Original paper

Artificial intelligence in cardiology: evaluating the accuracy of ChatGPT-o1-preview in a medical specialisation exam

Michał Bielówka[1,2], Adam Mitręga[2], Dominika Kaczyńska[2], Natalia Denisiewicz[2], Marcin Rojek[2], Mikołaj Magiera[2], Adam Mastalerz[3], Maja Dreger[2], Jakub Kufel[4]

[1]Department of Biophysics, Faculty of Medical Sciences in Zabrze, Medical University of Silesia in Katowice, Poland
[2]Students' Scientific Association of Computer Analysis and Artificial Intelligence at the Department of Radiology and Nuclear Medicine of the Medical University of Silesia, Katowice, Poland
[3]Faculty of Automatic Control, Electronics, and Computer Science, Department of Distributed Systems and Informatic Devices, Silesian University of Technology, Gliwice, Poland
[4]Department of Radiodiagnostics, Interventional Radiology and Nuclear Medicine, Medical University of Silesia, Katowice, Poland

Medical Studies
DOI: https://doi.org/10.5114/ms/209889

Key words: ChatGPT, artificial intelligence, cardiology, natural language processing.

Abstract

Introduction: Recent advancements in artificial intelligence (AI) language models, such as ChatGPT, have sparked interest in their application across various domains, including medicine.
Aim of the research: This study evaluates the performance of the ChatGPT-o1-preview model on the Polish Specialist Examination (PES) in cardiology.
Material and methods: A total of 120 single-choice questions from the Spring 2023 PES cardiology exam were analysed. Each question was submitted to the model five times in separate sessions using a standardised prompt. We assessed the accuracy of the responses and the model's consistency by introducing a "Confidence Index of Language Model" (defined as the ratio of the frequency of the most commonly selected answer to the total number of sessions). Questions were classified according to Bloom's Taxonomy, and statistical relationships between model performance and question characteristics were analysed.
Results: ChatGPT-o1-preview provided 97 correct answers (80.83%), well above the passing threshold of 60%. Statistically significant correlations were observed between answer accuracy and difficulty indices (both the Human Difficulty Index and the Confidence Index of Language Model). Questions that were more difficult for human examinees also proved more challenging for the model. No significant differences in model performance were found based on question type (clinical versus other) or content (comprehension and critical thinking versus memory).
Conclusions: ChatGPT-o1-preview successfully completed a specialised cardiology exam. It is essential to conduct further, more extensive research based on a significantly larger and more diverse set of questions. Only then will it be possible to reliably determine the true potential and limitations of AI in medical applications.

Introduction

Polish Specialist Examination (PES) in Cardiology is a standardised test administered twice a year for resident physicians. It represents the final stage of specialist training in cardiology. To pass the examination, candidates must correctly answer at least 60% of the questions (i.e. a minimum of 72 out of 120). Successful completion of the PES confers the title of cardiology specialist. ChatGPT (Chat Generative Pre-trained Transformer) is a tool based on artificial intelligence (AI) and natural language processing, made publicly available in November 2022 by the company OpenAI [1]. Currently, ChatGPT is being utilised across various fields of knowledge, including customer service, content creation, and computer programming. Its undeniable popularity may stem from its ability to deliver expected outcomes in significantly less time than manual efforts would require. In medicine, ChatGPT has drawn considerable attention from the medical community due to its potential to advance the field [2].

In clinical practice, AI assistance may prove beneficial across multiple domains. It can aid in assessing the clinical status of patients and support decision-making regarding diagnosis and treatment, potentially minimising cognitive bias. For both physicians and patients, ChatGPT may serve as a helpful tool for acquiring medical knowledge because it is capable of answering a wide range of medical questions [3].

Researchers worldwide have also attempted to harness AI in cardiology. ChatGPT was evaluated on cardiology-related questions from the United States Medical Licensing Examination (USMLE), achieving a 60% accuracy rate. Artificial intelligence has also been studied for its usefulness in radiology, particu-

larly in echocardiography and computed tomography (CT), for measuring the dimensions of various anatomical structures [4, 5].

However, due to concerns about bias and inaccuracies – such as the generation of incorrect or fabricated information (e.g. false citations) – the clinical use of AI remains controversial [6]. Currently, the utility of ChatGPT is limited by legal concerns [7].

Inspired by the results of our previous efforts to assess the performance of ChatGPT, we conducted an evaluation comparing responses generated by ChatGPT-o1-preview with the correct answers from PES in cardiology. The PES is a single-choice examination, each with four distractors and one correct option. This analysis aims to draw conclusions about the accuracy of AI-generated responses and to compare the capabilities of ChatGPT with human cognitive performance.

Materiał and methods

Data collection and analysis

The questions were obtained, with permission, from the Examination Centre in Lodz using a custom Python-based web scraper. The extracted data included the question content, its status (valid or excluded due to factual or structural inaccuracy), and session-specific statistics such as the difficulty index, calculated based on the number of correct responses provided by examinees for each question.

Three custom classification schemes inspired by Bloom's Taxonomy [8] were developed:

- Type: distinguishing between memory-based and comprehension/critical-thinking questions;
- Subtype: including categories such as clinical procedures, disease-related, diagnostic, and treatment questions;
- Clinical relevance: separating clinical questions from other questions.

Examination and questions

All 120 questions from the Spring 2023 session of the PES in cardiology were used in the analysis [9]. Each question – preceded by a standardised prompt – was submitted five times in independent sessions to a large language model. The prompt was designed to clearly define the rules and format of the examination, ensuring a reliable simulation of the single-best-answer condition. The Human Difficulty Index (HDI) values were obtained from the official archival database, where they had been calculated by the Polish National Examination Centre (Centrum Egzaminów Medycznych, CEM) according to their standard formula. The HDI is defined as $HDI = (Ns + Ni)/(2n)$, where n is the number of examinees in each of the extreme groups (the upper 27% and the lower 27% of total test scores), Ns is the number of correct responses in the high-scoring group, and Ni is the number of correct responses in the low-scoring group. The HDI ranges from 0 (extremely difficult items) to 1 (extremely easy items). This index differs from the raw percentage of correct responses because it does not take into account answers given by examinees with average test performance (CEM, 2023).

To assess model consistency, we used the Confidence Index for the Language Model (CI-LM), adapted from a previously established confidence measure in the literature. CI-LM is defined as the frequency of the most commonly selected answer across five sessions, divided by the total number of sessions. This metric reflects the model's certainty in its responses and aligns with methodologies reported in prior studies.

Statistical analysis

All statistical analyses were performed with a significance threshold of $p < 0.05$. The analyses were conducted using Python version 3.11.1 with the following libraries: NumPy, SciPy, Pandas, Matplotlib, and Plotly.

The following statistical methods were used:

- To examine correlations between continuous variables, such as the CI-LM and the Human Difficulty Index (HDI), Spearman's rank correlation was applied due to the non-normal and discrete nature of the variables.
- To assess relationships between continuous variables and binary categorical variables (e.g. Type, Clinical relevance, and Correctness of ChatGPT's response), the Mann-Whitney U test was used.
- For associations between categorical variables, chi-square (χ^2) tests were conducted.

Results

The number of correct answers provided by ChatGPT was 97 out of 120 points, corresponding to 80.83%, while 23 (19.17%) answers were incorrect.

The results of the quantitative analyses compared with nominal binary variables are presented in Table 1. For all comparisons, the Mann–Whitney U test was used.

Two statistically significant associations were identified ($p < 0.05$): A significant relationship was found between the Human Difficulty Index and ChatGPT's response accuracy (median difficulty index for correct responses: $\bar{x} = 0.742$; for incorrect responses: $\bar{x} = 0.468$), as illustrated in Figure 1.

A similar relationship was observed for the Confidence Index of Language Model (CI-LM) (median for correct responses: $\bar{x} = 1.0$; for incorrect responses: $\bar{x} = 0.8$), as shown in Figure 2.

Spearman's rank correlation was used to analyse the relationship between the CI-LM and the HDI; however, the correlation did not reach statistical significance ($p = 0.06$).

Table 1. Results of the analysis of quantitative variables with binary nominal variables

Metric	Category	*U* Statistic	*P*-value
CI-LM	Type	1576.5	0.6616
CI-LM	Answer correctness	654.5	0.0002
CI-LM	Division into clinical and other questions	930	0.3687
HDI	Type	1497	0.9334
HDI	Answer correctness	516.5	0.00006
HDI	Division into clinical and other questions	1185	0.316

CI-LM – Confidence Index of Language Model, HDI – Human Difficulty Index.

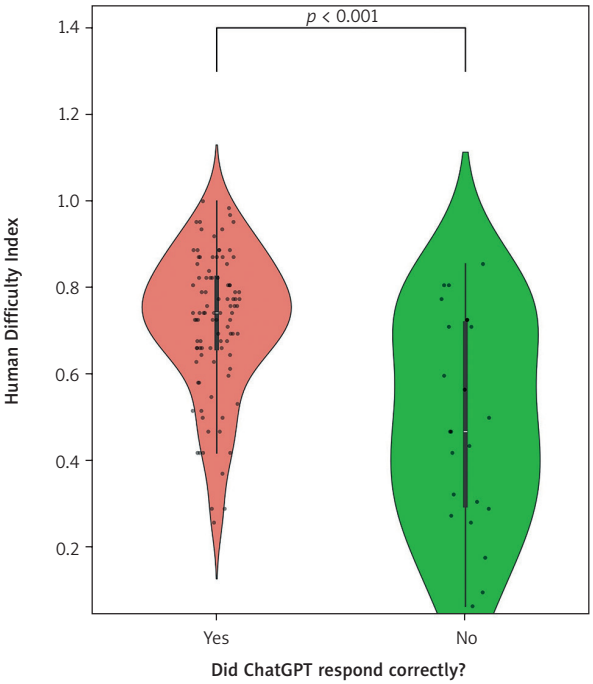

Figure 1. Association between the Human Difficulty Index and response accuracy of the language model

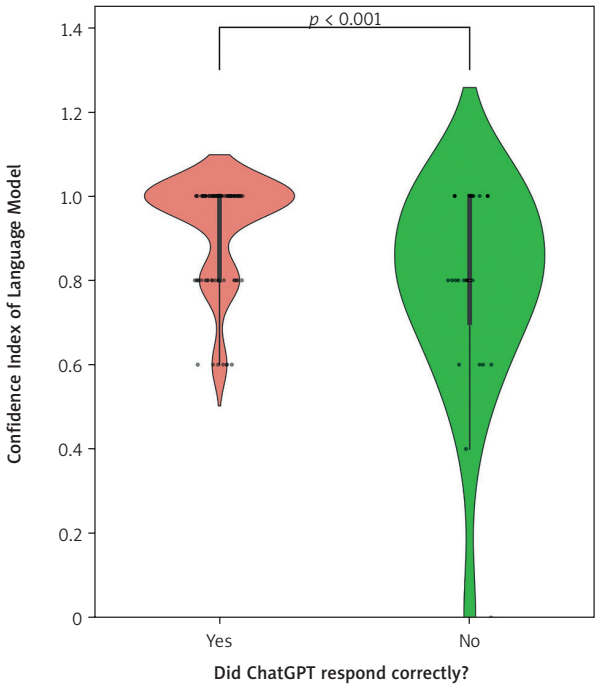

Figure 2. Association between the Confidence Index of Language Model and the model's response accuracy

Table 2. Results of the exam with the division into "memory questions", "comprehension and critical thinking questions", "clinical questions", and "other questions"

Category of questions	Correct answer			
	Yes	%	No	%
Comprehension and critical thinking	29	80.56	7	19.44
Memory	68	80.95	16	19.05
Clinical	79	79.8	20	20.2
Other	18	85.7	3	14.3

To evaluate whether the correctness of ChatGPT's responses was associated with question categories – specifically clinical vs. other, and comprehension/critical thinking vs. memory – two chi-square (χ^2) tests were conducted. No statistically significant associations were found (clinical vs. other: $p = 0.76$; memory vs. comprehension/critical thinking: $p = 1.00$). The distribution of correctness by question type is summarised in Table 2.

Discussion

The PES evaluates both theoretical and practical competencies in diagnosing, treating, and preventing cardiovascular diseases. It covers the interpretation

of diagnostic tests, clinical decision-making, and current cardiology guidelines. Between 2009 and 2018, a total of 2333 candidates sat the exam, of whom 2210 passed, yielding a success rate of 94.73% [9].

In our study, we assessed the performance of the ChatGPT-o1-preview language model on the PES exam from the Spring 2023 session. The model was presented with all 120 single-choice questions and provided correct responses to 97 of them, corresponding to a score of 80.83% – well above the passing threshold.

Statistically significant relationships were found for both HDI and our introduced metric, CI-LM. The median HDI for correctly answered questions was 0.742, compared to 0.468 for incorrectly answered ones. A similar pattern was observed for CI-LM, with a median of 1.0 for correct responses and 0.8 for incorrect ones. These results suggest that questions more challenging for physicians were also more likely to be answered incorrectly by the model.

However, no statistically significant associations were found between model performance and question categories, such as clinical vs. other, or memory vs. comprehension/critical thinking.

The model demonstrated advanced data processing capabilities, enabling a more nuanced contextual understanding and analysis of complex medical information. These improvements over previous iterations probably contributed to the model's higher answer accuracy and increased precision in interpreting intricate cardiological concepts.

A 2024 study by Huwiler *et al.* evaluated the performance of several AI-powered chatbots available in Switzerland (as of June 2023) on a cardiology multiple-choice exam and compared their results to those of cardiology residents. The residents achieved a very high median success rate of 98%, with all passing the test. In contrast, the chatbots performed significantly worse, with a median score of 47%. Only one model, Jasper Quality, surpassed the passing threshold (73%). When image-based questions were excluded, performance slightly improved, and ChatGPT Plus 4.0 also exceeded the passing score with 76%. In comparison, our study demonstrated superior performance by the newer ChatGPT-o1-preview model (80.83%), suggesting a rapid improvement in language model capabilities in the context of cardiology-focused medical assessments. It should be noted that the Swiss examination included image-based questions, which may have posed additional challenges for AI systems [10].

Another relevant study by Skalidis *et al.* examined the performance of ChatGPT 3.5 on the European Exam in Core Cardiology (EECC), using 362 multiple-choice questions from three sources: the European Society of Cardiology (ESC), StudyPRN, and Braunwald's Heart Disease Review and Assessment (BHDRA). ChatGPT achieved an overall score of 58.8%, with specific results of 61.7% (ESC), 52.6% (BHDRA), and 63.8% (StudyPRN), thereby approaching or slightly surpassing the typical passing threshold of 60%. In contrast, the ChatGPT-o1-preview model achieved a considerably higher score in our study, highlighting the improved accuracy of newer versions on similar cardiology exams [11].

Finally, a 2024 study by Bielówka *et al.* evaluated ChatGPT-3.5's performance on the PES in Pathomorphology. The study analysed 119 questions by type (memory vs. comprehension/critical thinking) and topic. The model achieved a score of 45.38%, which fell below the passing threshold and was markedly lower than the score achieved in our cardiology-focused study. Bielówka *et al.* also found that the model performed better on comprehension/critical thinking questions (50%) than on memory-based questions (42.03%). In contrast, our findings showed comparable performance across both categories, possibly indicating improved integration of medical knowledge in newer language model versions [12].

This study highlights the growing potential of AI in medical education and evaluation. Future research involving broader question sets, diverse medical specialties, and comparisons across different language models will be essential to better understand their capabilities and define their potential role within healthcare systems.

Conclusions

This study evaluated the performance of the ChatGPT-o1-preview language model on the PES in cardiology. The model achieved a score of 80.83%, substantially exceeding the passing threshold of 60%. A statistically significant relationship was observed between question difficulty for human examinees and model accuracy; items that posed greater challenges for physicians were also more frequently answered incorrectly by the model. These findings demonstrate that, while the model can achieve performance well above the passing score in a specialist-level cardiology examination, its accuracy is still influenced by the inherent difficulty of the questions. Future research should assess its performance across broader exam formats, different medical specialties, and varied question types to further define its role in medical education and evaluation.

Funding

The publication of this article was supported by the Metropolitan Science Support Fund of the Górnośląsko-Zagłębiowska Metropolis (GZM). The research itself received no external funding.

Ethical approval

Not applicable.

Conflict of interest

The authors declare no conflict of interest.

References

1. OpenAI. Introducing ChatGPT. OpenAI (2022). Available at: https://openai.com/index/chatgpt/
2. Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article. Int J Surg. 2024; 110: 3701-3706.
3. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res. 2023; 25: e48568.
4. Sharma A, Medapalli T, Alexandrou M, Brilakis E, Prasad A. Exploring the role of ChatGPT in cardiology: a systematic review of the current literature. Cureus. 2024; 16: e58936.
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digital Health. 2023; 2: e0000198.
6. The Lancet Digital Health. ChatGPT: friend or foe? Lancet Digital Health. 2023; 5: e102.
7. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intelligence. 2023; 6: 1169595.
8. Anderson LW, Krathwohl DR. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Longman 2001.
9. Centrum Egzaminów Medycznych, "Specjalizacja – Kardiologia," CEM (2025). Available at: https://www.cem.edu.pl/aktualnosci/spece/spece_stat2.php?nazwa=Kardiologia
10. Huwiler J, Oechslin L, Biaggi P, Tanner FC, Wyss CA. Experimental assessment of the performance of artificial intelligence in solving multiple-choice board exams in cardiology. Swiss Medical Weekly. 2024; 154: 3547.
11. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, Fournier S. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digital Health. 2023; 4: 279-281.
12. Bielówka M, Kufel J, Rojek M, Kaczyńska D, Czogalik Ł, Mitręga A, Bartnikowska W, Kondoł D, Palkij K, Mielcarska S. An investigative analysis – ChatGPT's capability to excel in the Polish speciality exam in pathology. Pol J Pathol. 2024; 75: 236-240.

Address for correspondence

Natalia Denisiewicz
Students' Scientific Association
of Computer Analysis and
Artificial Intelligence at the
Department of Radiology
and Nuclear Medicine
Medical University of Silesia
Katowice, Poland
E-mail: dndenisiewicz@gmail.com

Received: 22.05.2025
Accepted: 25.08.2025
Online publication: 18.08.2026