



Scientific Letter

Can an Artificial Intelligence Model Pass an Examination for Medical Specialists?



To the Director,

ChatGPT, Chat Generative Pre-Trained Transformer,¹ is an artificial intelligence (AI) language model trained on a massive set of internet text data using machine learning and natural language processing (NLP) algorithms. It can generate human-like automatic responses to a variety of questions and stimuli in multiple languages and subject areas.² The main aim of this study was to evaluate how ChatGPT operates when responding to multiple-choice questions in a highly specialized area of state-of-the-art medical expertise.

We conducted a descriptive analysis of the performance of ChatGPT (OpenAI, San Francisco; Version: 9) in the 2022 competitive exams for the post of Specialist in thoracic surgery announced by the Andalusian Health Service.³ This particular exam was chosen because it uses a multiple-choice format with 4 possible answers, only one of which is correct. Participants answer 2 sets of questions: one, consisting of 100 direct questions, is theoretical, and the other, consisting of 50 questions, is practical and addresses clinical scenarios focused on critical reasoning.

ChatGPT answered the questions on its online platform between 10/02/2023 and 15/02/2023, in response to the following wording: "ANSWER THE FOLLOWING MULTIPLE-CHOICE QUESTION:" Separate sessions were used for each of the theoretical questions, while the practical questions were answered in the same session, using the memory retention bias of the artificial intelligence model to increase its performance. The definitive official template published by the public administration³ was used as a model answer. The examination consisted of 146 questions (theoretical section: 98/practical section: 48) after the Andalusia Health Service excluded 7 questions and included another 3 reserve questions.

ChatGPT answered 58.90% (86) of the answers correctly: inferential analysis revealed significant differences with respect to the rate of correct answers that could be attributed to chance (25%) with a level of statistical significance of 99% ($p < 0.001$). The pass rate for the theoretical section was 63.2% (62) compared to 50% for the practical section (24). Scoring criteria were applied for each correct question, including a penalty of -0.25 for each incorrect answer and weighting criteria for each section of the examination phase. The threshold specified in the official call for passing the exam was 60% of the average of the best 10 scores.³ A pass mark of 40 points was set. The artificial intelligence model would therefore have passed this part of the access examination for thoracic surgery physician/specialist with a score of 45.79 points.

Our results are in line with the existing literature on the potential of ChatGPT for completing question-answer tasks in different areas of knowledge, including the medical field. For example, evidence has been published of a correct response rate of over 60% for the United States Medical Licensing Examination (USMLE) Step 1, over 57% for the USMLE Step-2,⁴ and over 50% in the access examination for specialist residency posts in Spain in 2022.⁵

In our study, the AI tool performed worse when answering practical questions compared to theoretical questions, suggesting that it has difficulties in responding to clinical practice scenarios that require critical reasoning. As limitations of the study, it should be noted that we did not analyze the model's ability to respond correctly depending on the way the questions were formulated, nor did we analyze the justification that the AI model gave for each correct or incorrect answer (Fig. 1).

In conclusion, the ChatGPT model was capable of passing a competitive exam for the post of specialist in thoracic surgery, although differences were observed in its performance in areas in which critical reasoning is required. The emergence of AI tools that can resolve a variety of questions and tasks, including in the health field, their potential for development, and their incorporation into our training and daily clinical practice are a challenge for the scientific community.

A

ANSWER THE FOLLOWING MULTIPLE-CHOICE QUESTION:

In fibrosing mediastinitis, it is FALSE to say that:

- (A) It is characterized macroscopically by diffuse and poorly defined fibrotic infiltration of mediastinal structures.
- (B) It may be associated with a number of syndromes, such as retroperitoneal fibrosis.
- (C) Patients are usually young at the time of presentation, but cases have been observed in the fourth to fifth decades of life.
- (D) Once diagnosed, progress is rapid, with a 5-year mortality rate of 60%.

The false answer is D. Fibrosing mediastinitis is characterized by diffuse and poorly defined fibrotic infiltration of mediastinal structures. It may be associated with a number of syndromes, such as retroperitoneal fibrosis. Patients are young at the time of presentation, but cases have been observed in the fourth to fifth decades of life. Disease progress is variable and can be slow or rapid, and mortality varies depending on the extent and severity of the disease and the treatment received. The general 5-year mortality rate for all cases of fibrosing mediastinitis is not 60%. Therefore, option D is false.

B

ANSWER THE FOLLOWING MULTIPLE-CHOICE QUESTION:

PRACTICAL CASE 2:

An infant presents with severe, life-threatening respiratory dysfunction. Chest X-ray reveals a fluid-filled mass with smooth, clearly defined margins in the subcarinal region, causing displacement and stenoses of both main bronchi.

Your first diagnostic suspicion would be:

- (A) Cystic adenomatoid malformation.
- (B) Bronchogenic cyst.
- (C) Intrapulmonary sequestration.
- (D) Non-Hodgkin lymphoma.

The correct answer is B) Bronchogenic cyst.

The description of a fluid-filled mass on chest X-ray, with smooth margins and displacement of both main bronchi, is suggestive of a bronchogenic cyst. Bronchogenic cysts are congenital lesions that develop from the cells of the bronchial bud and can occur anywhere in the trachea or bronchi. They may contain air, fluid, or a combination of both.

Cystic adenomatoid malformation, intrapulmonary sequestration, and non-Hodgkin lymphoma are other possible causes of chest masses, but their clinical and radiological presentation usually differs from that of a bronchogenic cyst. Therefore, the most likely response in this case is B) Bronchogenic cyst.

Fig. 1. Example of questions from the competitive exam and responses of the ChatGPT model. (A) Theoretical section. (B) Practical section.

Conflict of Interests

The authors state that they have no conflict of interests.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi: 10.1016/j.arbres.2023.03.017](https://doi.org/10.1016/j.arbres.2023.03.017).

References

1. ChatGPT [Web]. <https://openai.com/blog/chatgpt/2023> [accessed 18.02.23].
2. Scott K. Microsoft teams up with Open AI to exclusively license GPT-3 language model. The Official Microsoft Blog [Web]. <https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/2020> [accessed 18.02.23].
3. Boletín Oficial de la Junta de Andalucía (2021, June 22). 118 – Tuesday, June 22, 2021. Depósito Legal: SE-410/1979. ISSN: 2253-802X.
4. 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:e45312, <http://dx.doi.org/10.2196/45312> [PMID: 36753318; PMCID: PMC9947764].

5. 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, <http://dx.doi.org/10.6018/edumed.556511revistas.um.es/edumed>.

Álvaro Fuentes-Martín^{a,*}, Ángel Cilleruelo-Ramos^a,
Bárbara Segura-Méndez^a, Julio Mayol^b

^a Hospital Clínico Universitario de Valladolid, Universidad de Valladolid, Spain

^b Hospital Clínico San Carlos, IdISSC, Universidad Complutense de Madrid, Spain

Corresponding author.

E-mail address: alvarofuentesmartin@gmail.com

(Á. Fuentes-Martín)

 [@alvarofuentes](https://twitter.com/alvarofuentes).