Open Respiratory Archives 6 (2024) 100340

Sociedad Española de Neumología y Cirugía Torácica SEPAR

Open Respiratory Archives

www.elsevier.es/ora

Carta al Editor

Estudio comparativo de la capacidad de aprendizaje de ChatGPT en la resolución de preguntas de especialización médica

A comparative study on the learning capabilities of ChatGPT in medical specialization query resolution

Estimado editor,

El reciente estudio «*Can an Artificial Intelligence Model Pass an Examination for Medical Specialists?*»[1] publicado en la revista *Archivos de Bronconeumología*, refleja la capacidad de ChatGPT (OpenAI, San Francisco, EE. UU.), un modelo del Procesamiento del Lenguaje Natural (PLN) entrenado mediante algoritmos de aprendizaje automático, en la resolución de preguntas de medicina especializada mediante la superación de una fase opositiva de cirugía torácica.

El objetivo de esta carta es realizar una reflexión sobre la capacidad actual de aprendizaje de dichos modelos de Inteligencia Artificial Generativa (IAG). Para ello, hemos evaluado su capacidad de mejora en la resolución de dichas preguntas de temática médica en un intervalo de 90 días.

Se ha realizado un análisis descriptivo de la capacidad de resolución de ChatGPT-3.5 frente a ChatGPT-4 respecto al mismo examen de oposición de la especialidad de cirugía torácica en la convocatoria de 2022 del Servicio Andaluz de Salud.

La resolución de preguntas por ChatGPT se realizó a través de su plataforma *online* en dos intervalos: 10/02/2023-15/02/2023 y 11/05/2023-13/05/2023, utilizando el siguiente *prompt*: «*RESPONDE LA SIGUIENTE PREGUNTA TEST*:». Se utilizaron sesiones independientes para cada pregunta del cuestionario teórico, utilizándose la misma sesión para las series de preguntas basadas en el mismo escenario, aumentando el rendimiento del modelo mediante la utilización del sesgo de retención de memoria del mismo. Se utilizó como patrón de respuesta la plantilla oficial definitiva publicada por la administración pública. El examen contó con 146 preguntas (cuestionario teórico: 98/cuestionario práctico: 48).

ChatGPT-3.5 alcanzó una tasa de acierto global del 58,9% (86), desglosada en un 63,2% (62) en el cuestionario teórico y un 50% (24) en el práctico. Por otro lado, ChatGPT-4 obtuvo una tasa de acierto global del 65,7% (96), con un 71,43% (70) en el cuestionario teórico y un 54,16% (26) en el práctico. Aplicando los criterios de puntuación, ChatGPT-4, como ya consiguió ChatGPT-3.5, aprobaría este examen de oposición; sin embargo, el análisis inferencial no reveló diferencias estadísticamente significativas ($p > 0.05$) con respecto a la tasa de respuestas correctas entre ambas versiones.

Nuestro estudio contrasta con otras publicaciones que han evaluado de forma reciente la capacidad de aprendizaje de dichos modelos de IAG respecto a la resolución de escenarios específicos dentro del ámbito de la medicina, entre ellos, por ejemplo, se ha evidenciado una mejoría en la capacidad de resolución de ChatGPT-3.5 frente a ChatGPT-4 en el ámbito de la oncología radioterápica[2] o de la oftalmología[3,4].

Estos hallazgos nos deben hacer reflexionar sobre la magnitud del progreso de los modelos de IAG al enfrentarse a áreas de razonamiento crítico complejo. Es crucial puntualizar que la precisión y la validez de la información generada por estos modelos de IAG dependen no solo de los algoritmos aplicados y su capacidad computacional, sino también de forma directa de la veracidad de los datos de los que aprenden estos modelos[5,6].

Como conclusión, la capacidad de aprendizaje en los modelos de IAG puede ser significativa y de valor para la práctica médica en contextos específicos. Los autores consideramos imperativo que la comunidad científica desempeñe un papel activo en garantizar la precisión y validez de la información generada y de los datos utilizados en el entrenamiento de este tipo de modelos de IAG, así como en la evaluación del progreso y en la aplicación de estos al ámbito de la medicina.

Financiación

Este trabajo no ha recibido ningún tipo de financiación.

Contribuciones de los autores

Todos los autores participaron en la concepción y diseño del trabajo. Todos los autores creen que el manuscrito representa un trabajo válido, lo han leído y lo han aprobado completamente. Los autores garantizan que el artículo es original y no ha sido enviado a otra revista para su publicación.

Conflictos de interés

Los autores declaran no tener ningún conflicto de intereses.

Bibliografía

1. 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):534–6, http://dx.doi.org/10.1016/j.arbres.2023.03.017.
2. Huang Y, Gomaa A, Semrau S, Haderlein M, Lettmaier S, Weissmann T, et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology. Front Oncol. 2023;13:1265024, http://dx.doi.org/10.3389/fonc.2023.1265024.
3. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine. 2023;95:104770, http://dx.doi.org/10.1016/j.ebiom.2023.104770.
4. Raimondi R, Tzoumas N, Salisbury T, Di Simplicio S, Romano MR, North East Trainee Research in Ophthalmology Network (NETRiON). Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. Eye (Lond). 2023;37(17):3530–3, http://dx.doi.org/10.1038/s41433-023-02563-3.

https://doi.org/10.1016/j.opresp.2024.100340

5. Lukoianova T, Rubin V. Veracity Roadmap: Is Big Data Objective, Truthful and Credible? Advances in Classification Research Online. 2014;24(1), http://dx.doi.org/10.7152/ACRO.V24I1.14671.
6. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, 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):e0000198, http://dx.doi.org/10.1371/journal.pdig.0000198.

Álvaro Fuentes-Martín[a,*], Ángel Cilleruelo-Ramos[a], Bárbara Segura-Méndez[a] y Julio Mayol[b]

[a] *Hospital Clínico Universitario de Valladolid, Universidad de Valladolid, Valladolid, España*

[b] *Hospital Clínico San Carlos, IdISSC, Universidad Complutense de Madrid, Madrid, España*

[*] Autor para correspondencia.
Correo electrónico: alvarofuentesmartin@gmail.com
(Á. Fuentes-Martín).
𝕏 @alvar0fuentes (Á. Fuentes-Martín), @angelcilleruelo (Á. Cilleruelo-Ramos), @juliomayol (J. Mayol)