© Pol J Radiol 2023; 88: e430-e434
DOI: https://doi.org/10.5114/pjr.2023.131215

Received: 06.06.2023
Accepted: 07.06.2023
Published: 18.09.2023

Polish Journal of Radiology

http://www.polradiol.com

Original paper

Will ChatGPT pass the Polish specialty exam in radiology and diagnostic imaging? Insights into strengths and limitations

Jakub Kufel[1,A,B,D,E,F,G], Iga Paszkiewicz[2,D,E,F], Michał Bielówka[3,A,B,D,F], Wiktoria Bartnikowska[4,D,E,F], Michał Janik[3,B,D,F], Magdalena Stencel[3,B,D,F], Łukasz Czogalik[3,B,D,F], Katarzyna Gruszczyńska[5,E,F], Sylwia Mielcarska[6,C,D,E]

[1]Department of Biophysics, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Zabrze, Poland
[2]Tytus Chałubiński Hospital, Zakopane, Poland
[3]Professor Zbigniew Religa Student Scientific Association at the Department of Biophysics, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Poland
[4]Faculty of Medical Sciences in Katowice, Medical University of Silesia, Katowice, Poland
[5]Department of Radiology and Nuclear Medicine, Faculty of Medical Sciences in Katowice, Medical University of Silesia, Katowice, Poland
[6]Department of Medical and Molecular Biology, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Zabrze, Poland

Abstract

Purpose: Rapid development of artificial intelligence has aroused curiosity regarding its potential applications in medical field. The purpose of this article was to present the performance of ChatGPT, a state-of-the-art language model in relation to pass rate of national specialty examination (PES) in radiology and imaging diagnostics within Polish education system. Additionally, the study aimed to identify the strengths and limitations of the model through a detailed analysis of issues raised by exam questions.

Material and methods: The present study utilized a PES exam consisting of 120 questions, provided by Medical Examinations Center in Lodz. Questions were administered using openai.com platform that grants free access to GPT-3.5 model. All questions were categorized according to Bloom's taxonomy to assess their complexity and difficulty. Following the answer to each exam question, ChatGPT was asked to rate its confidence on a scale of 1 to 5 to evaluate the accuracy of its response.

Results: ChatGPT did not reach the pass rate threshold of PES exam (52%); however, it was close in certain question categories. No significant differences were observed in the percentage of correct answers across question types and sub-types.

Conclusions: The performance of the ChatGPT model in the pass rate of PES exam in radiology and imaging diagnostics in Poland is yet to be determined, which requires further research on improved versions of ChatGPT.

Key words: ChatGPT, deep learning, large language model, artificial intelligence.

Introduction

ChatGPT is a massive artificial intelligence (AI)-based model that has taken the world by storm [1]. In just over 5 days, ChatGPT recorded over 1 million users, achieving a milestone that Facebook took 10 months to reach, Instagram accomplished in 2.5 months, and Netflix achieved in 41 months. This model was built upon the foundation of a large language model (LLM), and was trained on a huge amount of text data, exceeding 45 terabytes [2]. LLM utilize deep neural networks to analyze and generate text based on data inputted into the model. While ChatGPT

Correspondence address:
Jakub Kufel, Department of Biophysics, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Zabrze, Poland, e-mail: jakubkufel92@gmail.com

Authors' contribution:
A Study design · **B** Data collection · **C** Statistical analysis · **D** Data interpretation · **E** Manuscript preparation · **F** Literature search · **G** Funds collection

has yet to be trained specifically for medical use in Poland, efforts have already commenced in the United States to use this tool in more efficient patients' descriptions, medical education, and retrieval of medical information [3].

In the field of radiology, AI has primarily been employed for automating the generation of imaging study descriptions and image analysis. However, these applications, though more advanced in Western countries, are still in testing phase in Poland. Nevertheless, ChatGPT has gained interest in its potential applications in education, differential diagnosis, and disease classification, utilizing LLM capabilities [4].

ChatGPT uses a deep learning model to recognize patterns and relationships between words. Its functionality relies on extensive training databases to generate human-like responses. However, it is important to note that the model's reactions can sometimes be inaccurate or incorrect. Despite this, an intriguing study evaluating ChatGPT performance on United States medical licensing exam (USMLE) yielded surprising results. The model achieved a score of 60%, equivalent to passing the exam [5].

Apart from the broad applicability of LLM and ChatGPT, on which it is based, its use in medicine and radiology has not yet been defined. National specialty examination (NSE) aims to comprehensively assess the knowledge, reasoning abilities, and decision-making skills of specialist trainees in specific clinical scenarios. Our study aimed to evaluate ChatGPT's effectiveness in answering radiological NSE questions, and analyze its strengths and weaknesses in comparison with human cognition.

Material and methods

Examination and questions

This prospective study was conducted from May 5th, 2023 to May 24th, 2023. The focus of the study was one specialty exam in radiology and diagnostic imaging (Spring, 2023), which was randomly selected from available exams in the question archive database of Medical Examinations Center in Lodz, Poland. The selected exam comprised of 120 single-choice questions, each having one correct answer and four distractors (wrong answers). One question was excluded by the Board of Examiners, as it was not in line with current knowledge. Therefore, a total of 119 questions were analyzed.

To ensure comprehensive analysis, all questions were classified according to Bloom's taxonomy [6,7]. The classification included memory questions, comprehension, and critical thinking questions as well as further sub-divisions, such as calculations and classifications, description of imaging results, disease-related questions, and clinical management questions. Additionally, each question was categorized as physical, clinical, or topography-related. Two independent researchers performed the classification, and any disagreements were resolved by a third independent researcher. Inter-observer agreement was evaluated using a statistical test, with Cohen's coefficient of 0.95 (agreement, 97.9%), indicating near-perfect agreement [8].

Data collection and analysis

Prior to presenting the questions, ChatGPT-3.5 was provided with exam rules, including number of questions, number of answer options, and number of correct answers. Furthermore, after each question, an additional query was posed to ChatGPT, asking, "On a scale of 1 to 5, how confident are you in this answer?" This was done to assess ChatGPT's level of confidence in its chosen response. The scale was defined as follows: 1 represented "definitely not sure", 2 "not very sure", 3 indicated "almost sure", 4 "very sure", and 5 meant "definitely sure". Each question was inputted into ChatGPT, and all chat interactions were documented (see Supplement 1). To maintain consistency with the content of exam questions, the chat dialogue was conducted in Polish.

Statical analysis

The results obtained from ChatGPT were compared with correct answers and statistics published by the Medical Examinations Center in Lodz. The evaluation focused on determining the percentage of correct answers provided by ChatGPT. The percentage of correct answers provided by ChatGPT for different question types and sub-types was also compared. The difficulty of questions that were answered correctly and incorrectly by ChatGPT was also analyzed.

To assess the significance between distributions of correct and incorrect answers, question type, and other qualitative variables, Pearson χ^2 test was applied. Shapiro-Wilk test was used to evaluate the distribution of quantitative variables, such as question difficulty and RBPI (relative Bloom's proficiency index). For comparing quantitative variables between groups, Mann-Whitney *U* test was employed. R Studio (Integrated Development for R. Studio, PBC, Boston, MA, USA) was used for all statistical analyses.

In addition, to compare the confidence level of responses between correct and incorrect answers, Mann-Whitney *U* test was applied. *P*-values of less than 0.05 were considered significant.

Results

In the radiology NSE, ChatGPT received an overall failing grade, scoring 52%. The passing threshold for the exam was set at 60% (Table 1). For statistical analysis, specific groups of questions, on which ChatGPT was assessed were selected. These groups included "critical thinking" and "knowledge" questions. The "critical thinking" group was further divided into sub-categories, such as "clinical management", "describing imaging studies", "calculating

Table 1. Percentage of correct and false answers submitted by ChatGPT in the whole test

Correct answer	Number	Percentage
Yes	63	0.52940
No	56	0.47060

Table 2. Distribution of correct/false answers and types of questions, χ^2 test ($p = 0.31$)

Question type	Correct answer	
	Yes, *n* (%)	No, *n* (%)
Comprehension and critical thinking	50 (55.56)	40 (44.44)
Memory	13 (44.83)	16 (55.17)

and classifying", and "disease-related" questions. The questions were also categorized into three types, such as "clinical", "physical", and "topography".

ChatGPT's performance in the exam was as follows: for "critical thinking" questions, it scored 55.56% (50 out of 90), and for "knowledge" questions, it scored 44.83% (13 out of 29) (Table 2). In the sub-categories, ChatGPT scored 75% (6 out of 8) for "clinical management" questions, 62.86% (22 out of 35) for "clinical examination" descriptions, 51.43% (18 out of 35) for "disease-related" questions, 33.33% (4 out of 12) for "calculations and classifications", and 33.33% (3 out of 9) for "topography-related" questions (Table 3). Regarding question types, ChatGPT correctly answered 54.55% (54 out of 99) of clinical questions and 54.55% (6 out of 11) of physical questions (Table 4). No significant differences were observed in the percentage of correct answers among questions types and sub-types (χ^2 test; Tables 2-4).

A significant difference in the difficulty index was observed between questions that ChatGPT answered correctly and those answered incorrectly. The difficulty index (Table 5, Figure 1) was significantly higher for questions that ChatGPT answered correctly. Additionally, the confidence index between correct and incorrect answers was analyzed (Table 4, Figure 2), revealing that questions answered correctly by ChatGPT had a higher confidence index. No correlation was observed between the question difficulty index and the certainty of answer on a five-point

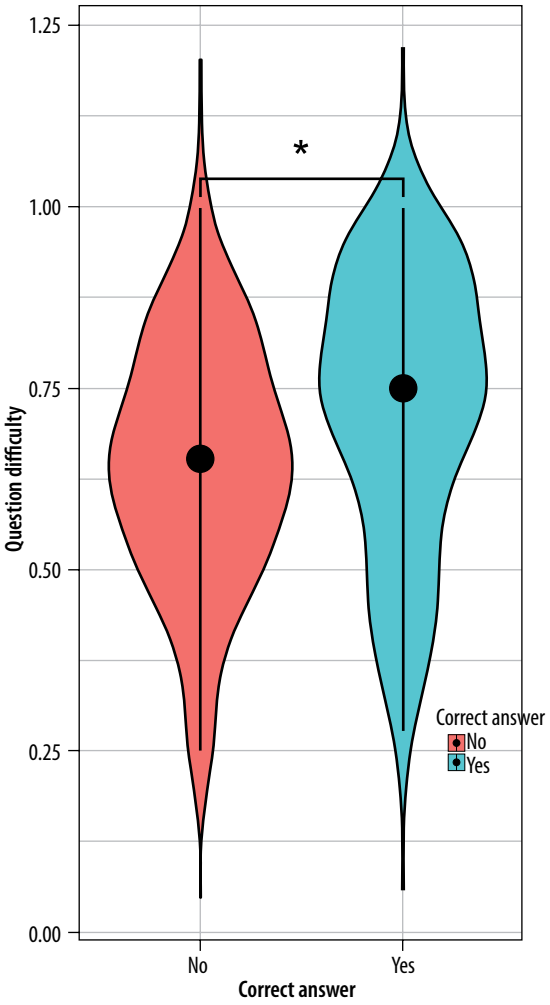

Figure 1. Comparison of questions difficulty between correct and false answers submitted by ChatGPT (Mann-Whitney *U* test, $p < 0.05$)

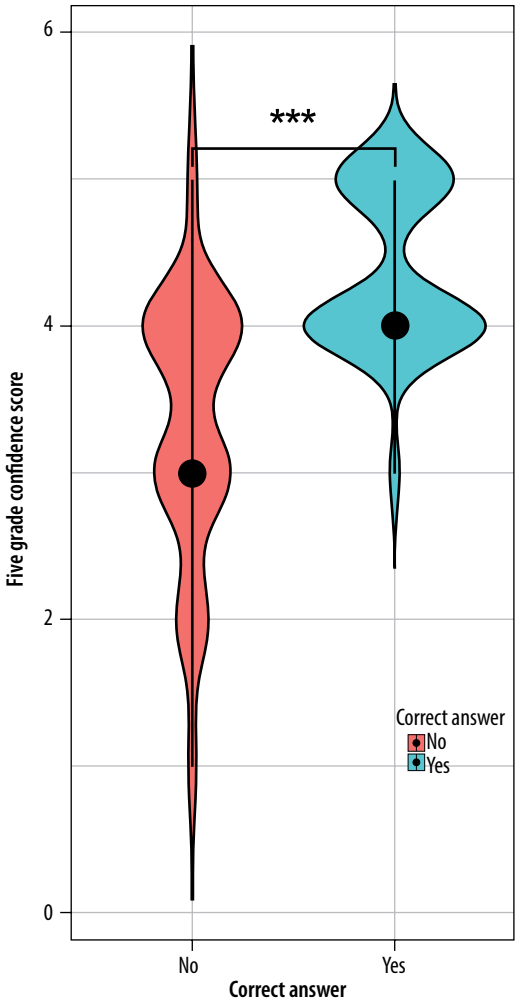

Figure 2. Comparison of five-grade confidence score between correct and false answers submitted by ChatGPT (Mann-Whitney *U* test, $p < 0.001$)

Table 3. Distribution of correct/false answers and types of questions, χ^2 test ($p = 0.20$)

Question type	Correct answer	
	Yes, *n* (%)	No, *n* (%)
Description of imaging results	22 (62.86)	13 (37.14)
Related to diseases	18 (51.43)	17 (48.57)
Calculation and classification	4 (33.33)	8 (66.67)
Clinical management	6 (75.00)	2 (25.00)

Table 4. Distribution of correct/false answers and types of questions, χ^2 test ($p = 0.47$)

Question type	Correct answer	
	Yes, *n* (%)	No, *n* (%)
Physical	6 (54.55)	5 (45.45)
Clinical	54 (54.55)	45 (45.45)
Topography	3 (33.33)	6 (66.67)

Table 5. Comparison of questions difficulty, and five-grade confidence score between correctly and incorrectly answered questions by ChatGPT (Mann-Whitney *U* test)

	Correct answer			False answer			*p*-value
	Median	Q1	Q3	Median	Q1	Q3	
Questions difficulty	0.75	0.61	0.89	0.65	0.54	0.78	< 0.05
Five-grade confidence score	4.00	4.00	5.00	3.00	3.00	4.00	< 0.001

scale. There was no association between the type of question and the frequency of correct answers submitted by ChatGPT.

Discussion

The specialist examination in radiology and imaging diagnostics is a crucial assessment for individuals seeking to complete their specialist training and become specialists in this field of medicine. The examination comprises both practical and theoretical component. In Poland, a score of 60% or higher is considered a passing grade for the specialist examination in radiology and imaging diagnostics. However, if a candidate achieves a score of 75% or higher in the theoretical part, oral examination is waived. Similar qualification examinations are used in many countries worldwide. For instance, experiments similar to ours have been conducted in Canada and the United States.

In our study, ChatGPT performed significantly worse (52%) than in a study by Bhayana *et al.,* who investigated the pass rate of the Canadian Royal College examination in diagnostic radiology [9]. This means that ChatGPT did not attain a score high enough to pass the exam. In our study, ChatGPT achieved a higher score on questions requiring critical thinking (55%) compared with questions needing knowledge (44%). However, it performed worse than in Bhayan et al. study, where it scored 84% on lower-order thinking questions and 60% on higher-order thinking questions ($p = 0.002$). It is quite remarkable that ChatGPT achieved superior outcomes in questions, which required critical thinking as opposed to questions needed knowledge. One explanation might be that the PES is the most difficult exam a potential radiologist will take, which makes PES questions very complex and difficult. It is possible that such a scenario may arise where a question requiring knowledge is structured in a highly intricate manner, resulting in inadequate responses. ChatGPT performed best on clinical questions (75%), but its performance was better in Bhayana *et al.* study (89%). Similarly, the AI model performed poorly on calculation and classification questions in both studies, achieving a score of 33% in our study and 25% in Bhayana *et al.* research. Moreover, theoretically, ChatGPT as an AI tool should exhibit superior performance in questions regarding calculation and classification. However, this was not the case in this instance. Many questions in this category required ChatGPT to classify complex and difficult radiological findings into many different categories, which resulted in very poor results. It performed better on physical questions in Bhayana *et al.* study (55% compared with 30% in our study). Conversely, it performed worse on clinical questions in the present study (55% compared with 73% in Bhayana *et al.*). Additionally, we observed that ChatGPT consistently used certain language patterns to present its answers, which might have influenced user's perception of correctness, even when the answer was incorrect.

A study by Gilson *et al.* [10] demonstrated that the latest version of the ChatGPT model outperformed previous versions by 8.15% in answering questions from AMBOSS (a publicly available database of questions for medical students) and National Board of Medical Examiners (NBME) database. The study also found that logical reasoning for ChatGPT's answer choices was present in 100% of the results from the NBME datasets, even when they did not align with the answer key.

In our study, we identified that questions, for which ChatGPT provided correct answer, had a significantly higher confidence index. Therefore, the confidence index can be considered a parameter indicating a higher likelihood of ChatGPT providing a correct answer (Figure 2). Discrepancies between the obtained results are

likely attributed to differences in the quality, complexity, and specificity of the presented test questions as well as a potential language barrier since Polish is not the primary language for the ChatGPT model. It can be concluded that, given equal access to training resources and training time, the Polish specialist exam is significantly more challenging than the Canadian Royal College exam. Furthermore, it is reasonable to assume that if the Polish radiology textbooks recommended for studying for the specialty exam were more readily available in an online format, ChatGPT would be more likely to utilize the data from these textbooks in its training, therefore improving its ability to handle questions from the NSE.

Conclusions

Based on the presented results, it can be concluded that the performance of the ChatGPT model in passing the specialist in radiology and imaging diagnostics examination in Poland remains uncertain. In our experiment, the model did not achieve the minimum score required for a passing grade, although it came close in certain categories. To truly assess the effectiveness of ChatGPT in successfully passing the specialty exam, further testing using official questions provided by the Centre for Medical Examinations is necessary.

When evaluating the usefulness of ChatGPT, factors such as, the level of difficulty of the exam, the specific types of questions that posed the greatest challenges, and the availability of recommended scientific sources for studying for the NSE, should also be taken into consideration. It is possible that future versions of the ChatGPT model may perform better in addressing requirements of the NSE, but currently, there is no evidence to support this claim. Further research and testing of ChatGPT on the passing rates of state examinations in radiology are necessary to gain a more comprehensive understanding of its capabilities.

Conflict of interest

The authors report no conflict of interest.

References

1. Duarte F. Number of ChatGPT Users (2023). Available at: https://explodingtopics.com/blog/chatgpt-users (Accessed: 04.06.2023).
2. Sallam M. The utility of ChatGPT as an example of large language models in healthcare education, research and practice: systematic review on the future perspectives and potential limitations. Healthcare 2023; 11: 887. doi: 10.3390/healthcare11060887.
3. The Doctor's Digital Path to Treatment. Available at: https://www.thinkwithgoogle.com/marketing-strategies/search/the-doctors-digital-path-to-treatment/ (Accessed: 04.06.2023).
4. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023; 307: e230163. doi: 10.1148/radiol.230163.
5. 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: e0000198. doi: 10.1371/journal.pdig.0000198.
6. Anderson LW, Krathwohl DR. A Taxonomy for Learning, Teaching, and Assessing: a Revision of Bloom's Taxonomy of Educational Objectives. Pearson; 2001.
7. Sawin EI. Taxonomy of educational objectives: the classification of educational goals. Handbook 1. Committee of College and University Examiners, Benjamin S. Bloom. Elem Sch J 1957; 57: 343-344.
8. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012; 22: 276-282.
9. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023; 307: e230582. doi: 10.1148/radiol.230582.
10. 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. doi: 10.2196/45312.