ORIGINAL ARTICLE

VIA MEDICA

Reshaping medical education: Performance of ChatGPT on a PES medical examination

Simona Wójcik[1], Anna Rulkiewicz[1], Piotr Pruszczyk[2], Wojciech Lisik[3],
Marcin Poboży[4], Justyna Domienik-Karłowicz[1, 2]

[1]LUX MED Llc, Warsaw, Poland
[2]Department of Internal Medicine and Cardiology with The Center for Diagnosis and Treatment
of Thromboembolism, Medical University of Warsaw, Poland
[3]Department of General and Transplantation Surgery, Medical University of Warsaw, Poland
[4]Cichowski Pobozy Healthcare Facility, Maciejowice, Poland

Abstract

Background: *We are currently experiencing a third digital revolution driven by artificial intelligence (AI), and the emergence of new chat generative pre-trained transformer (ChatGPT) represents a significant technological advancement with profound implications for global society, especially in the field of education.*

Methods: *The aim of this study was to see how well ChatGPT performed on medical school exams and to highlight how it might change medical education and practice. Recently, OpenAI's ChatGPT (OpenAI, San Francisco; GPT-4 May 24 Version) was put to the test against a significant Polish medical specialization licensing exam (PES), and the results are in. The version of ChatGPT-4 used in this study was the most up-to-date model at the time of publication (GPT-4). ChatGPT answered questions from June 28, 2023, to June 30, 2023.*

Results: *ChatGPT demonstrates notable advancements in natural language processing models on the tasks of medical question answering. In June 2023, the performance of ChatGPT was assessed based on its ability to answer a set of 120 questions, where it achieved a correct response rate of 67.1%, accurately responding to 80 questions.*

Conclusions: *ChatGPT may be used as an assistance tool in medical education. While ChatGPT can serve as a valuable tool in medical education, it cannot fully replace human expertise and knowledge due to its inherent limitations.* (Cardiol J 2024; 31, 3: 442–450)

Keywords: ChatGPT, innovations, artificial intelligence, AI in medicine, health IT, medical education, language processing, virtual teaching assistant

Introduction

Over the past half-century, society has experienced two digital revolutions. The first was in communications, taking us from analogue phones to the Internet. The second revolution, centered around computation, introduced personal computers and smartphones. Now, we have entered a third digital revolution built around artificial intelligence (AI).

While there have been various incarnations of AI systems, AI assistants, and chatbots over the years, including the notorious yet notable ELIZA AI chatbot from 1966, none have exerted as much influence or possessed as much intuitive an interface

as the chat generative pre-trained transformer (ChatGPT) [1]. Owing to its simplicity, remarkable capability to address virtually any query, and capacity to generate a diverse range of content, ChatGPT has emerged as a formidable new disruptive technology, with some envisioning it as the successor to Google or a revolutionary force within the Internet realm [2]. ChatGPT signifies a pivotal new technological development with far-reaching implications for the global society, particularly in education [3].

When the Stanford Accelerator for Learning and the Stanford Institute for Human-Centered AI were in the initial stages of organizing the inaugural AI + Education Summit in 2022, public discourse and controversy around AI had not yet escalated to the levels observed today. Even so, intensive research was already underway across Stanford University to understand AI's vast potential, including transforming education as we know it. By the time the summit was held on February 15, ChatGPT had reached over 100 million unique users, making it the fastest-growing consumer application in history and certainly in educational settings [4].

Generative AI is a powerful tool that can help address pressing societal challenges and business problems by augmenting, expanding, and extending the human experience rather than replicating or replacing it.

ChatGPT-4, created by OpenAI, is an advanced language model at the forefront of technology. It stands as the fourth evolution in the ChatGPT series, specifically engineered to produce text responses that resemble human conversation. ChatGPT-4 expands on the progress made by its earlier versions by enhancing its grasp of language, its text generation capabilities, and its capacity to furnish pertinent responses in diverse subject areas and assignments. This model marks a noteworthy achievement in the evolution of conversational AI, finding practical uses in customer assistance, content creation, and a wide array of fields reliant on natural language understanding and processing.

The far-reaching consequences of ChatGPT, among other Large Language Models (LLMs), can be described as a paradigm shift in academia and healthcare practice. Discussing its potential benefits, future perspectives, and limitations appears timely and relevant

ChatGPT: Implications for education

This initial panic from the education sector was understandable. ChatGPT, available to the public via a web app, can answer questions and generate slick, well-structured blocks of text several thousand words long on almost any topic it is asked about, from string theory to Shakespeare. Each essay it produces is unique, even when given the same prompt again, and its authorship is (practically) impossible to spot. ChatGPT would undermine how we test what students have learned, a cornerstone of education [5].

A recent study conducted by McGee (2023) [6] shed light on the widespread utilization of ChatGPT among American college students for academic purposes. The findings revealed that 89% of students rely on ChatGPT to assist them in completing their homework assignments. Notably, approximately 53% of these students also employ ChatGPT as a tool for writing academic papers. The impact of ChatGPT extends beyond the realm of homework, as 48% of students reported using the tool during exams, indicating its potential to aid in exam preparation. Furthermore, 22% of students utilize ChatGPT to generate paper outlines, demonstrating its versatility in facilitating the writing process [6]. It is noteworthy that the use of ChatGPT not only assists students in completing their assignments but also contributes to their academic success. Some students have demonstrated their ability to achieve high scores while employing this AI chatbot as a resource. In a separate survey conducted in Japan, a research group found that approximately 32% of university students acknowledged using the AI chatbot ChatGPT. Many of these students stated that using ChatGPT enhances their thinking abilities. Among the participants, students in the science, technology, and agriculture departments showed the highest usage rates, accounting for 45.5% overall.

Interestingly, male students utilized ChatGPT more frequently, with 44.8%, compared to female students at 27.1%. When asked about the influence of using ChatGPT on their thinking abilities, 70.7% of participants reported a positive or somewhat positive impact. Conversely, 15.4% indicated a negative or rather negative influence. These findings emphasize the potential benefits and varied perceptions of utilizing ChatGPT among university students.

ChatGPT: A "game changer in assessments?"

Sometimes, it is suggested that ChatGPT is a "game changer" with the potential to end some traditional sorts of assignments and assessments. The ChatGPT answer is sufficient for the student to receive the minimum passing grade. Some educators even claim that the chatbot answers somewhat

better than the "average" Master of Business Administration [7]. A University of Minnesota Law School professor's team found that ChatGPT would underperform the average law school student but could skirt by with a passing grade on final exams in four courses [7]. The researchers found that the bot would be a mediocre law student, but it could assist students with their assignments [8, 9]. This AI system has excellent potential to help students train and prepare so that they can use it to study for exams and improve their knowledge.

In the area of healthcare education, ChatGPT also has massive transformative potential. The need to rethink and revise the current assessment tools in health care education comes in light of ChatGPT's ability to pass reputable exams (e.g., United States Medical Licensing Examination [USMLE]) [7–11]. We specifically question what education should be offered to students and what adjustments to education are needed to fulfil the needs of students. Therefore, the current review aims to explore the future perspectives of ChatGPT as a prime example of LLMs in healthcare education.

Thus, the aim herein, was to see how well ChatGPT performed on medical school exams to highlight how it might change medical education and practice. When asked hundreds of questions on the standard tests that United States universities use to grant or deny licensure to practice medicine, ChatGPT proved more than capable of passing the exam. The experiment was conducted by a group of researchers from several United States universities that put the AI program through a standardized exam, the USMLE, without special training or tutoring. The exam consists of three sections with questions covering most medical subjects, from biochemistry to diagnostic reasoning to bioethics. Before the test, the researchers reviewed the questions and deleted those related to interpreting radiological and other images, leaving 350 out of 376 questions. As the researchers explain in the journal PLoS Digital Health, ChatGPT achieved a score ranging from 52.4% to 75% correct answers, which is notable.

Considering the success rate is around 60%, the prospects of success for the AI system are quite high. In addition, ChatGPT had an unusually high percentage of coherent responses — 94.6%. No answers contradicted each other, and in 88.9% of its solutions, it offered explanations that were not trivial or obvious, revealing some intuition [4]. It is worth noting that a similar system developed by the world's largest database of scientific publications, PubMed (PubMedGPT), could only answer up to 50% of the answers correctly in the same USMLE exam, far behind ChatGPT [11].

In a study conducted by Antaki et al. [12] in the field of ophthalmology, the performance of ChatGPT was assessed and found to be comparable to that of an average first-year resident. This demonstrates the remarkable capabilities of large language models like ChatGPT.

Furthermore, recent studies have revealed the potential of ChatGPT to perform well on radiology board-style examinations. In a prospective exploratory study, 150 multiple-choice questions were designed to resemble radiology examinations' style, content, and difficulty level. These questions were categorized based on their thinking levels (lower-order and higher-order) and topics (physics and clinical). The higher-order thinking questions were further classified into specific types, such as the description of imaging findings, clinical management, application of concepts, calculation and classification, and disease associations. Despite not receiving specific training in radiology, ChatGPT demonstrated impressive performance, nearly passing the radiology board-style examination without images. It excelled in answering lower-order thinking questions and clinical management questions. However, it faced challenges with higher-order thinking questions that involved the description of imaging findings, calculation and classification, and application of concepts. The study conducted a univariable analysis and reported that ChatGPT correctly answered 69% of the questions (104 out of 150) [13].

Another study aimed to evaluate the performance of ChatGPT on the Plastic Surgery In-Service Examination and compare it to the performance of plastic surgery residents nationally. The Plastic Surgery In-Service Examinations from 2018 to 2022 were used as the source of questions, and ChatGPT was given access to both the stem and multiple-choice options for each question. The performance of ChatGPT on the 2022 examination was then compared to the performance of plastic surgery residents across different training levels. The final analysis included 1129 questions, and ChatGPT answered 630 (55.8%) correctly. ChatGPT achieved its highest score on the 2021 exam (60.1%) and the comprehensive section (58.7%). When compared to the performance of plastic surgery residents in 2022, ChatGPT would rank in the 49th percentile for first-year integrated plastic surgery residents, 13th percentile for second-year residents, 5th percentile for third- and fourth-year

residents, and 0th percentile for fifth- and sixth-year residents [14].

Consequently, within the purview of proper academic guidance, ChatGPT could be beneficial in advancing communication competencies within the healthcare education sector [15, 16].

Methods

Recently, the present study tested OpenAI's ChatGPT (OpenAI, San Francisco; GPT-4 May 24 Version) against a significant Polish medical specialization licensing exam (PES), and the results are now available. The version of ChatGPT used in this study was the most up-to-date model at the time of publication (GPT-4). ChatGPT answered questions from June to 28, 2023, and on June 30, 2023.

Państwowy Egzamin Specjalizacyjny (PES) is a national specialized examination in Poland. It is regulated by the Articles 16rb-16x of the Act of December 5, 1996, on the professions of physician and dentist (consolidated text: Journal of Laws of 2018, item 617, as amended). A physician can obtain a specialist title in a specific field of medicine: after completing specialist training and passing the National Specialization Examination, referred to as "PES" (written and oral parts) or after recognition of an equivalent specialist title obtained abroad. The exams cover the scope of completed specialist training. The tests and test questions are developed and established by the Center for Medical Education (CEM) in consultation with the national consultant responsible for the respective field of medicine [17]. The entrance test in the PES consists of solving 120 questions with five answer options, of which only one is correct. The test portion of PES is considered passed with a positive result when the physician achieves at least 60% of the maximum possible score.

In the case of this study, ChatGPT was deployed to address the cardiology section of the examination (Spring 2023) [18]. The CEM has been publishing test questions and the correct answers after they have been used on a given exam within 7 days of conducting that exam since the Spring 2023 session. For this study, individualized prompts for each question were not employed; instead, an initial investigation was conducted to identify the prompts that yielded the most favorable responses. This practice, known as "prompt engineering," has gained substantial interest since the integration of language models like ChatGPT. The ethical implications of AI tools being potentially misused

for cheating or gaining unfair advantages in admission tests raise significant concerns. In this study, the focus was on evaluating ChatGPT as an educational tool for test preparation, emphasizing the importance of responsible usage and discouraging its application during actual exams.

Beyond evaluating the overall performance, ChatGPT was also tasked with estimating the "Level of correctness" (LOC) for each answer provided. Accuracy assessment was carried out by comparing ChatGPT's responses with the answer key derived from the CEM question banks. Five randomly selected questions, for which GPT-4 presented an explanation and rationale for the chosen answer, were subjected to further analysis. The responses were collected and archived for subsequent research. Two expert cardiology academicians examined the replies on a zero to five scale.

Results

ChatGPT marks a significant improvement in natural language processing models on the tasks of medical question answering. In June 2023, the performance of ChatGPT was assessed based on its ability to answer a set of 120 questions, where it achieved a correct response rate of 67.1%, accurately responding to 80 questions.

The number of questions to which ChatGPT-4 provided correct answers is shown in green, while the questions where the answers did not match the key are marked in red. The level of confidence that ChatGPT-4 assigned to each question is marked on the y-axis (Fig. 1).

Although this success rate may be considered a passing score, it is important to acknowledge the narrow margin by which the AI chatbot met the criteria. Internal information to the question was present in 99.2% (119/120) of all questions. Significantly, the unvarying use of assertive language by GPT-4 can potentially mislead, even when it is incorrect, particularly in crucial sectors like healthcare. As such, employing a more nuanced language that mirrors the degree of confidence could be more suitable and safer.

Additionally, it should be highlighted that ChatGPT's capacity to provide logic and informational context across most answers. These facts, taken together, make a compelling case for the potential applications of ChatGPT as an interactive medical education tool to support learning. ChatGPT has been known to exhibit tendencies to "hallucinate" by generating fabricated facts. However, in this study, it was generally observed that ChatGPT
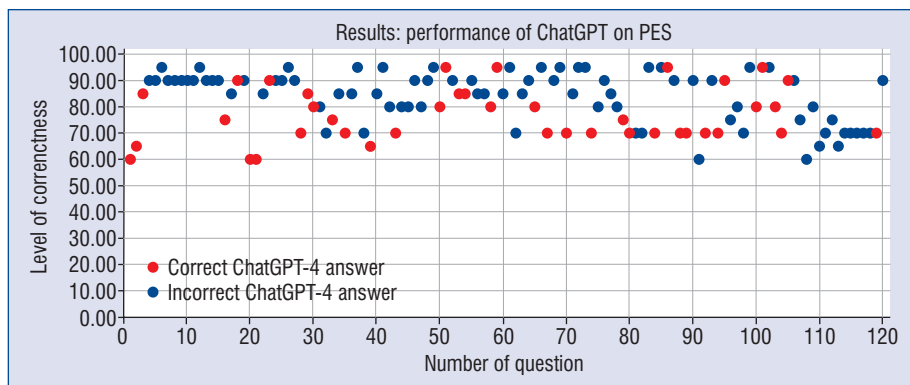
Figure 1. The results of the conducted experiment: performance of ChatGPT on PES; ChatGPT — chat generative pre-trained transformer; PES (*Państwowy Egzamin Specjalizacyjny*) — national specialization examination in Poland.

refrained from such tendencies, possibly due to the prompts provided, which explicitly instructed ChatGPT not to fabricate cases and, when necessary, provide a specific universe of issues for the model to work with.

A few of the 120 prompts with correct answers and explanations by ChatGPT-4 are present in Table 1.

Nevertheless, it is vital to acknowledge the limitations of this study. Firstly, ChatGPT was initially trained on a dataset of information generated on or before 2021. Consequently, the prompts used with the model may only contain information available up until that date, so they do not comply with actual European Society of Cardiology guidelines. Secondly, due to the closed nature of the experiment and the absence of a public application programming interface, fine-tuning the model specifically for task-related purposes is not currently feasible. Lastly, ChatGPT receives regular updates resulting from continued training on user-provided inputs. It is reasonable to hypothesize that each subsequent iteration of the model will not significantly impair its performance on the outlined task and may improve performance.

Discussion

Multiple-choice evaluations are the predominant form of assessment in medical education [19]. Medical students commonly resort to third-party question banks to prepare for licensure examinations. These resources typically explain answers, revealing specific teaching points related to each question. Concurrently, medical institutions often disseminate practice exams to facilitate students' preparations and steer their study strategies. Chat-GPT could function as a "virtual pedagogical aide" [20], offering insights into each question, providing feedback, and elucidating concepts that students might struggle with. Further, it could also be harnessed to create interactive knowledge verification questions, thereby reinforcing students' conceptual comprehension in an engaging manner [21].

Academic teachers will transition from being information custodians to facilitators. Presently, the mandate for educators is to guide students in locating information, discern trustworthy sources from unreliable ones, and discriminate between the two. Nonetheless, this innovative approach demands time and resources, which many educators may not readily have, given their heavy workloads, limited resources, and adherence to strict performance measures. This makes the full exploitation of opportunities offered by chatbots challenging. Artificial intelligence stands on the precipice of transforming education in ways yet to be fully anticipated. The advent of ChatGPT underscores educators' need to remain flexible and prepared to acclimate swiftly to rapid and substantial technological advancements [22].

ChatGPT may be used as an assistance tool in medical education. However, it cannot be considered as a replacement for human capability and knowledge, as it is still plagued by the limitations that AI faces. However, we are witnessing a quantum leap in information technology, machine learning, and AI. At this pace, it will transform our approach to medical education and clinical management in a few days. These changes should be seen and adopted with an open mind to make good use of them for improving medical education and clinical management. As the efficacy of ChatGPT augments, potentially via strategic prompts and

Table 1. Examples of prompts with correct answers and explanations by ChatGPT-4 (correct/incorrect).

Number of question from CEM	Prompt (content of the question)	Correct answer	Explanation by ChatGPT-4
15	In a patient with heart failure, in NYHA functional class III despite optimal therapy, with persistent atrial fibrillation, with a left ventricular ejection fraction of 30%:consider implanting a resynchronization pacemaker (CRT-P or CRT-D) regardless of the width of their own QRS complexes; consider implanting CRT-P or CRT-D if the width of their own QRS complexes is > 130 ms; an ICD should be implanted without CRT for primary prevention of sudden death; in patients who have already undergone CRT-P or CRT-D implantation, additional consideration (class IIa) should be given to ablation of the atrioventricular junction if the percentage of biventricular stimulation is < 90%; in those who have symptoms of atrial fibrillation and an uncontrollable, despite maximum doses of drugs, heart rate, CRT should be implanted regardless of the width of their own QRS complexes. The correct answer is: A. 1, 4, 5 B. Only 3 C. 2, 4, 5 D. 3, 4, 5 E. 4, 5	C	C. 2, 4, 5 Resynchronization pacing (CRT) is one of the fundamental tools in the treatment of patients with advanced heart failure. However, not every patient qualifies for this type of therapy. The width of the QRS complexes is a key factor that affects the effectiveness of CRT therapy. CRT is indicated when the QRS complex width is above 130 ms (2). When the width of the QRS complexes is smaller, there is not enough conduction delay for CRT therapy to be effective. For patients with heart failure and persistent atrial fibrillation, whose percentage of biventricular stimulation is below 90%, atrioventricular junction ablation may be beneficial (4). This procedure aims to ensure constant biventricular stimulation, which is key to the effectiveness of CRT therapy. In relation to patients with heart failure who have symptoms of atrial fibrillation and an uncontrollable, despite maximum doses of drugs, heart rate, an individual approach is necessary. In such cases, CRT therapy may be considered regardless of the width of the QRS complexes (5). That's because with persistent atrial fibrillation, the goal is heart rate control. In some cases, CRT can help achieve this goal, even if the width of the QRS complexes is within the normal range. All this underscores how important it is to carefully consider all clinical factors before deciding on CRT therapy for each individual patient. With proper selection, CRT is an effective tool that can significantly improve the quality of life for patients with advanced heart failure.
7	Which of the following questions is most appropriate to ask a 60-year-old woman, previously untreated cardiologically, with overweight and registered sinus bradycardia (HR 52/min) during the consultation, in order to determine the cause of bradycardia? A. Does she snore at night and has she been diagnosed for sleep apnea syndrome (SAS)? B. Has she been diagnosed with an upper gastrointestinal tract infection? C. Has she had a tick bite and does she remember seeing erythema migrans on her skin? D. Is she being treated for depression? E. All of the above questions should be asked	E	E. All of the above questions should be asked. Bradycardia, or a slow heart rate, can have many causes. Direct examination of the patient and a thorough medical history are key to determining the true source of the problem. A. SAS, or sleep apnea syndrome, is a condition where the patient experiences frequent pauses in breathing during sleep. It is a common problem among overweight and obese individuals and can lead to a slower heart rate. B. An upper gastrointestinal tract infection can lead to electrolyte disturbances which can affect the heart's function. C. Lyme disease, which is transmitted by ticks, can cause heart disturbances including bradycardia. D. Some medications used in the treatment of depression can cause bradycardia as a side effect. So, all of these questions are pertinent in the context of bradycardia. Without answers to these questions, it might be difficult to accurately determine the cause of the problem.

$\rightarrow$

Table 1 (cont.). Examples of prompts with correct answers and explanations by ChatGPT-4 (correct/incorrect).

Number of question from CEM	Prompt (content of the question)	Correct answer	Explanation by ChatGPT-4
49	Which of the following patients can undergo extracorporeal membrane oxygenation (ECMO)? A. A 45-year-old patient in cardiogenic shock — in the case of ineffectiveness of previous treatment methods — including the use of inotropic drugs in increased doses B. A 70-year-old patient in terminal cancer C. A 75-year-old patient with another infectious exacerbation of chronic heart failure — before the use of inotropic drugs D. An 84-year-old patient with low cardiac output syndrome after cardiac surgery, mechanically ventilated for 10 days with $FiO_2 > 95\%$ and plateau pressure > 40 cm H_2O. E. A 75-year-old patient with heart failure treated with high doses of immunosuppressive drugs due to lupus and a history of heparin-induced thrombocytopenia	A	A. In a 45-year-old patient in cardiogenic shock — in the case of ineffectiveness of previous treatment methods — including the use of inotropic drugs in increased doses. Extracorporeal membrane oxygenation (ECMO) is an intensive care method used in the treatment of severe, potentially reversible respiratory and (or) circulatory failures that are unresponsive to standard treatment methods. Cardiogenic shock that does not respond to conventional treatment methods, such as inotropic drugs, may be an indication for ECMO. The other described situations are usually not considered typical indications for ECMO. Patients in the terminal stage of cancer (B), with advanced comorbidities, such as severe heart failure, and especially multi-organ failure (C, E), and patients who are already undergoing intensive mechanical ventilation therapy with high parameters (D), may not be good candidates for ECMO due to the risk of complications and limited potential for improvement.
57	42-year-old man was admitted to the cardiology department on an emergency basis due to strong atypical chest pain occurring periodically for about 7 days. He also suffers from ulcerative colitis. Laboratory tests revealed slightly elevated markers of cardiac muscle damage. Sinus tachycardia was recorded in the ECG, with no significant changes in the ST segment. Transthoracic echocardiography showed normal dimensions of the heart chambers, a competent valve apparatus, without segmental contractility disorders, LVEF 52%. Indicate the procedure that will facilitate the diagnosis and treatment of the patient: 1. You should ask if the patient is using sulfasalazine for ulcerative colitis 2. You should ask about a recent viral infection 3. You should perform coronary angiography urgently 4. You should determine CRP 5. If the results suggest a diagnosis of acute pericarditis, ibuprofen should be administered at a dose of 2×400 mg The correct answer is: A. Only 1 B. 1, 2, 3 C. 1, 2, 4 D. 2, 4 E. All listed	C	C. 1, 2, 4. Given the presented symptoms and accompanying diseases, the most important steps to establish a diagnosis seem to be: Asking about medications the patient may be taking. Sulfasalazine, used in the treatment of ulcerative colitis, is a known drug that can cause cardiomyopathy. Asking about a recent viral infection. Viruses are often the cause of myocarditis, a condition that can cause symptoms similar to those described by the patient. Determining CRP (C-reactive protein). An increase in CRP is often observed in inflammatory conditions and can help confirm or rule out an inflammatory process, such as myocarditis. While coronary angiography (3) is important for assessing coronary artery disease, the described symptoms and context do not indicate it as an urgent diagnostic step. Similarly, initiating ibuprofen treatment (5) may be premature until we confirm a diagnosis of a condition that would require it, such as pericarditis.

Table 1 (cont.). Examples of prompts with correct answers and explanations by ChatGPT-4 (correct/incorrect).

Number of question from CEM	Prompt (content of the question)	Correct answer	Explanation by ChatGPT-4
1	Indicate the most beneficial course of action for a 70-year-old patient with sinus venosus type ASD, a pulmonary-to-systemic flow ratio (Qp/Qs) > 1.5, and an elevated pulmonary resistance value during heart catheterization (3–5 WU): A. The patient should be subjected to immediate surgical treatment (class IIa) B. The patient should only be subjected to surgical treatment in a situation where the benefits of surgery outweigh its risk C. The patient requires targeted pharmacotherapy for pulmonary arterial hypertension for several months and re-evaluation of resistance values in the pulmonary artery bed D. In such a situation, it is recommended to abandon surgical proceedings in favor of implanting a so-called "fenestrated" type of occluder E. Surgical proceedings are ruled out. It is recommended to initiate vasodilatory agents, including those from the group of endothelin receptor antagonists (ERA)	B	C. The patient requires targeted pharmacotherapy for pulmonary arterial hypertension for several months and re-evaluation of resistance values in the pulmonary artery bed. Even though it seems the patient has a clear indication for surgery (sinus venosus type ASD with Qp/Qs > 1.5), elevated pulmonary vascular resistance presents a significant challenge. The operation may increase the load on the right ventricle of the heart, which can lead to heart failure. Before proceeding with the operation, the optimal approach would be to implement targeted pharmacotherapy to reduce pulmonary arterial hypertension. This could include drugs such as type 5 phosphodiesterase inhibitors, endothelin receptor antagonists, or prostanoids. After several months of therapy, the pulmonary vascular resistance can be reassessed. If the resistance has decreased, the patient may be more suitable for surgery. However, it is important to remember that every patient is different, and the final decision depends on a risk-benefit assessment as well as patient preferences. This decision should be made after consultation with an experienced medical team.

ECG — electrocardiogram; CEM — Center for Medical Education; CRP — C-reactive protein; CRT — cardiac resynchronization therapy; ICD — implantable cardioverter-defibrillator; LVEF — left ventricular ejection fraction; NYHA — New York Heart Association

OpenAI updates, it becomes paramount that we adopt a collective approach in instituting safety measures for our patients and medical education [12, 23]. These would encompass protecting susceptible groups from inherent biases and assessing potential detrimental implications or risks associated with implementing recommendations dispensed by LLMs, such as ChatGPT. This is especially critical for high-stakes decisions [12], as well as for formulating queries that may prove challenging to train due to the ambiguity of training data available on the Internet, which reflects the diversity in research data and global practice trends. We hold great enthusiasm for the potential contribution of ChatGPT in medical education, yet we exercise prudence in contemplating potential clinical applications of this burgeoning technology [22].

Conversely, ChatGPT does not possess the complex reasoning capabilities of a human, and its achievement in passing medical exams predominantly underscores the fact that present medical examinations largely rely on the rote learning of mechanistic models of health and diseases. However, this stands in stark contrast to the practice of medicine, which is fundamentally based on human interactions. For these reasons, AI will never supplant the role of nurses, doctors, and other frontline healthcare professionals. Although AI and LLMs are certain to revolutionize every aspect of our work, ranging from research and writing to graphic design and medical diagnosis, their current success in repeatedly passing standardized tests is a critique of our educational methods and how we train our doctors, lawyers, and students in general [21].

Conclusions

The profound consequences of ChatGPT, along with other LLMs, can be described as a paradigm shift in academia and healthcare practice. Discussing its potential benefits, future perspectives, and limitations appears timely and relevant.

Conflict of interest: None declared.

References

1. Adamopoulou E, Moussiades L. An overview of chatbot technology. In IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer 2020: 373–383.

2. Ritson M. Is ChatGPT the next significant threat to Google's dominance in the AI market? Marketing Weekly. https://www.marketingweek.com/ritson-chatgpt-google-ai (2022, December 9).

3. Brent AA. "Why ChatGPT is such a big deal for education". C2C Digital Magazine. 2023; 1(18): 14.

4. AI Will Transform Teaching and Learning. Let's Get it Right. https://hai.stanford.edu/news/ai-will-transform-teaching-and-learning-lets-get-it-right (Accessed June 2023).

5. Epstein RH, Dexter F. Variability in Large Language Models' Responses to Medical Licensing and Certification Examinations. Comment on "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: e48305, doi: 10.2196/48305, indexed in Pubmed: 37440293.

6. McGee R. Is chat GPT biased against conservatives? An empirical study. SSRN Electron J. 2023, doi: 10.2139/ssrn.4359405.

7. Kelly SM. ChatGPT passes exams from law and business schools. https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/ (2023, January 26).

8. Choi J, Hickman K, Monahan A, et al. ChatGPT Goes to Law School. SSRN Electron J. 2023, doi: 10.2139/ssrn.4335905.

9. Carbone C. How ChatGPT could make it easy to cheat on written tests and homework: 'You can NO LONGER give take-home exams or homework'. https://www.dailymail.co.uk/sciencetech/article-11513127/ChatGPT-OpenAI-cheat-testshomework.htm (2022, December 7).

10. Terwiesch C. Would Chat GPT3 Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course. https://mackinstitute.wharton.upenn.edu/2023/would-chat-gpt3-get-a-wharton-mba-newwhite-paper-by-christian-terwiesch (2023, January 17).

11. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023; 2(2): e0000198, doi: 10.1371/journal.pdig.0000198, indexed in Pubmed: 36812645.

12. 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; 3(4): 100324, doi: 10.1016/j.xops.2023.100324, indexed in Pubmed: 37334036.

13. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology. 2023; 307(5): e230582, doi: 10.1148/radiol.230582, indexed in Pubmed: 37191485.

14. Humar P, Asaad M, Bengur FB, et al. ChatGPT is equivalent to first year plastic surgery residents: evaluation of ChatGPT on the plastic surgery in-service exam. Aesthet Surg J. 2023 [Epub ahead of print], doi: 10.1093/asj/sjad130, indexed in Pubmed: 37140001.

15. Kumar AHS. Analysis of ChatGPT tool to assess the potential of its utility for academic writing in biomedical domain. Biol Eng Med Sci Rep. 2023; 9(1): 24–30, doi: 10.5530/bems.9.1.5.

16. Benoit J. ChatGPT for clinical vignette generation, revision, and evaluation. medRxiv. 2023, doi: 10.1101/2023.02.04.23285478.

17. https://isap.sejm.gov.pl/isap.nsf/download.xsp/WDU1997-0280152/O/D19970152.pdf (Accessed June 2023).

18. https://cem.edu.pl/pytcem/wyswietl_pytania_pes. (Accessed June 2023).

19. Murphy JFA. Assessment in medical education. Ir Med J. 2007; 100(2): 356, indexed in Pubmed: 17432806.

20. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ. 2023 [Epub ahead of print], doi: 10.1002/ase.2270, indexed in Pubmed: 36916887.

21. Tsang R. Practical applications of ChatGPT in undergraduate medical education. J Med Educ Curric Dev. 2023; 10: 23821205231178449, doi: 10.1177/23821205231178449, indexed in Pubmed: 37255525.

22. Mbakwe AB, Lourentzou I, Celi LA, et al. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health. 2023; 2(2): e0000205, doi: 10.1371/journal.pdig.0000205, indexed in Pubmed: 36812618.

23. Nath S, Marie A, Ellershaw S, et al. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. Br J Ophthalmol. 2022; 106(7): 889–892, doi: 10.1136/bjophthalmol-2022-321141, indexed in Pubmed: 35523534.