GPT. ¿Usted leería a un autor artificial? *Medicina (B Aires)* 2023; 83: 329-332.

9. Lüthy IA. Inteligencia artificial y aprendizaje de máquina en diagnóstico y tratamiento del cáncer. *Medicina (B Aires)* 2022; 82: 798-800.