Can ChatGPT pass the thoracic surgery exam?

Adem Gencer and Suphi Aydin

Department of Thoracic Surgery, Afyonkarahisar Health Sciences University, Faculty of Medicine, Zafer Sağlık Külliyesi, Dörtyol Mah. 2078 Sok. No:3 A Blok, Afyonkarahisar, Turkey

ABSTRACT

Background: The capacity of ChatGPT in academic environments and medical exams is being discovered more and more every day. In this study, we tested the success of ChatGPT on Turkish-language thoracic surgery exam questions.

Methods: ChatGPT was provided with a total of 105 questions divided into seven distinct groups, each of which contained 15 questions. Along with the success of the students, the success of ChatGPT-3.5 and ChatGPT-4 architectures in answering the questions correctly was analyzed.

Results: The overall mean score of students was 12.50 ± 1.20, corresponding to 83.33%. Moreover, ChatGPT-3.5 managed to surpass students' score of 12.5 with an average of 13.57 ± 0.49 questions correctly on average, while ChatGPT-4 answered 14 ± 0.76 questions correctly (83.3%, 90.48%, and 93.33%, respectively).

Conclusions: When the results of this study and other similar studies in the literature are evaluated together, ChatGPT, which was developed for general purpose, can also produce successful results in a specific field of medicine. AI-powered applications are becoming more and more useful and valuable in providing academic knowledge.

Key Indexing Terms: Artificial intelligence (AI); ChatGPT; Large language models; Medical education; Thoracic surgery. **[Am J Med Sci 2023;366(4):291–295.]**

INTRODUCTION

Artificial intelligence (AI) is the term used to describe the use of computers and technology to simulate intelligent behavior and critical thinking comparable to that of a human being.[1] These machines have the capacity to learn, reason, and solve problems in a manner comparable to human cognition, and they can conduct tasks that typically require human intelligence, such as language comprehension, image recognition, and decision-making.[2]

One of the most important areas of artificial intelligence is natural language processing (NLP). Large Language Models (LLMs) utilize AI models that have been trained to understand and generate natural language.[2] Notable examples of this type of technology include OpenAI's chatbot ChatGPT, Google's chatbots LaMDA and Bard, and Stability AI's and OpenAI's imagery generators Stable Diffusion and Dall-E.[3]

Chat Generative Pre-Trained Transformer (ChatGPT) is a large language model developed by OpenAI (San Francisco, California, USA) and trained on a massive dataset of text from the internet up to 2021.[4] It can generate human-like responses to a variety of questions and prompts, in multiple languages and subject areas.[5] ChatGPT was released to the public on November 20, 2022, and is able to engage in complex dialog as well as provide information on nearly all topics.[6] The most recent version of ChatGPT is based on the GPT-4 architecture and was published for paid users on March 14, 2023. GPT-3.5 has 175 billion parameters, whereas GPT-4 has 100 trillion parameters and can process images.[7]

One of the key benefits of ChatGPT is its ability to provide instant, accurate, and personalized responses to a wide range of health care questions.[8] ChatGPT can be used to recommend academic journals,[9,10] discharge summaries,[11,12] teaching clinical judgment in nursing,[13] seeking colonoscopy,[14] arthroplasty,[6] pediatric palliative care,[15] plastic surgery,[16] and cancer information.[17]

Several media articles state that ChatGPT passes exams with open questions or paper assignments without revealing exact results or research methodologies. In addition, scientific research papers regarding the success of ChatGPT in various exams are also published. There are publications reporting that ChatGPT has successfully passed a Law School exam,[18] a Bar exam[19] a Master of Business Administration (MBA) exam[20] and numerous standardized admission tests in the United Kingdom.[21]

The success of ChatGPT in medical exams has been investigated in many studies. American Heart Association Life Support Exams,[4] Andalusian Health Service Thoracic Surgery Exam,[5] Antwerp University Family Medicine Exam,[22] United States Medical License Exam (USMLE),[23,24] Ophthalmic Knowledge Assessment

Program Exam,[25] Case-Based Clinical Reasoning Final Exam[26] and National Council Licensure Examination – Registered Nurse questions[27] are some of these studies.

To our knowledge, the performance of ChatGPT in the medical field in Turkish has not yet been investigated. In this study, we tested the accuracy of ChatGPT's responses to Turkish-language thoracic surgery exam questions and compared the students' performance to ChatGPT's.

METHODS

The study did not involve human or animal subjects and therefore did not require approval from the Institutional Ethical Board.

In Turkey, medical students are required to complete an internship in numerous medical departments in their fourth or fifth year. After receiving theoretical and practical training during the internship, students must successfully complete theoretical and practical exams at the end of each internship. One of these mandatory internships that must be completed is an internship in thoracic surgery. After completing a one-week thoracic surgery internship at our university, students are required to pass both theoretical and practical exams. The students' achievement scores are determined by adding the results of their theoretical and practical examinations in proportional amounts.

In our university, training in thoracic surgery is provided to 5th-semester medical students in seven distinct groups so far, with a test containing diverse questions for each group for the 2022–2023 academic year. The theoretical thoracic surgery exam consists of 15 multiple-choice questions per group. Each question has five possible answers, but only one is the correct response.

In our study, the paid "May 12 version" of ChatGPT was utilized. Both the GPT-3.5 architecture and the GPT-4 architecture were tested separately. The Safari® (Apple Inc., California, USA) web browser was used to access ChatGPT on May 18–21, 2023. On ChatGPT, a separate session has been initiated for each group. Each of the 15 questions and choices were copied from the exam file to the ChatGPT interface. In cases where ChatGPT provided an incorrect response, it was given a second opportunity and instructed to generate a new response using the "regenerate response" command. If the precise response is given in ChatGPT after repeated attempts, the result is considered successful.

The paid version of ChatGPT using GPT-3.5 and GPT-4 architectures was accessed from "https://chat.openai.com". Microsoft Excel® for Mac v16.73 (Microsoft Corp., USA) was used to analyze the results.

RESULTS

The number of correct answers given to the questions was defined as the score. Since there are 15 questions for each group, the highest score that can be obtained in a group is 15.

So far, 158 students in 7 groups have completed the thoracic surgery theoretical exam at our university in the 2022–2023 academic year. The average number of students in each group was 22.57 ± 1.29. The scores of the students are presented in Table 1 and illustrated in Fig. 1. The overall mean score of students was 12.50 ± 1.20, corresponding to 83.33%.

Table 2 shows the success of ChatGPT-3.5 and ChatGPT-4 in each group. While ChatGPT-3.5 answered 13.57 ± 0.49 questions correctly on average, ChatGPT-4 answered 14 ± 0.76 questions correctly on average (90.48%, 93.33, respectively). While there were 10 questions that ChatGPT-3.5 couldn't answer correctly out of a total of 105 questions, ChatGPT-4 could not answer 7 questions correctly (9.52% and 6.67%, respectively).

The mean scores of the students and the number of correct answers given by ChatGPT-3.5 and ChatGPT-4 were compared for each group in Fig. 2.

DISCUSSION

Every day, advancements in artificial intelligence technologies like ChatGPT continue to surprise us with their success and capabilities. As seen in the results of our study, ChatGPT-3.5 managed to surpass the students' score of 12.5 with an average of 13.57 correct answers. Moreover, ChatGPT-4 achieved the highest success with 14 correct answers for a Turkish-language thoracic surgery exam (83.33%, 90.48%, and 93.33%, respectively). This success is compatible with many studies in the literature.

Table 1. The success of students.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
N	22	24	22	21	22	25	22
Min	9	11	10	14	7	10	6
Max	15	15	13	15	14	15	14
IQR	12–13.75	11–14	12–13	15–15	11–13	12–13	9.25–12
Median	13	12	13	15	12	13	11
Mean	12.64	12.50	12.45	14.95	11.91	12.48	10.59
SD	1.55	1.26	0.72	0.21	1.73	1.33	1.95

Abbreviations: IQR, Interquartile range; SD, standard deviation.

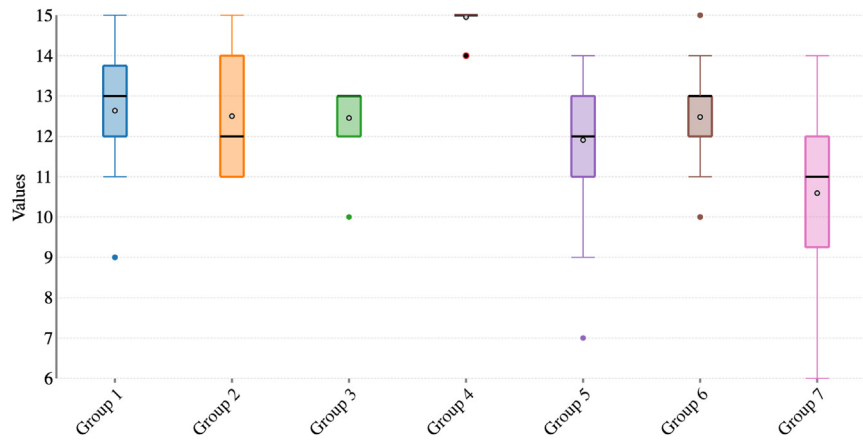

FIG. 1. Statistics of students.

Table 2. The success of students and ChatGPT.

	Students score*	GPT-3.5 score**	GPT-4 score**
Group 1	12.64 ± 1.55 (84.24%)	14 (93.33%)	15 (100%)
Group 2	12.50 ± 1.26 (83.33%)	13 (86.67%)	13 (86.67%)
Group 3	12.45 ± 0.72 (83.03%)	14 (93.33%)	14 (93.33%)
Group 4	14.95 ± 0.21 (99.68%)	13 (86.67%)	14 (93.33%)
Group 5	11.91 ± 1.73 (79.39%)	13 (86.67%)	13 (86.67%)
Group 6	12.48 ± 1.33 (83.20%)	14 (93.33%)	14 (93.33%)
Group 7	10.59 ± 1.95 (70.61%)	14 (93.33%)	15 (100%)
Overall*	**12.50 ± 1.20 (83.33%)**	**13.57 ± 0.49 (90.48)**	**14.00 ± 0.76 (93.33)**

* Mean, standard deviation, and percentage.
** Score and percentage.

ChatGPT answered 58.90% of the thoracic surgery questions correctly on the exam administered by the Andalusian Health Service.[5] In another study, ChatGPT achieved 68% and 64% accuracy in the 25-question American Heart Association (AHA) Basic Life Support (BLS) exams and 68.4% and 76.3% accuracy in the two 38-question Advanced Cardiovascular Life Support (ACLS) exams and did not reach the passing threshold for any of the exams.[4]

It is obvious that ChatGPT is getting better with each update, and new versions are having better success. Namkee et al.[28] tested the performance of ChatGPT

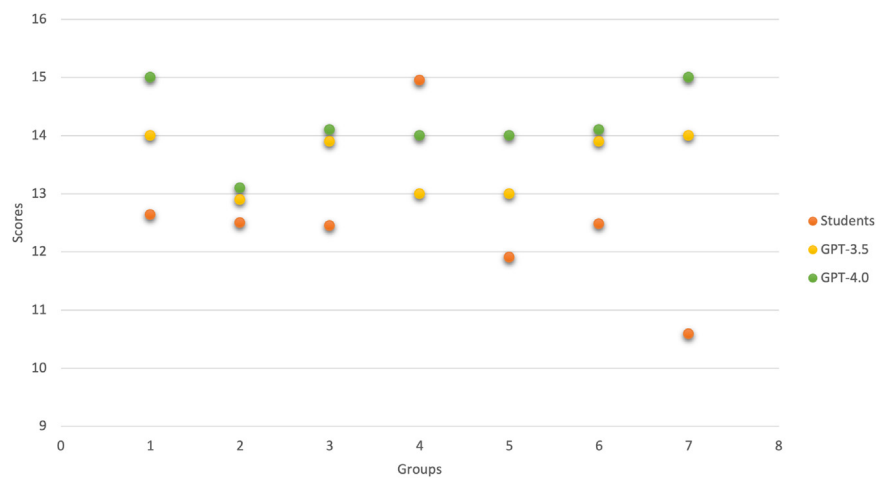

FIG. 2. Mean scores of ChatGPT and students.

using questions from the Korean general surgery board exam and observed that the model achieved an accuracy of 76.4% with GPT-4 and 46.8% with GPT-3.5. In another study in ophthalmology,[25] the legacy model achieved 55.8% accuracy on the Basic and Clinical Science Course (BCSC) set and 42.7% on the OphthoQuestions set. With ChatGPT Plus, accuracy increased to 59.4% ± 0.6 and 49.2% ±1.0, respectively. Similarly, in our study, ChatGPT-4 achieved much better success than ChatGPT-3.5 (93.33%, 90.48%, respectively). It can be thought that the high rates we obtained in the study are due to the diversity in the question patterns and the updates in ChatGPT. Moreover, during our research, we discovered that ChatGPT-4 provides more flexible, more interpretive, and more accurate answers than its predecessor.

Gilson et al. and Kung et al.[23,24] tested the success of ChatGPT in USML (United States Medical Licensing) exams. Gilson et al.[24] demonstrated that ChatGPT's accuracy for the four data sets AMBOSS-Step1, AMBOSS-Step2, NBME-Free-Step1, and NBME-Free-Step2 was 44%, 42%, 64.4%, and 57.8%, respectively. In a similar study, Kung et al. (23) reported that ChatGPT accuracy for USMLE Steps 1, 2CK, and 3 was 64.5% / 41.2%, 52.4% / 49.5%, and 65.2% / 59.8%, respectively, and proved that ChatGPT can be successful on USML multiple choice questions and can exceed the passing grade. Despite the absence of a specific passing grade for the thoracic surgery theoretical exam at our university, the over 60% success rate of ChatGPT can be interpreted as passing the exam for us.

Due to the nature of artificial intelligence, ChatGPT's answers to the same question may be different each time, and it may make a different interpretation in each attempt. According to Eric et al.'s[26] study on the clinical reasoning exam, ChatGPT's performance ranged between 56% and 81% after 20 repetitions of the same case. Similarly, in our study, the answers given by ChatGPT differed each time, and ChatGPT was sometimes able to reach the correct answer after a few tries. Moreover, sometimes ChatGPT repeated its answers with great confidence and with clear explanations, even if it was a completely wrong answer. This is technically called a hallucination.[22]

A particular limitation of ChatGPT is its reliance on a static database with a specific knowledge termination date (November 2021) for both ChatGPT-3.5 and ChatGPT-4. Google® has recently released Bard, an additional AI chatbot. This AI chatbot has an infrastructure that is continuously updated and will soon be compatible with the Google Search Engine.[29] However, BARD or other artificial intelligence-supported chatbots do not have Turkish support yet. Since the dataset we used in the study is in Turkish, we could only perform this study on ChatGPT. In the near future, more comprehensive and comparative studies will be possible with other popular chatbots supporting local languages more.

CONCLUSIONS

Despite its inconsistencies and limitations, ChatGPT has achieved a high success rate in the Turkish-language thoracic surgery exam for medical faculty students. Even though ChatGPT's efficacy is inconsistent, it can be utilized for academic purposes. When the results of this study and other similar studies in the literature are considered together, it can be concluded that as a general large language model, ChatGPT (especially ChatGPT-4) possesses at least the knowledge of an average medical student. The future development of ChatGPT and other AI-supported models will produce more precise, significant, and accurate outcomes. In addition, the use of more reliable data sources such as scientific literature in natural language processing models will increase the success of large language models in specific areas of medicine. Increasingly, artificial intelligence-supported applications are proving more and more useful and valuable in providing academic information. However, in our opinion, it should always be questioned whether the information produced and predicted by artificial intelligence is reliable in terms of academic and medical information.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Adem Gencer: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft. **Suphi Aydin:** Investigation, Writing – review & editing.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

1. **Amisha Malik P, Pathania M, et al.** Overview of artificial intelligence in medicine. *J Fam Med Prim Care*. 2019;8(7):2328.
2. **Aubignat M, Diab E.** Artificial intelligence and ChatGPT between worst enemy and best friend: the two faces of a revolution and its impact on science and medical schools. *Rev Neurol (Paris)*. 2023. [Internet]Mar [cited 2023 May 22]; Available from: https://linkinghub.elsevier.com/retrieve/pii/S0035378723008809.
3. **Harrer S.** Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*. 2023;90: 104512.
4. **Fijačko N, Gosak L, Štiglic G, et al.** Can ChatGPT pass the life support exams without entering the American heart association course? *Resuscitation*. 2023;185: 109732.
5. **Fuentes-Martín Á, Cilleruelo-Ramos Á, Segura-Méndez B, et al.** Can an artificial intelligence model pass an examination for medical specialists? *Archivos de Bronconeumología*. 2023. Mar;S0300289623001163.

6. **Dubin JA, Bains SS, Chen Z, et al.** Using a Google web search analysis to assess the utility of ChatGPT in total joint arthroplasty. *J Arthroplasty*. 2023. Apr;S0883540323003522.
7. **Ray PP.** ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys Syst*. 2023;3:121–154.
8. **Odom-Forren J.** The role of ChatGPT in PeriAnesthesia Nursing. *J Perianesth Nurs*. 2023;38(2):176–177.
9. **Haman M, Školník M.** Exploring the capabilities of ChatGPT in academic research recommendation. *Resuscitation*. 2023;187:109795.
10. **Alser M, Waisberg E.** Concerns with the usage of ChatGPT in academia and medicine: a viewpoint. *Am J Med Open*. 2023;9:100036.
11. **Byrne MD.** Generative artificial intelligence and ChatGPT. *J Perianesth Nurs*. 2023;38(3):519–522.
12. **Patel SB, Lam K.** ChatGPT: the future of discharge summaries? *The Lancet Digital Health*. 2023;5(3):e107–e108.
13. **Seney V, Desroches ML, Schuler MS.** Using ChatGPT to teach enhanced clinical judgment in nursing education. *Nurse Educ*. 2023;48(3):124—124.
14. **Lee TC, Staller K, Botoman V, et al.** ChatGPT answers common patient questions about colonoscopy. *Gastroenterology*. 2023. May;S0016508523007047.
15. **Almazyad M, Aljofan F, Abouammoh NA, et al.** Enhancing expert panel discussions in pediatric palliative care: innovative scenario development and summarization with ChatGPT-4. *Cureus*. 2023. [Internet]Apr 28 [cited 2023 May 22]; Available from: https://www.cureus.com/articles/152782-enhancing-expert-panel-discussions-in-pediatric-palliative-care-innovative-scenario-development-and-summarization-with-chatgpt-4.
16. **Seth I, Cox A, Xie Y, et al.** Evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. *Aesthet Surg J*. 2023:sjad140.
17. **Hopkins AM, Logan JM, Kichenadasse G, et al.** Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectrum*. 2023;7(2):pkad010.
18. **Choi JH, Hickman KE, Monahan A, et al.** ChatGPT goes to law school. *SSRN J*. 2023. [Internet][cited 2023 May 22]; Available from: https://www.ssrn.com/abstract=4335905.
19. **Bommarito MJ, Katz DM.** GPT takes the bar exam. *SSRN J*. 2022. [Internet][cited 2023 May 22]; Available from: https://www.ssrn.com/abstract=4314839.
20. **Eke DO.** ChatGPT and the rise of generative AI: threat to academic integrity? *J Responsible Technol*. 2023;13:100060.
21. **Giannos P, Delardas O.** Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. *JMIR Med Educ*. 2023;9:e47737.
22. **Morreel S, Mathysen D, Verhoeven V, Aye AI!.** ChatGPT passes multiple-choice family medicine exam. *Med Teach*. 2023;45(6):665–666.
23. **Kung TH, Cheatham M, Medenilla A, et al.** Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. Dagan A, editor *PLOS Digit Health*. 2023;2(2): e0000198.
24. **Gilson A, Safranek CW, Huang T, 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.
25. **Antaki F, Touma S, Milad D, et al.** Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci*. 2023 100324.
26. Strong E., DiGiammarino A., Weng Y., et al. Performance of ChatGPT on free-response, clinical reasoning exams [Internet]. Medical Education; 2023 Mar [cited 2023 May 22]. Available from: http://medrxiv.org/lookup/doi/10.1101/2023.03.24.23287731.
27. **Qi X, Zhu Z, Wu B.** The promise and peril of ChatGPT in geriatric nursing education: what we know and do not know. *Aging Health Res*. 2023;3(2):100136.
28. **Oh N, Choi GS, Lee WY.** ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Ann Surg Treat Res*. 2023;104(5):269.
29. **Sanchez-Ramos L, Lin L, Romero R.** Beware of references when using ChatGPT as a source of information to write scientific articles. *Am J Obstet Gynecol*. 2023. Apr;S000293782300234X.

Submitted June 7, 2023; accepted August 2, 2023.

Corresponding author. Adem Gencer, Department of Thoracic Surgery, Afyonkarahisar Health Sciences University, Faculty of Medicine, Zafer Sağlık Külliyesi, Dörtyol Mah. 2078 Sok. No:3 A Blok, Afyonkarahisar, Turkey (E-mail: dr.ademgencer@gmail.com).