Revista Española de Patología xxx (xxxx) xxx–xxx

SEAP-IAP

REVISTA ESPAÑOLA DE Patología

www.elsevier.es/patologia

ORIGINAL

ChatGPT es un estudiante por encima de la media en la Facultad de Medicina de la Universidad de Zaragoza y un colaborador excelente en la elaboración de material docente

Clara Cabañuz[a,b] y Mar Garcia Garcia[a,b,*]

[a] *Anatomía Patológica, Hospital Clínico Universitario Lozano Blesa, Zaragoza, España*
[b] *Departamento de Cirugía, Facultad de Medicina, Universidad de Zaragoza, Zaragoza, España*

Recibido el 8 de noviembre de 2023; aceptado el 12 de enero de 2024

PALABRAS CLAVE
Grado de medicina;
Anatomía patológica;
Inteligencia artificial;
ChatGPT;
Docencia

Resumen

Introducción y objetivo: La inteligencia artificial se halla plenamente presente en nuestras vidas. En educación las posibilidades de su uso son infinitas, tanto para alumnos como para docentes.

Material y métodos: Se ha explorado la capacidad de ChatGPT a la hora de resolver preguntas tipo test a partir del examen de la asignatura Procedimientos Diagnósticos y Terapéuticos Anatomopatológicos de la primera convocatoria del curso 2022-2023. Además de comparar su resultado con el del resto de alumnos presentados, se han evaluado las posibles causas de las respuestas incorrectas. Finalmente, se ha evaluado su capacidad para realizar preguntas de test nuevas a partir de instrucciones específicas.

Resultados: ChatGPT ha acertado 47 de las 68 preguntas planteadas, obteniendo una nota superior a la de la media y mediana del curso. La mayor parte de preguntas falladas presentan enunciados negativos, utilizando las palabras «no», «falsa» o «incorrecta» en su enunciado. Tras interactuar con él, el programa es capaz de darse cuenta de su error y cambiar su respuesta inicial por la correcta. Finalmente, ChatGPT sabe elaborar nuevas preguntas a partir de un supuesto teórico o bien de una simulación clínica determinada.

Conclusiones: Como docentes estamos obligados a explorar las utilidades de la inteligencia artificial, e intentar usarla en nuestro beneficio. La realización de tareas que suponen un consumo de tipo importante, como puede ser la elaboración de preguntas tipo test para evaluación de contenidos, es un buen ejemplo.

* Autor para correspondencia.
Correo electrónico: margg94972@gmail.com (M. Garcia Garcia).

https://doi.org/10.1016/j.patol.2024.01.003
1699-8855/

Cómo citar este artículo: C. Cabañuz and M. Garcia Garcia, ChatGPT es un estudiante por encima de la media en la Facultad de Medicina de la Universidad de Zaragoza y un colaborador excelente en la elaboración de material docente, Revista Española de Patología , https://doi.org/10.1016/j.patol.2024.01.003

KEYWORDS
Medical degree;
Pathology;
Artificial intelligence;
ChatGPT;
Teaching

Abstract
Introduction and objective: Artificial intelligence is fully present in our lives. In education, the possibilities of its use are endless, both for students and teachers.
Material and methods: The capacity of ChatGPT has been explored when solving multiple choice questions based on the exam of the subject «Anatomopathological Diagnostic and Therapeutic Procedures» of the first call of the 2022-23 academic year. In addition, to comparing their results with those of the rest of the students presented the probable causes of incorrect answers have been evaluated. Finally, its ability to formulate new test questions based on specific instructions has been evaluated.
Results: ChatGPT correctly answered 47 out of 68 questions, achieving a grade higher than the course average and median. Most failed questions present negative statements, using the words «no», «false» or «incorrect» in their statement. After interacting with it, the program can realize its mistake and change its initial response to the correct answer. Finally, ChatGPT can develop new questions based on a theoretical assumption or a specific clinical simulation.
Conclusions: As teachers we are obliged to explore the uses of artificial intelligence and try to use it to our benefit. Carrying out tasks that involve significant consumption, such as preparing multiple-choice questions for content evaluation, is a good example.

Introducción

Desde sus inicios a mitad del siglo xx, la inteligencia artificial (IA) está presente en la actualidad en una amplia variedad de campos. Por IA entendemos el desarrollo de máquinas que son capaces de realizar tareas que normalmente requieren inteligencia humana, como pueden ser la capacidad de aprender, adaptarse, racionalizar, comprender conceptos abstractos, así como reaccionar ante atributos humanos complejos como la atención, la emoción, la creatividad, etc.[1]. De sus distintas ramas, cabe destacar la IA generativa (IAG)[2], que se ocupa de crear contenidos originales a partir de datos o instrucciones. Una de las herramientas más conocidas hoy en día de este tipo de IA es «ChatGPT». Desde su lanzamiento en noviembre de 2022, desarrollado por OpenAI (OpenAI, L.L.C., San Francisco, CA, EE.UU.), es un programa capaz de comprender y generar respuestas utilizando una interfaz basada en texto, conocida como chatbot, entrenada en grandes conjuntos de datos en varios idiomas y con la capacidad de generar respuestas similares a las humanas a partir de instrucciones escritas[3].

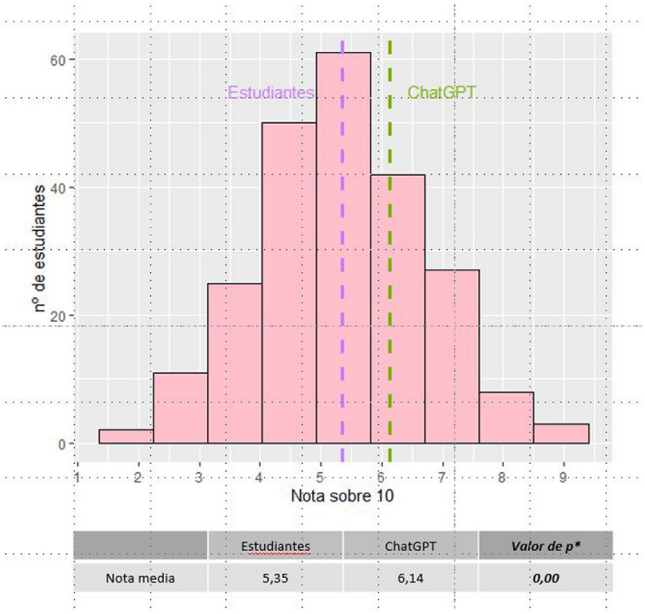
En el campo de la educación médica en ámbito académico, existen en la literatura ya algunos trabajos que describen el potencial uso de la IAG. Las posibilidades son infinitas[1]. El estudiante puede utilizarlo como fuente de información, realizar tareas escritas, disponer de un tutor personalizado que va adaptando los contenidos a su ritmo, con lo que se promueve un mayor compromiso y motivación, o incluso tener lo que se conoce como tutorías inteligentes, con orientación y apoyo personalizado durante su proceso de aprendizaje. Pero las aplicaciones de esta tecnología no solo se limitan en beneficio del estudiante, sino que también pueden ser de utilidad para el docente. La generación de contenidos educativos puede beneficiar la metodología docente basada en simulaciones clínicas, entre otras ventajas.

En esta revisión se ha explorado el uso potencial de la IA mediante la herramienta ChatGPT en la elaboración de pruebas de examen para estudiantes del grado de medicina a través de la resolución de un examen tipo test basado en preguntas centradas en conocimientos teóricos.

Material y métodos

Se ha realizado una revisión con análisis descriptivo de la capacidad de resolución de ChatGPT, en su versión gratuita (GPT-3.5), del examen de la asignatura de Procedimientos Diagnósticos y Terapéuticos Anatomopatológicos correspondiente a la primera convocatoria del curso 2022-2023. El examen consta de 100 preguntas: 70 de contenido teórico y 30 de contenido práctico. Se han excluido las 30 preguntas de la parte práctica del examen, ya que están basadas en imágenes, y estas no se pueden copiar en esta versión del programa. También se han excluido otras dos preguntas de índole teórica; estas preguntas fueron impugnadas por los estudiantes y finalmente anuladas en dicha convocatoria por distintos motivos.

Una vez realizado este paso, se ha abierto un chatbot en el que se le ha indicado la siguiente instrucción o prompt: «por favor, contesta a la siguiente pregunta test». Se han comparado sus respuestas con la plantilla de respuestas correctas utilizada en dicha convocatoria. Se han recalificado las notas de todos los alumnos presentados en base a los resultados en las 68 preguntas evaluadas, con penalización de −0,25 por pregunta fallada. Finalmente, dichos



	Estudiantes	ChatGPT	*Valor de p**
Nota media	5,35	6,14	*0,00*

Figura 1 Distribución de las calificaciones obtenidas por los alumnos y comparación con ChatGPT. En el histograma de frecuencias se observa que las calificaciones obtenidas por los alumnos siguen una distribución normal. Existe una diferencia estadísticamente significativa entre la nota media de los alumnos y la nota obtenida por ChatGPT.
t de Student para muestra única; *: diferencias significativas si $p < 0,05$).

resultados han sido ponderados a una nota máxima de 10. Para los cálculos estadísticos realizados se ha utilizado el programa IBM SPSS Statistics 29.0.

En una segunda fase se han evaluado las respuestas incorrectas para ver el posible motivo de error. En esta fase también se ha interaccionado con el programa con el fin de aportar nuevos argumentos que él pudiera utilizar para responder correctamente a la pregunta planteada. Finalmente, se le ha propuesto una tarea final en la que ChatGPT ha generado preguntas test en base a un planteamiento o contexto clínico y a la información aportada, en ocasiones con inclusión de las distintas variables distractoras en las respuestas y en otras con elección libre de las cinco opciones de respuesta.

Resultados

Un total de 228 alumnos se presentaron al examen de la convocatoria de enero-febrero del curso 2022-2023 de la asignatura Procedimientos Diagnósticos y Terapéuticos Anatomopatológicos. La nota media del curso ha sido de 5,35, y la mediana, de 5,29, siendo la nota más baja 1,99 y la más alta 9,15. ChatGPT acertó un total de 47 preguntas, muy por encima de las esperadas por azar (13,6), y obtuvo 21 fallos, lo que le otorgó una puntuación final de 41,75 puntos, correspondientes a una nota ponderada a 10 de 6,14, por encima de la media y la mediana del curso (fig. 1).

ChatGPT ha razonado cada una de sus respuestas, tanto en las que ha respondido correctamente como en las que ha fallado, aportando siempre la base teórica en la que se ha basado al responder en cada caso (figs. 2 y 3). En cuanto a las preguntas respondidas erróneamente, la mayor parte (11 de 21 [52%]) corresponden a preguntas con enunciado negativo, ya sea utilizando «no» o «falsa» o «incorrecta». En otros casos se ha valorado la posibilidad de ambigüedades en el enunciado de la pregunta o en las respuestas posibles, mientras que en otros el problema creemos ha sido de interpretación al no ser capaz de distinguir algunos matices. En cualquier caso, cuando se han formulado nuevas preguntas relacionadas con la base teórica de la pregunta con respuesta incorrecta, ChatGPT ha sido capaz de aportar nueva información y, finalmente, reconsiderar su respuesta inicial falsa por la respuesta verdadera (ver anexo 1). Por último, se ha demostrado la capacidad del programa para elaborar preguntas tipo test a partir de una entrada de texto con las instrucciones necesarias (ver anexo 2).

Discusión

Desde su lanzamiento, ChatGPT ha recibido en la comunidad científica y académica reacciones mixtas que reflejan la historia de controversia sobre los beneficios frente a los riesgos de las tecnologías de IA avanzadas. De entre los riesgos, los más frecuentemente mencionados son el bias de la fuente de información con la que se nutre el programa, así como la manipulación de la información tras ciberataques[1]. Cuando se refiere al uso de esta tecnología en el campo de la medicina, se incluyen además otro tipo de desafíos y riesgos, como pueden ser la precisión y la validez de la información generada, la ética sobre el origen y el uso de los datos, así como la privacidad y la seguridad de los datos tratados.

En el ámbito académico se añaden, también, otros factores independientes a la herramienta y directamente relacionados con los planes de estudio actuales y la resistencia al cambio. En cuanto a los beneficios, no queda duda de las oportunidades que ofrece a los estudiantes en distintas facetas durante su formación. Además, perciben la información aportada por esta herramienta como más estructurada y más clara en comparación con otras fuentes[4]. Pero hay que educarles en su uso; es imperativo advertirles de que la información recabada tiene que ser contrastada, más cuando el conocimiento del tema puede ser nulo o superficial. Como se ha demostrado en este trabajo, algunas de las respuestas generadas por ChatGPT no son correctas, y, de ser utilizadas por alguien sin capacidad de reconocer y/o corregir errores, pueden plantear interrogantes sobre su fiabilidad y su precisión. La IA tampoco tiene que ser un sustituto del trabajo de aprendizaje del alumno, y el inculcarles un pensamiento crítico es fundamental para su futuro.

Como docentes, la IAG también puede sernos de gran ayuda, especialmente en tareas que suponen un gran consumo de tiempo, como puede ser la elaboración de material didáctico o la confección de exámenes basados en preguntas tipo test. Se ha demostrado previamente que ChatGPT es capaz de superar pruebas de examen de conocimien-

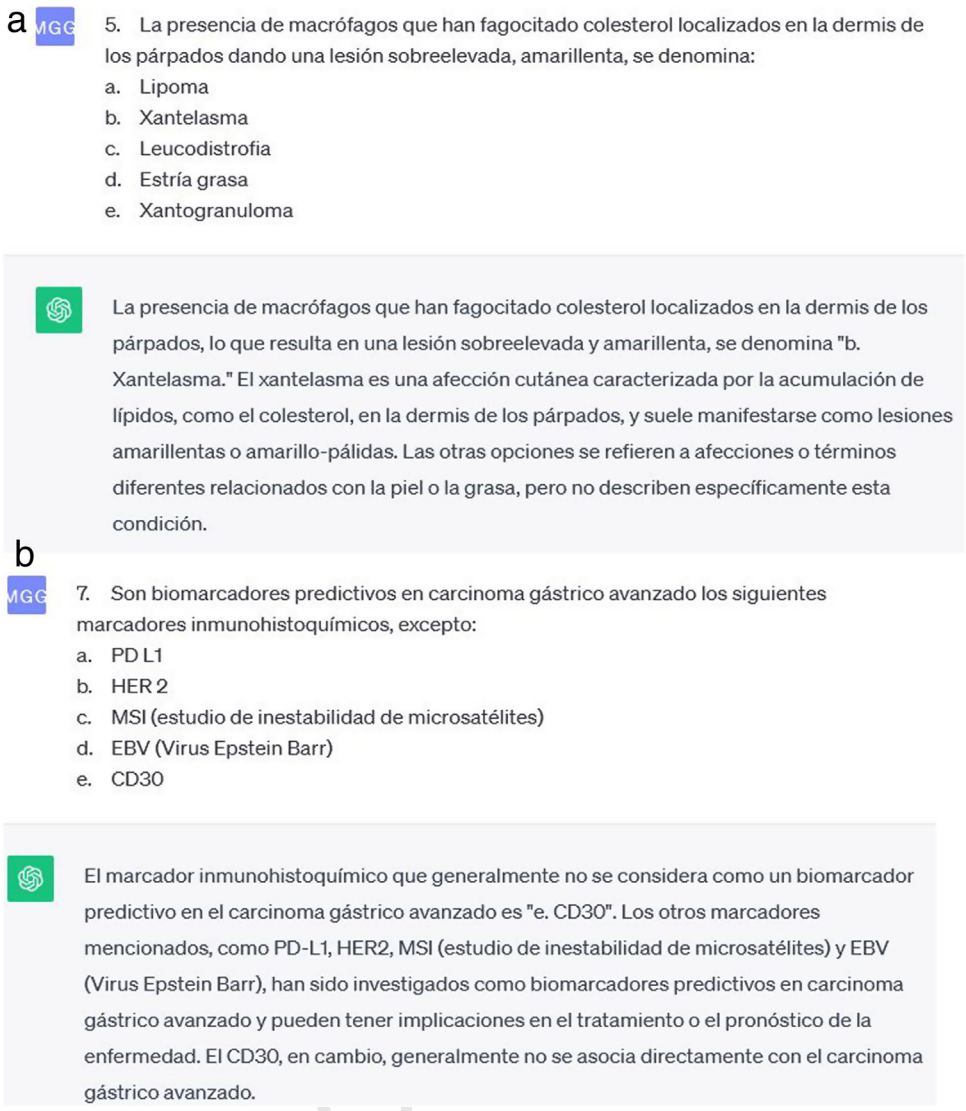


Figura 2 Ejemplos de preguntas con respuesta correcta, correspondientes a las preguntas 5 (2a) y 7 (2b) del examen. Se incluye, además, el razonamiento en el que basa su respuesta.

tos médicos, tanto de grado como de especialidad[5,6]. Esta capacidad ha sido expuesta también en este trabajo, obteniendo 47 respuestas correctas del total de las 68, muy por encima de los aciertos que se podrían esperar como resultado del azar. De hecho, la probabilidad de acertar 47 preguntas es prácticamente nula. Además, hemos podido observar que, en las respuestas falladas, tras conversar mediante la introducción de nuevas preguntas, el programa ha sido capaz de darse cuenta del error inicial y de responder finalmente de forma correcta. Esta capacidad puede ser explotada con el objetivo de testar la calidad de las preguntas previamente preparadas y evitar así que algunas de las cuestiones del examen puedan ser impugnadas por los estudiantes. Por otro lado, se ha demostrado que ChatGPT puede elaborar preguntas tipo test a partir de unas instrucciones específicas en forma de texto, basadas en contextos clínicos[7]. Dichas instrucciones pueden ir refinándose y mejorándose en función de nuestro objetivo. Esta competencia también ha sido demostrada aquí, y puede ser útil en la elaboración de nuevas preguntas tipo test para evaluación de conocimientos, por ejemplo, con el objetivo de elaborar un banco de preguntas. Lo mejor de todo esto es que estas tareas son realizadas en pocos segundos de tiempo.

a

17. El esófago de Barret NO se caracteriza por:
a. Metaplasia de glándulas gástricas en esófago distal sin metaplasia intestinal
b. Parches de mucosa rosada en esófago distal
c. Metaplasia intestinal con células Globet en esófago distal
d. Está en relación con el reflujo gastro-esofágico
e. Es una lesión precursora de adenocarcinoma de esófago distal

El esófago de Barrett se caracteriza por "c. Metaplasia intestinal con células de Goblet en el esófago distal."

El esófago de Barrett es una condición en la cual el revestimiento normal del esófago es reemplazado por tejido similar al revestimiento intestinal, con células de Goblet. Esto se debe a la exposición crónica al ácido gástrico y el reflujo gastroesofágico. Es una lesión precursora de adenocarcinoma de esófago distal (opción e) y es importante debido al riesgo aumentado de cáncer asociado a esta condición. Las opciones a, b y d son características de esta condición y están relacionadas con los cambios en la mucosa esofágica debido al esófago de Barrett.

b

57. ¿Cuál de las siguientes características NO es propia de una neumonía intersticial usual (NIU)?
a. Patrón radiológico de panalización pulmonar
b. Nódulos fibroblásticos
c. Fibrosis intersticial difusa y bilateral
d. Etiología desconocida
e. Todas son características habituales en una NIU

La característica que NO es propia de una neumonía intersticial usual (NIU) es:

b. Nódulos fibroblásticos

Las características típicas de una neumonía intersticial usual (NIU) incluyen un patrón radiológico de panalización pulmonar, fibrosis intersticial difusa y bilateral, y una etiología desconocida. Los nódulos fibroblásticos no son una característica típica de la NIU, por lo que la respuesta correcta es la opción b.

Figura 3 Ejemplos de preguntas con respuesta incorrecta, correspondientes a las preguntas 17 (3a) y 57 (3b). Se incluye, además, el razonamiento, aunque incorrecto, en el que basa su respuesta.

Conclusiones

La IA se halla presente plenamente en nuestras vidas, y debemos verla como una oportunidad. Como docentes no podemos dar la espalda a esta herramienta, y la vez utilizarla en nuestro favor, sobre todo en aquellas tareas en las que su uso nos ahorra tiempo, como puede ser la preparación de exámenes en la asignatura del grado de Medicina, a la vez que reduce la probabilidad de error, con la consecuente anulación de preguntas. Queda claro que para algunas tareas su uso es más limitado, en especial para abordar preguntas con enunciados negativos o en la resolución de cuestionarios complejos que requieran niveles de comprensión y análisis mucho más profundos. Finalmente, debemos enseñar unas

buenas prácticas a nuestros estudiantes en el uso de esta herramienta, así como hacerles conscientes de sus limitaciones.

Consideraciones éticas

Este trabajo es un artículo original basado en una revisión de datos que no implica sujetos y por lo tanto no implica modificación en su manejo clínico. Por otro lado, a pesar de que el trabajo está basado en inteligencia artificial, no se ha utilizado dicha tecnología para escribirlo.

Financiación

La presente investigación no ha recibido ayudas específicas provenientes de agencias del sector público, sector comercial o entidades sin ánimo de lucro

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Agradecimientos

A Rocío Bermúdez Cameo.

Anexo. Material adicional

Se puede consultar material adicional a este artículo en su versión electrónica disponible en https://doi.org/10.1016/j.patol.2024.01.003.

Bibliografía

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11:887, http://dx.doi.org/10.3390/healthcare11060887.
2. Mayol J. Inteligencia artificial generativa y educación médica. Educación Médica. 2023;24:1–3, http://dx.doi.org/10.1016/j.edumed.2023.100851.
3. OpenAI. OpenAI: Models GPT-3.5. Disponible en: https://openai.com/chatgpt
4. Breeding T, Martinez B, Patel H, Nasef H, Arif H, Nakayama D, et al. The utilization of ChatGPT in reshaping future medical education and learning perspectives: A curse or a blessing? Am Surg. 2023, http://dx.doi.org/10.1177/00031348231180950.
5. Fuentes-Martín A, Cilleruelo-Ramos A, Segura-Méndez B, Mayol J. Can an artificial intelligence model pass an examination for medical specialists? Arch Bronconeumol. 2023;59:534–6, http://dx.doi.org/10.1016/j.arbres.2023.03.017.
6. 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;1:55–69 https://doi.org/10.6018/edumed.556511
7. Kıyak YS. A ChatGPT prompt for writing case-based multiple-choice questions. Rev Esp Edu Med. 2023;3:98–103 https://doi.org/10.6018/edumed.587451