ChatGPT's Potential in Navigating the Complexity of the Polish Anaesthesiology Specialist Examination

Original Article

Michał Bielówka[1], Jakub Kufel[1], Marcin Rojek[1], Adam Mitręga[1], Dominika Kaczyńska[1], Łukasz Czogalika[1], Dominika Kondoł[2], Kacper Palkij[2], Wiktoria Bartnikowska[1]

[1]Department of Radiology and Nuclear Medicine, Faculty of Medical Sciences in Katowice, Medical University of Silesia;[2]Department of Radiology and Nuclear Medicine of the Medical University of Silesia in Katowice, Poland.

ABSTRACT

Purpose: This study aims to assess the capability of an artificial intelligence (AI) model, specifically ChatGPT-3.5, in answering questions from the test section of the Polish National Specialist Examination (PES) in anaesthesiology and intensive care.

Materials and Methods: A pool of 118 questions from the spring 2023 PES exam was utilized. Bloom's classification was employed to categorize questions based on comprehension, critical thinking, and memory. The questions were then presented to ChatGPT-3.5 in five independent sessions to evaluate its performance. Statistical analyses were conducted to assess correlations between the model's confidence, question difficulty, and correctness of answers.

Results: ChatGPT-3.5 achieved an overall accuracy of 47.5%, with variations observed across different question types and subtypes. Significant correlations were found between the model's confidence and answer correctness. However, no correlation was observed between the certainty index and question difficulty or answer correctness based on category or subcategory.

Conclusions: While ChatGPT-3.5 exhibited moderate performance, it fell short of the 60% threshold required to pass the PES exam. Comparison with similar AI studies in Japan suggests superior performance by the Polish AI model, albeit with limitations in expertise level. Human candidates consistently outperformed the AI model, indicating the current superiority of human expertise in this domain. Despite current limitations, continued research and collaboration offer promising prospects for AI integration in medical practice, supporting diagnostics, therapeutics, and patient care.

Key Words: Anaesthesiology, artificial intelligence, ChatGPT, intensive care, medical education, specialty examinations.

Received: 18 May 2024**, Accepted:** 23 August 2024

Corresponding Author: Michał Bielówka, MSc, Student Scientific Association of Computer Analysis and Artificial Intelligence at the Department of Radiology and Nuclear Medicine of the Medical University of Silesia in Katowice, Poland,**Tel.:** 501439954, **E-mail:** michalbielowka01@gmail.com

ISSN: 2090-925X, Vol.17, No.1, 2025

INTRODUCTION

Artificial intelligence (AI), is a branch of science that deals with solving logical problems using machines that mimic the work of the human brain. AI is tasked with interpreting given commands, formulating hypotheses, making logical analyses, interpreting images or planning[1].

The first definition of AI dates back to 1955 and was proposed by John McCarthy[1]. Since then, scientists have increasingly turned to this tool to automate specific areas of science and life. In medicine, too, new ideas are emerging for applying AI to diagnostic and therapeutic processes. As a result, the number of scientific papers treating AI is even growing at a logarithmic rate - between 2014 and 2019, the number of publications on AI in health care increased dramatically[2]. The current popularity of AI technology is due to the company OpenAI, which in November 2022 released the ChatGPT application, trained to provide complex and detailed answers to questions posed to it[3,4].

The authors of this article decided to test the accuracy of ChatGPT in providing correct answers to questions from the test part of the National Specialist Examination (PES) in anaesthesiology and intensive care, since the authors are increasingly addressing the use of AI in this field of medicine in the literature[5-7].

Obtaining a passing score on the test included in the PES is one of the components necessary to become a specialist in anaesthesiology and intensive care. This exam consists of solving 120 single-choice tasks with a minimum of 60% correct answers. The questions are designed to test the

DOI: 10.21608/ASJA.2024.290716.1111

young student's theoretical preparation, logical thinking skills and ability to draw conclusions[8].

The following text is intended to present the weaknesses and strengths of AI during the PES examination process. The authors undertook an analysis of the results of the ChatGPT in relation to the achievements that young doctors have. The task of this process is to outline the potential advantages of using AI technology in medicine, to point out the disadvantages and problems (often of an ethical nature), and to identify the direction of change that is taking place in modern medicine.

METHODOLOGY

Examination and questions

A pool of 120 questions from the spring 2023 Polish National Specialist Examination (PES) was used to conduct a study to assess the ability of an artificial intelligence model to provide correct answers in the anaesthesiology and intensive care specialty exam. The latest publicly available set was selected. Two questions were excluded, one as incompatible with modern medical knowledge and the other due to graphic content, leaving a pool of 118 questions[5]. The qualified questions were subjected to Bloom's classification and two parallel author's divisions: the first one divided the set of all questions into comprehension and critical thinking questions or memory questions. The second one, however, involved classifying the range of information to which the questions referred. Thus, subcategories such as: "anatomy and physiology", "anaesthesia", "medical guidelines", "medical procedures", "medication", "related to diseases" and "treatment".

Data collection and analysis

An analysis was conducted using the GTP-3.5 language model with an update date of June 1, 2023. In order to determine the parameter, the certainty factor, each question was asked five times in independent sessions to exclude the evaluation of previous answers by ChatGPT-3.5 and to examine the probabilistic nature of this language model. The study consisted of five sessions in which a total of 118 different questions were asked, each preceded by a prompt, which was a facilitator for collecting answers, limiting them to one letter and presenting the general concept of a single-letter test.

Static analysis

A series of statistical analyses were carried out on the set of responses obtained, using Statistica software (Statistica 13.1- StatSoft Poland) and the mathplotlib, scipy and plotly libraries of the Python language (Operated on data in the Jupyter Notebook environment). Based on the answers to the qualified questions, the author's parameter - the certainty index (Equation 1) was calculated. It determines the ratio of the most frequent answer in consecutive sessions to the number of independent sessions (n=5). The index determined in this way provides information about the "internal belief" of the model about the correctness of any of the answers.

$$P_{GPT} = \frac{{}_{j}\max \sum_{j=1}^{n} \delta(xj - xj)}{n=5}$$

Equation 1. Author's formula describing the certainty index.

RESULTS

The language model studied achieved a score of 47.5% correct answers (Table 1).

Performance in each question type and subtype was counted.

For the purpose of statistical analysis, the results in each type and subtype were compared (Tables 2, 3)

Table1: Number of correct and incorrect answers:

Correct answer	Number of questions	%
No	62	50,42%
Yes	56	49,58%

Table 2: Comparison of correct and incorrect answers by type.

Type	Comprehension and critical thinking questions	Memory questions
No	23	39
Yes	17	39
% of correct answers	42,5%	50%

Table 3: Comparison of correct and incorrect answers by subtype

Subtype	Number of incorrect answers	Number of correct answers	% of correct answers
anatomy and physiology	13	11	45,8%
anaesthesia	18	9	33,3%
medical guidelines	7	9	56,25%
medical procedures	4	4	50%
medication	8	15	65,2%
related to diseases	9	3	25%
treatment	3	5	62,5%

Among the tests conducted, it was noted that one relationship fulfilled the recognized threshold of significance. There was a statistically significant correlation between the model confidence index and the correctness of the question answer (p<0.0001). No correlation was shown between the certainty index and difficulty index (r=0.08). No significant correlation was shown between the correctness of an answer and its belonging to a category (p=0.56) or subcategory (p=0.25). There was also no significant relationship between the difficulty factor and the correctness of the answer (p=0.017) (Figure 1)

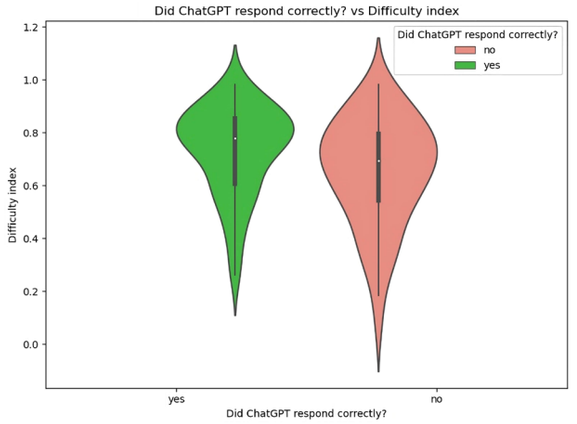

Fig. 1 Comparing the correctness of the answers of the tested language model with the difficulty index of questions.

DISCUSSION

Specialization in anaesthesiology and intensive care in Poland lasts six years. Over the past few years, the number of residency spots for this specialty has increased significantly. For comparison, in 2020, 279 residencies were granted in the fall and spring recruitment combined, while in 2023 as many as 525, which is related to the inclusion of this specialty in the list of priority medical fields. This means that the number of specialists in the country is insufficient in relation to patient needs. The Specialty Examination in Anaesthesiology and Intensive Care is a single-choice exam, and passing it is equivalent to receiving the title of specialist in this field upon completion of training. Like every specialty exam in Poland, it consists of a theoretical and practical part. In order to pass the theoretical exam, it is necessary to correctly answer at least 60% of the questions, except that a score above 75% exempts you from the practical exam. In 2018, 126 people out of 141 who took the practical exam received the diploma of specialist in this field of medicine[9]. Taking advantage of the opportunities offered by the development of artificial intelligence has been one of the more rapidly growing fields of research in medicine for years. Scientists around the world are asking themselves whether the level of AI expertise will be able to match that of a doctor. Specialty exams are some of the most difficult tests in a doctor's career. To our knowledge, our study is the only one to examine how ChatGPT fared in passing such an exam in anaesthesiology and intensive care.

In a study conducted by J. Kufel *et al*., the identical language model was examined and achieved an overall score of 56% when answering PES questions related to nuclear medicine. Interestingly, this score is slightly higher than the one observed in our study, implying a potential proficiency in nuclear medicine over pathology. This discrepancy could be attributed to the abundance of online resources accessible to the ChatGPT[10].

The subsequent study conducted by Kinoshita M. *et al*., examined how Chat GPT-3.5 and ChatGPT-4 would fare in answering questions on the written portion of the 2021 and 2022 JSA-Certified Anaesthesiologist examinations. As in our study, the criteria for excluding questions were those containing diagrams and figures, and those removed by the Japanese Society of Anaesthesiologists (JSA) due to errors in question content. The study used 163 questions from 2021. (132 general, 31 clinical) and 93 questions from 2022 (71 general, 27 clinical). Questions were originally asked in Japanese. ChatGPT-3.5 scored 23.3% on the 2021 exam and 21.4% on the 2022 exam. In comparison, the GPT-4 in the Konoshita M. *et al*., study scored 51.5% on the 2021 test and 49.0% on the 2022 test. Neither GPT-3.5 nor GPT-4 showed significant differences in accuracy between the general and clinical questions. 1.0% (GPT-3.5) and 2.3% (GPT-4) of responses were categorized as "beyond my knowledge." The official results of the JSA exam are not known, only data is given that the average scores for the newly created questions answered by the examinees ranked between 45-67%. The study's authors note that in the study conducted by Tanaka *et al*., after translating the Japanese Medical Licensing Exam (JMLE) questions into English, ChatGPT-4 did better in answering them. This may have been due to the language barrier and the more specialized knowledge required to answer correctly on the JSA exam[11].

The results achieved by ChatGPT 3.5 in the Kinoshito M. *et al*., study are significantly worse than those achieved by ChatGPT 3.5 in our study. The artificial intelligence model answered 118 questions on the Polish specialty exam, of which it did worst on the subcategories "anaesthesia" and "related to diseases," while it did best on the subcategories "medication" and "treatment." This may reflect the greater availability of materials applicable to the online version of the exam in Polish than in Japanese, the differences in the level of difficulty of the exam in the two countries, or the language barrier, which the aforementioned authors also point out. They also emphasize that the ChatGPT needs to be improved in specialized areas of medicine and should be used by those with medical knowledge. The authors of this study also believe that when using ChatGPT in certain medical fields, which include anaesthesiology and intensive care, special care should be taken. The results of the study indicate that, for the time being, the ChatGPT is not a tool whose knowledge is comparable to that of specialists who pass final exams. This means that the knowledge gained during training is essential to ensure maximum safety for patients during their hospital stay, and precludes the possibility of replacing anaesthesiologists in patient care. Perhaps in the future AI will become a support for intensive care physicians in patient care, which would relieve their workload. Due to the small number of studies on the use of ChatGPT in the field of anaesthesiology and intensive care, the possibility of comparing our results with other researchers is limited. We agree that more research is needed to see how ChatCPT performs in this medical field, and how medics could use it in their clinical work in the future.

CONCLUSIONS

A study was conducted using the GPT-3.5 language model, which answered questions from the test portion of the PES exam in anaesthesiology and intensive care. The results show that the model achieved 47.5% correct answers, with different results depending on the type and subtype of questions. The score achieved by ChatGPT-3.5 is not enough to reach the 60% threshold required by the PES. Statistical analysis showed significant correlations between model confidence and correctness of answers, in addition, it can be noted that the model performed better on 'memory questions' (50%) than 'comprehension and critical thinking questions' (42.5%). A comparison with the results of another AI study in Japan suggests that the Polish AI model performed better, although there are still limitations related to the level of expertise.

Between 2009 and 2018, 1,680 people took the specialist exam, while 1,610 doctors achieved a positive result, giving us a pass rate of 95.8%. This shows the significant advantage of humans over the artificial intelligence model tested in this study.

Despite current limitations and challenges, the future prospects for the use of AI in medicine are promising. Through further research, collaboration between scientists, doctors and technologists, and the development of appropriate regulations and standards, AI can become an even more effective tool to support the diagnostic, therapeutic and patient care process.

CONFLICT OF INTERESTS

There are no conflicts of interest

REFERENCES

1. Różanowski K. Sztuczna inteligencja rozwój, szanse i zagrożenia. Zeszyty Naukowe Warszawskiej Wyższej Szkoły Informatyki. (2007). nr 2. Accessed April 17, 2024. http://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-1cd1832b-24aa-4187-9c90-4bed4cd6eb97.
2. Guo Y, Hao Z, Zhao S, Gong J, Yang F. (2020). Artificial Intelligence in Health Care: Bibliometric Analysis. J Med Internet Res. 22(7):e18228. doi:10.2196/18228.
3. Introducing GPTs. Accessed December 21, (2023). https://openai.com/blog/introducing-gpts.
4. Introducing ChatGPT. Accessed April 17, (2024). https://openai.com/blog/chatgpt.
5. Paiste HJ, Godwin RC, Smith AD, Berkowitz DE, Melvin RL. (2024). Strengths-weaknesses-opportunities-threats analysis of artificial intelligence in anesthesiology and perioperative medicine. Front Digit Health. 6:1316931. doi:10.3389/fdgth.2024.1316931.
6. Bellini V, Valente M, Gaddi AV, Pelosi P, Bignami E. (2022). Artificial intelligence and telemedicine in anesthesia: potential and problems. Minerva Anestesiol. 88(9):729-734 doi:10.23736/S0375-9393.21.16241-8.
7. Langeron O, Castoldi N, Rognon N, Baillard C, Samama CM. (2024). How anesthesiology can deal with innovation and new technologies? Minerva Anestesiol. 90(1-2):68-76. doi:10.23736/S0375-9393.23.17464-5.
8. Centrum Egzaminów Medycznych. (2023). Accessed December 21 . https://www.cem.edu.pl/spec.php.
9. Centrum Egzaminów Medycznych. Accessed April 17, (2024). https://www.cem.edu.pl/aktualnosci/spece/spece_stat2.php?nazwa=Anestezjologia%20i%20intensywna%20terapia.
10. Kufel, J., Bielówka, M., Rojek, M., Mitręga, A., Czogalik, L., Kaczyńska, D., Kondoł, D., Palkij, K., & Mielcarska, S. (2024). Assessing ChatGPT's performance in national nuclear medicine specialty examination: An evaluative analysis. Iranian Journal of Nuclear Medicine, 32(1), 60-65. doi: 10.22034/irjnm.2023.129434.1580.
11. Kinoshita M, Komasaka M, Tanaka K. (2024). ChatGPT's performance on JSA-certified anesthesiologist exam. J Anesth. 38(2):282-283. doi:10.1007/s00540-023-03275-4.