SCIE DOI: 10.14744/scie.2024.91129
South. Clin. Ist. Euras. 2024;35(3):209-212

Original Article

Large Language Model and Medical Education: Evaluation of Human and Artificial Intelligence Responses to Thoracic Surgery Questions

Mesut Buz, Recep Demirhan

Department of Thoracic Surgery, Health Sciences University, Kartal Dr. Lütfi Kırdar City Hospital, İstanbul, Türkiye

Submitted: 05.08.2024
Revised: 05.08.2024
Accepted: 13.08.2024

Correspondence: Mesut Buz, Department of Thoracic Surgery, Health Sciences University, Kartal Dr. Lütfi Kırdar City Hospital, İstanbul, Türkiye
E-mail: mesutbuzmd@gmail.com

Keywords: Artificial intelligence; ChatGPT-4; large language models; thoracic surgery.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ABSTRACT

Objective: This study aimed to evaluate the performance of ChatGPT-4, a large language model, in answering thoracic surgery questions compared to 5th-year medical students. The goal was to assess the potential of ChatGPT-4 as an educational tool in medical training.

Methods: A retrospective comparative analysis was conducted involving 10 fifth-year medical students and ChatGPT-4. Each participant answered 40 multiple-choice questions related to thoracic surgery. The students' scores were compared to the scores generated by ChatGPT-4. Statistical analysis was performed using an independent sample t-test to determine the significance of the differences in performance.

Results: The students' scores ranged from 80% to 97.5%, with an average score of 88.25% (SD=5.63). ChatGPT-4 scored 95% on the same set of questions. The t-test results indicated a statistically significant difference between the students' scores and ChatGPT-4's score (t=-3.98, p=0.00088).

Conclusion: The study demonstrated that ChatGPT-4 can provide accurate answers to thoracic surgery questions, surpassing the performance of 5th-year medical students. This indicates the potential of large language models as valuable educational tools in medical training. However, further research is needed to evaluate the model's performance across different medical disciplines and question types.

INTRODUCTION

Thoracic surgery is a medical specialty that involves the surgical treatment of organs within the thoracic cavity. This field deals with the surgical treatment of diseases and disorders affecting vital organs such as the lungs, esophagus, chest wall, and diaphragm. Thoracic surgery encompasses numerous serious health issues, including cancer, infections, trauma, and congenital anomalies. Therefore, accurate and timely interventions in the field of thoracic surgery play a critical role in enhancing patients' quality of life and improving survival rates.[1-5]

The significance of thoracic surgery is not limited to surgical techniques and applications alone. This field also requires working near complex anatomical structures and vital organs, demanding a high level of expertise and skill. Consequently, thoracic surgery training requires doctors to be well-equipped in both theoretical knowledge and practical skills.[6,7]

In recent years, the use of artificial intelligence (AI) and large language models (LLMs) in medical education and patient care has been increasing. LLMs have demonstrated significant potential in answering various medical questions, analyzing medical texts, and even providing diagnostic and treatment recommendations, thanks to their ability to learn from large datasets.[8,9] These models can be used as educational tools for medical students and doctors, aiding in the understanding of complex medical information.[10-12]

This study aimed to compare the responses of 5th-year medical students to thoracic surgery questions with those provided by an LLM, such as ChatGPT-4.

MATERIALS AND METHODS

This study was designed as a retrospective comparative analysis to compare the responses of 5th-year medical students from the the Health Sciences University Hamidiye International Faculty of Medicine thoracic surgery questions with those provided by a large language model like ChatGPT-4. The study includes 5th-year students enrolled in the Faculty of Medicine during the 2023-2024 academic year. A total of 10 students participated in the study. Ethics approval was obtained from the Kartal Dr. Lütfi Kırdar City Hospital Ethics Committee with the decision dated 26.07.2024 and numbered 2024/010.99/6/37.

To assess their knowledge in the field of thoracic surgery, participants were asked 40 multiple-choice questions. These questions were selected from those included in the medical school curriculum and prepared by the Division of Thoracic Surgery, covering theoretical knowledge and clinical applications. Each question was prepared in a five-choice multiple-choice format. The same questions were also posed to ChatGPT-4, and the model's responses were recorded. ChatGPT-4 provided answers by entering questions and options into the user interface. The model evaluated each of the selected options and determined the most appropriate answer.

ChatGPT-4 is a large language model developed by OpenAI and trained on millions of texts. The model is trained with deep learning algorithms to understand the structure and context of language using large datasets. ChatGPT-4 has the ability to generate text and answer questions on various topics using this pre-trained knowledge. The training of the model involves analyzing and processing a large amount of text data to develop the capacity to understand and produce human language.

Statistical Analysis

Statistical analyses were conducted using the Statistical Package for the Social Sciences software (Version 29, Chicago, IL, USA) for Windows. A significance level of $p<0.05$ was set for the analyses. An independent sample t-test was used to evaluate the differences in performance between the students and ChatGPT-4. The t-test was employed to assess whether the difference between the means of the two independent groups was due to chance. In this context, the difference in the number of correct answers given by the students and ChatGPT-4 was statistically analyzed.

RESULTS

In this study, the accuracy rates of responses to thoracic surgery questions given by 5th-year medical students and ChatGPT-4 were compared. When examining the scores of the 10 participating students (S1-S10) from 40 multiple-choice questions, it was observed that the students' performance ranged from 80% to 97.5% (Figure 1). The average exam performance of the students was calculated as 88.25% (SD=5.63). The accuracy rate of ChatGPT-4's responses to the same 40 questions was recorded as 95% (Figure 2). Although the individual scores of the students varied, ChatGPT-4's performance remained consistent and high. The results of the independent sample t-test indicated a statistically significant difference in performance between ChatGPT-4 and the students (t=-3.98, p=0.00088).

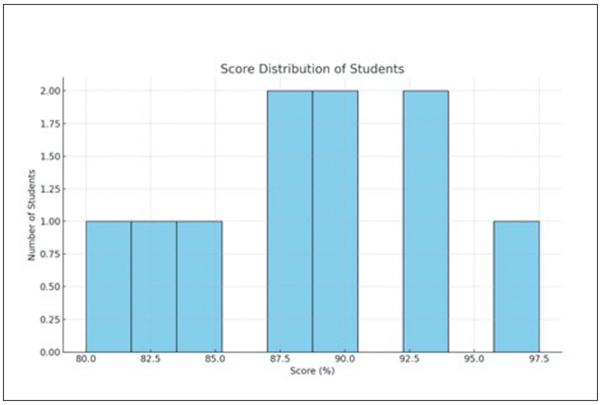

Figure 1. Score distribution of students.

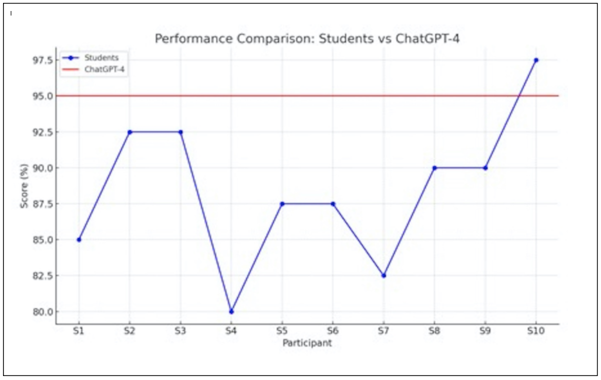

Figure 2. Performance comparison - Students vs ChatGPT-4.

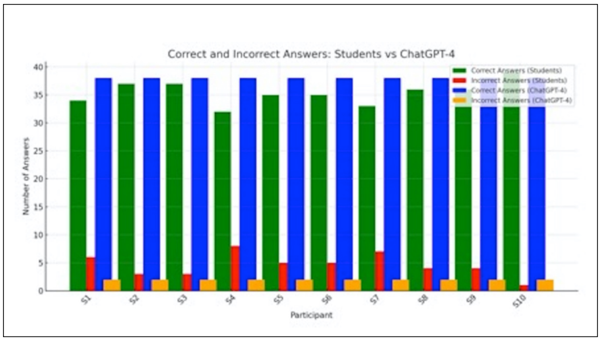

Figure 3. Correct and incorrect answers - Students vs ChatGPT-4.

Examining the distribution of student scores reveals that while the scores were spread over a wide range, they were generally high. The performance comparison showing ChatGPT-4 outperforming the students highlights the potential of the language model. The comparison of correct and incorrect responses supports that ChatGPT-4 has a higher accuracy rate than the students by providing more correct answers (Figure 3).

DISCUSSION

This study concluded that ChatGPT-4 provided significantly higher accuracy in responding to thoracic surgery questions compared to 5th-year medical students. This result indicates that LLMs can be used as valuable tools in medical education and can be effective in enhancing the knowledge level of medical students.

LLMs are deep learning systems trained on large datasets. These models learn the meaning and context of language using text data and can generate human-like text outputs. Models like GPT-4, developed by OpenAI, are highly advanced systems with millions of parameters, offering strong capabilities in language generation and understanding.[13-15]

LLMs have revolutionized the field of natural language processing (NLP) and have been used in various applications. These models demonstrate high performance in tasks such as text generation, translation, question answering, text summarization, and many more. ChatGPT, developed by OpenAI, is a chatbot built on large language models like GPT-4. ChatGPT has the ability to engage in natural and meaningful dialogues with users.[16,17]

In a study conducted in the field of neurology, the performance of GPT-4 and other large language models (Bard and Claude 2) on epilepsy examinations was evaluated. The study found that GPT-4 achieved the highest performance with an accuracy rate of 72%, while the other models showed lower performance.[18] This study highlighted the ability of large language models to answer medical exam questions, emphasizing their potential use in medical education and exam preparation.

In another study conducted in the field of oncology, the performance of LLMs on medical oncology exam questions was evaluated. The study found that a proprietary LLM, Proprietary LLM 2, achieved the highest performance with an accuracy rate of 85%. However, a significant portion of the incorrect answers was found to have a moderate to high potential for harm in clinical practice.[19] These findings suggest that LLMs can be effective in answering questions based on medical knowledge, but caution should be exercised when using them in clinical practice.

Limitations

This study has several limitations. First, the number of students participating is limited, which may hinder the generalization of the results. Second, only questions from the field of thoracic surgery were used; therefore, the results cannot be generalized to other medical fields. Third, the performance of ChatGPT-4 was evaluated only with multiple-choice questions; its performance on open-ended questions was not assessed.

Conclusion

This study demonstrates that ChatGPT-4 has a superior accuracy rate compared to medical students' performance on thoracic surgery questions. Large language models can be valuable tools in medical education and exam preparation. However, they need to be carefully evaluated and validated before being used in clinical practice. Future research should assess the performance of large language models across different medical fields and various types of questions and strive to better understand their integration into medical education.

Ethics Committee Approval

The study was approved by the Kartal Dr. Lütfi Kırdar City Hospital Ethics Committee (Date: 26.07.2024, Decision No: 2024/010.99/6/37).

Informed Consent

Retrospective study.

Peer-review

Externally peer-reviewed.

Authorship Contributions

Concept: M.B., R.D.; Design: M.B.; Supervision: R.D.; Fundings: M.B.; Materials: R.D. ; Data: M.B.; Analysis: M.B.; Literature search: M.B.; Writing: M.B., R.D.; Critical revision: M.B., R.D.

Conflict of Interest

None declared.

REFERENCES

1. Godoy LA, Hill E, Cooke DT. Social disparities in thoracic surgery education. Thorac Surg Clin 2022;32:91–102. [CrossRef]
2. Holmstrom AL, Meyerson SL. Obtaining meaningful assessment in thoracic surgery education. Thorac Surg Clin 2019;29:239–47. [CrossRef]
3. Shamji FM, Sekhon HJS, MacRae RM, Maziak DE. Guiding principles on the importance of thoracic surgical education on establishing integrated thoracic surgery program, interdisciplinary thoracic oncology conferences, and an interdisciplinary approach to management of thoracic malignancies. Thorac Surg Clin 2021;31:367–77. [CrossRef]
4. Faber LP, Liptay MJ, Seder CW. The history of the department of cardiovascular and thoracic surgery at rush. Semin Thorac Cardiovasc Surg 2016;28:687–99. [CrossRef]
5. Lou X. Thoracic surgery residents association ınaugural presidential address: Preserving the passion in cardiothoracic surgery training. J Thorac Cardiovasc Surg 2020;160:1002–3. [CrossRef]
6. Grossi S, Cattoni M, Rotolo N, Imperatori A. Video-assisted thoracoscopic surgery simulation and training: A comprehensive literature review. BMC Med Educ 2023;23:535. [CrossRef]
7. Van Schil PE. The present and future of thoracic surgery within the European Association for Cardio-Thoracic Surgery (EACTS). Eur J Cardiothorac Surg 2013;43:219–22. [CrossRef]
8. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29:1930–40. [CrossRef]
9. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ 2024;17:926–31.

10. Jowsey T, Stokes-Parish J, Singleton R, Todorovic M. Medical education empowered by generative artificial intelligence large language models. Trends Mol Med 2023;29:971–3. [CrossRef]
11. Song H, Xia Y, Luo Z, Liu H, Song Y, Zeng X, et al. Evaluating the performance of different large language models on health consultation and patient education in urolithiasis. J Med Syst 2023;47:125. [CrossRef]
12. Ahn S. The impending impacts of large language models on medical education. Korean J Med Educ 2023;35:103–7. [CrossRef]
13. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open 2023;6:e2330320. [CrossRef]
14. Nielsen JPS, von Buchwald C, Grønhøj C. Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. Acta Otolaryngol 2023;143:779–82. [CrossRef]
15. Kane MJ, King C, Esserman D, Latham NK, Greene EJ, Ganz DA. A compressed large language model embedding dataset of ICD 10 CM descriptions. BMC Bioinformatics 2023;24:482. [CrossRef]
16. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. [CrossRef]
17. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023;25:e48568.
18. Habib S, Butt H, Goldenholz SR, Chang CY, Goldenholz DM. Large language model performance on practice epilepsy board examinations. JAMA Neurol 2024;81:660–1. [CrossRef]
19. Longwell JB, Hirsch I, Binder F, Gonzalez Conchas GA, Mau D, Jang R, et al. Performance of large language models on medical oncology examination questions. JAMA Netw Open 2024;7:e2417641. [CrossRef]

Büyük Dil Modeli ve Tıp Eğitimi: Göğüs Cerrahisi Sorularında İnsan ve Yapay Zeka Yanıtlarının Değerlendirilmesi

Amaç: Bu çalışma, bir büyük dil modeli olan ChatGPT-4'ün, göğüs cerrahisi sorularına 5. sınıf tıp öğrencileri ile karşılaştırmalı olarak yanıt verme performansını değerlendirmeyi amaçlamaktadır. Çalışmanın hedefi, ChatGPT-4'ün tıp eğitiminde bir eğitim aracı olarak potansiyelini değerlendirmektir.

Gereç ve Yöntem: Çalışmada, 10 beşinci sınıf tıp öğrencisi ve ChatGPT-4'ün yer aldığı retrospektif karşılaştırmalı bir analiz yapıldı. Her katılımcı, göğüs cerrahisiyle ilgili 40 çoktan seçmeli soruyu yanıtladı. Öğrencilerin puanları, ChatGPT-4 tarafından üretilen puanlarla karşılaştırıldı. Performans farklarının anlamlılığını belirlemek için bağımsız örneklem t-testi kullanılarak istatistiksel analiz yapıldı.

Bulgular: Öğrencilerin puanları %80 ile %97.5 arasında değişmiş ve ortalama puan %88.25 (SD=5.63) olarak hesaplanmıştır. ChatGPT-4, aynı soru setinde %95 puan almıştır. T-testi sonuçları, öğrencilerin puanları ile ChatGPT-4'ün puanı arasında istatistiksel olarak anlamlı bir fark olduğunu göstermiştir (t=-3.98, p=0.00088).

Sonuç: Çalışma, ChatGPT-4'ün göğüs cerrahisi sorularına doğru yanıtlar verebildiğini ve 5. sınıf tıp öğrencilerinin performansını aştığını göstermiştir. Bu durum, büyük dil modellerinin tıp eğitiminde değerli eğitim araçları olarak potansiyelini ortaya koymaktadır. Ancak, modelin farklı tıbbi disiplinler ve soru türleri üzerindeki performansını değerlendirmek için daha fazla araştırmaya ihtiyaç vardır.

Anahtar Sözcükler: ChatGPT-4; büyük dil modelleri; göğüs cerrahisi; yapay zeka.