smw · swiss medical weekly

Original article | Published 02 October 2024 | doi:https://doi.org/10.57187/s.3547
Cite this as: Swiss Med Wkly. 2024;154:3547

Experimental assessment of the performance of artificial intelligence in solving multiple-choice board exams in cardiology

Jessica Huwiler[ab], Luca Oechslin[a], Patric Biaggi[ab], Felix C. Tanner[bc], Christophe Alain Wyss[abc]

[a] Heart Clinic Zurich, Zurich, Switzerland
[b] University of Zurich, Zurich, Switzerland
[c] Swiss Society of Cardiology, Berne, Switzerland

Summary

AIMS: The aim of the present study was to evaluate the performance of various artificial intelligence (AI)-powered chatbots (commercially available in Switzerland up to June 2023) in solving a theoretical cardiology board exam and to compare their accuracy with that of human cardiology fellows.

METHODS: For the study, a set of 88 multiple-choice cardiology exam questions was used. The participating cardiology fellows and selected chatbots were presented with these questions. The evaluation metrics included Top-1 and Top-2 accuracy, assessing the ability of chatbots and fellows to select the correct answer.

RESULTS: Among the cardiology fellows, all 36 participants successfully passed the exam with a median accuracy of 98% (IQR 91–99%, range from 78% to 100%). However, the performance of the chatbots varied. Only one chatbot, Jasper quality, achieved the minimum pass rate of 73% correct answers. Most chatbots demonstrated a median Top-1 accuracy of 47% (IQR 44–53%, range from 42% to 73%), while Top-2 accuracy provided a modest improvement, resulting in a median accuracy of 67% (IQR 65–72%, range from 61% to 82%). Even with this advantage, only two chatbots, Jasper quality and ChatGPT plus 4.0, would have passed the exam. Similar results were observed when picture-based questions were excluded from the dataset.

CONCLUSIONS: Overall, the study suggests that most current language-based chatbots have limitations in accurately solving theoretical medical board exams. In general, currently widely available chatbots fell short of achieving a passing score in a theoretical cardiology board exam. Nevertheless, a few showed promising results. Further improvements in artificial intelligence language models may lead to better performance in medical knowledge applications in the future.

Prof. Dr. Christophe Wyss
HerzKlinik Hirslanden
Witellikerstrasse 40
CH-8032 Zurich
christophe.wyss[at]
hirslanden.ch

Introduction

Artificial intelligence (AI) encompasses techniques enabling machines to replicate human cognitive functions such as learning, reasoning and language understanding. AI and chatbots have evolved since the mid-20th century, marked by milestones like ELIZA [1] in the 1960s and IBM's Watson [2] in 2011. Despite setbacks, machine learning advancements in the 1980s and 1990s reignited interest. The internet facilitated rudimentary chatbot development, leading to a surge in AI-powered chatbots across sectors. Natural language processing blends linguistics and machine learning to enable computers to understand, interpret and generate human language, leveraging deep learning models like recurrent neural networks (RNNs) and transformers. Natural language processing has numerous applications, ranging from chatbots to virtual assistants to content recommendation systems, language translation tools, sentiment analysis for social media monitoring, and even healthcare applications in education [3], research and practice (e.g. medical record analysis [4, 5] and diagnosis assistance [6]).

Skalidis et al. [7] and Kung et al. [8] have already proven that ChatGPT can pass medical exams such as the European Exam in Core Cardiology in May 2023 or the USMLE in February 2023. Both studies worked with ChatGPT to solve the exams and while ChatGPT may currently be the most famous chatbot, there are several other AI chatbots such as Jaspar, Notion, YouBot, Bearly and CopyAI that have not been tested in the same way.

This manuscript aims to investigate the performance of artificial intelligence in solving multiple-choice board exams in cardiology by evaluating the accuracy of different chatbots in comparison to human graders. By presenting the findings of this study, we aim to contribute to the ongoing dialogue regarding the integration of AI in medical education and assessment, discussing the challenges associated with implementing AI in this context, including the need for data transparency, ethical considerations and the importance of preserving the human touch in medical education.

Methods

In Switzerland, the Cardiological Board Exam consists of both a theoretical and a practical exam. The theoretical exam includes multiple-choice questions, while the practical exam involves hands-on assessments and oral evaluations. In the present study, our aim was to compare the performance of different chatbots with that of cardiology fellows in solving the theoretical exam. We analysed and assessed how well chatbots (commercially available in Switzerland up to June 2023) performed relative to human fellows in this task.

Dataset

Since 2018, the theoretical board exam is an online-proctored examination (European Examination in Core Cardiology [EECC]), endorsed by the European Society of Cardiology (ESC). These datasets are confidential and inaccessible. Therefore, for the present study, we used a "historical" Swiss theoretical board exam, specifically the one used in 2017. Until 2017, the theoretical multiple-choice Cardiological Board Exam in Switzerland was conducted using a selected set of 88 multiple-choice questions covering various cardiology topics and subfields from the 10th Edition of Braunwald's Heart Disease Review and Assessment [9]. Each question in the dataset was accompanied by five answer choices, with only one choice being the correct answer. The pass threshold was set at 64 correct answers (73% correct answers). Participants who scored 64 or above were considered to have successfully passed the cardiology board exam, while those who scored below were deemed unsuccessful. Since all the included chatbots were language-based AI models, we carried out a separate analysis using a reduced dataset that excluded the 12 image-based questions, giving 76 questions in total.

Study population

Cardiology fellows

In 2017, 36 cardiology fellows were registered for the board exam. Only those fellows who had fulfilled the prerequisite educational requirements were eligible to take the cardiology board exam.

Chatbots

We selected 9 chatbots that were commercially available in June 2023 in Switzerland. Some were free to use (ChatGPT, Bearly, You) and some required a paid subscription (monthly rates of 10–49 US dollars [$]). Since information on the technical background of chatbots is not easily accessible, we asked the chatbots to describe themselves (notably, the texts below have been manually revised):

ChatGPT (https://openai.com/): ChatGPT (generative pretrained transformer) is an artificial intelligence language model developed by OpenAI. OpenAI is a company founded as a non-profit in 2015 by various investors. The generative models use a technology called deep learning, which pulls large amounts of data to train an artificial intelligence system to perform a task. ChatGPT is designed to provide natural language processing capabilities for generating human-like text responses. The language model is trained on various diverse text data, including internet sources, books, articles etc. The model has been fine-tuned using advanced techniques to enhance its language generation capabilities, enabling it to generate coherent and contextually relevant responses. The knowledge incorporated into ChatGPT was up-to-date as of September 2021, which serves as the model's knowledge cut-off. This means that ChatGPT may not have access to information or events that have occurred after that time. The web-based application with different model options being the same for the 3.5 and the 4.0 model, the accuracy of the answers may vary between the two models. ChatGPT is free while ChatGPT plus 3.5 and ChatGPT plus 4.0 subscribers pay a monthly fee of $20.

- *ChatGPT:* ChatGPT is a free version of OpenAI ChatGPT based on the GPT-3.5 architecture.
- *ChatGPT plus 3.5:* ChatGPT 3.5 plus is also based on the GPT-3.5 architecture and has been trained on an immense amount of text data from diverse sources. Again, the knowledge of GPT-3.5's training data only goes up to September 2021.
- *ChatGPT plus 4.0:* ChatGPT-4.0 is based on the GPT-4 architecture, an enhanced version of its predecessors, boasting a more extensive training dataset and significant improvements in its model architecture and training process.

GPT-4 and GPT-3.5 differ in several key areas: (1) Architecture: GPT-4 has a more advanced and complex architecture, likely with more parameters, allowing it to process information better and handle complex tasks. (2) Training: GPT-4 was trained on a larger and more diverse dataset using improved techniques, including advanced methods like Reinforcement Learning from Human Feedback (RLHF), making it more accurate and context-aware. (3) Capabilities: GPT-4 generalizes better across tasks, is more consistent in maintaining context, and handles nuances more effectively than GPT-3.5. (4) Safety: GPT-4 includes stronger safety measures to reduce problematic outputs, making it more reliable in sensitive contexts. In short, GPT-4 is a more powerful, accurate, and safer version of GPT-3.5, with enhanced architecture and training methods.

Jasper (https://www.jasper.ai/): The proprietary natural language processing model used to power Jasper is called GPT-J. It is based on the GPT-3 models by OpenAI and has been trained on a diverse range of text sources; nonetheless the responses are based purely on the content input provided by the user and the data it was trained on. It is also important to note that Jasper's responses are generated based on the data that was fed into its model and therefore may contain biases. According to the company, Jasper Chat has learned from billions of articles and other pieces of information before mid-2021 in 29 languages. Jasper is available for $39/month with a free 7-day trail. Jasper offers two different settings to choose from: "speed" or "quality". Both "speed" and "quality" settings are part of the same GPT-3 model used in Jasper AI but configured differently to meet specific user needs.

- *Jasper "speed":* When users choose the "speed" setting, the GPT-3 model within Jasper AI is configured to prioritise faster response times, which may sacrifice some content quality. The "speed" setting is designed for users who prioritise quick results.
- *Jasper "quality":* When users choose the "quality" setting, the GPT-3 model within Jasper AI is configured to focus more on generating higher-quality content, even if it takes slightly longer to produce the response. The "quality" setting is geared towards those who value more polished and well-structured content.

Notion (https://www.notion.so/): The artificial intelligence language model of Notion is built on top of the GPT-3 model developed by OpenAI. It can perform various tasks such as language processing, image recognition and data analysis. In contrast to other language models, Notion is constantly being updated with the latest information and data. However, the specific knowledge and information available to it may vary depending on the sources and algorithms used to train its model. Therefore, it is difficult to determine an exact timeframe or end date for when Notion will be up to date. Notion is cheaper than other models, at $10/month.

Bearly (https://bearly.ai/): Bearly artificial intelligence assistant is powered by ChatGPT. Although Bearly does not undergo individual updates, the underlying ChatGPT model is periodically improved and updated by OpenAI. This allows Bearly to benefit from advancements and refinements made to the model, ensuring it stays up-to-date with the latest developments. Bearly is an application that users can download for free and use without a subscription.

Copy.ai (https://www.copy.ai): Copy.ai is an AI-powered chatbot that uses GPT-3 models to autogenerate content based on the user’s input and a few sentences of context. Copy.ai provides the user with the sources of the information it gives which makes it easier for understanding the answers. As Copy.ai uses GPT-3 models from OpenAI, its limitations are the same as those of ChatGPT. The information the chatbot uses is not up-to-date and needs to be considered while using it. Copy.ai is the priciest chatbot at $49 monthly or $15 for 24 hours.

YouBot (https://you.com/): The language model that YouBot is based on was built internally by You.com, and details about the specific architecture of the model are not available. YouBot is constantly updated with new data, but it does not have a specific timeline or schedule for these updates. The model has some capacity limitations and can only process a limited amount of information at once. YouBot is free to use after subscribing.

Evaluation metrics

To assess the performance of the participants and the chatbots, two evaluation metrics focusing on accuracy were employed (primary outcome measures):

- Top-1 Accuracy: This metric measured the percentage of questions for which the chosen answer was the correct answer. A higher top-1 accuracy indicates a better performance in selecting the correct answer choice.
- Pass Rate: The pass rate was calculated to determine the percentage of participants who achieved a score of 64 or more correct answers (73% or more). This pass rate served as the threshold for successful completion of the cardiology board exam.

To better understand the performance of the chatbots, Top-2 accuracy was also analysed (secondary outcome measure):

- Top-2 Accuracy (only for chatbots): This metric evaluated the percentage of questions for which the correct answer was among the top two chosen answer choices. It measures the ability to include the correct answer within the top two ranked choices.

Top-2 accuracy is not applicable to real-life exam settings so could not be applied to the performance of participants.

Evaluation procedure

The same set of 88 multiple-choice cardiology board exam questions was provided to the participating cardiology fellows and chatbots:

- Participants selected one answer for each question. The evaluation metrics mentioned above were then calculated using the participants' answers and the ground truth correct answers provided in the dataset.
- Chatbots generated two ranked answers based on its learned patterns and associations. The same prompt was used for all chatbots (“*Please answer these cardiology exam questions with the most accurate answer [listed as first] and the second most accurate [listed as second]*”). The evaluation metrics were then calculated using the model's ranked list and the ground truth correct answers.

The same procedure was used for the reduced dataset (exclusion of 12 questions with pictures).

Statistics and data interpretation

Continuous data are expressed as median and interquartile range (IQR) or as mean ± standard deviation (SD), as appropriate, and categorical data as counts and percentages. Given that the data was not normally distributed (Shapiro-Wilk Normality Test Results p <0.05), the nonparametric Kruskal-Wallis test (with post-hoc Bonferroni correction) was used for comparing the performance of different chatbots. A p-value <0.05 was considered statistically significant. Statistical analyses were performed on DATAtab (DATAtab Team [2023]. DATAtab: Online Statistics Calculator. DATAtab e.U. Graz, Austria. https://datatab.de).

Results

Cardiology fellows

Since all 36 fellows were well prepared and the entire question catalogue was publicly accessible in advance [9], all of them passed the theoretical exam (100% pass rate). The median percentage of correct answers was 98% (IQR 91–99%, range from 78% to 100%). The accuracy of the fellows compared to chatbots is shown in figure 1.

Chatbots

Out of the 9 chatbots, only one (Jasper quality) was successful in passing the cardiology board exam with a very close result: 64 correct answers out of 88 (73% correct answers, minimal requirement/pass rate threshold). Overall, most of the chatbots performed poorly – the median percentage of correct answers (Top-1 Accuracy) was 47% (IQR 44–53%, range from 42% to 73%). In direct comparison, there were overall no statistically significant differences between chatbot performances (p = 0.433). Only Jasper quality performed statistically significantly better than Jasper speed in post hoc testing (p = 0.037) (see appendix).

For Top-2 Accuracy, the probability of correct answers was augmented by an average of 18% (range 9–24% depending on the chatbot): the median percentage of correct answers (Top-2 Accuracy) was 67% (IQR 65–72%, range from 61% to 82%). Nevertheless, even with this advantageous assumption, only two chatbots (Jasper quality, ChatGPT plus 4.0) would have passed the theoretical exam. The comparative accuracy of the chatbots is shown in figure 2.

With the reduced dataset to 76 questions (exclusion of 12 questions with pictures), the results were quite similar: the median percentage of correct answers (Top-1 Accuracy) was 49% (IQR 46–51%, range from 45% to 80%). At a pass threshold of 73% correct answers, only Jasper quality (80%) and ChatGPT plus 4.0 (76%) would have passed the theoretical exam (see figure 3). Comparative results in both datasets are shown in table 1.

Figure 1: Accuracy of the fellows compared to chatbots (Top-1 and Top-2 accuracy). The dashed line is set at a pass rate of 64 correct answers out of 88 (73% correct answers, minimal requirement/pass rate threshold).

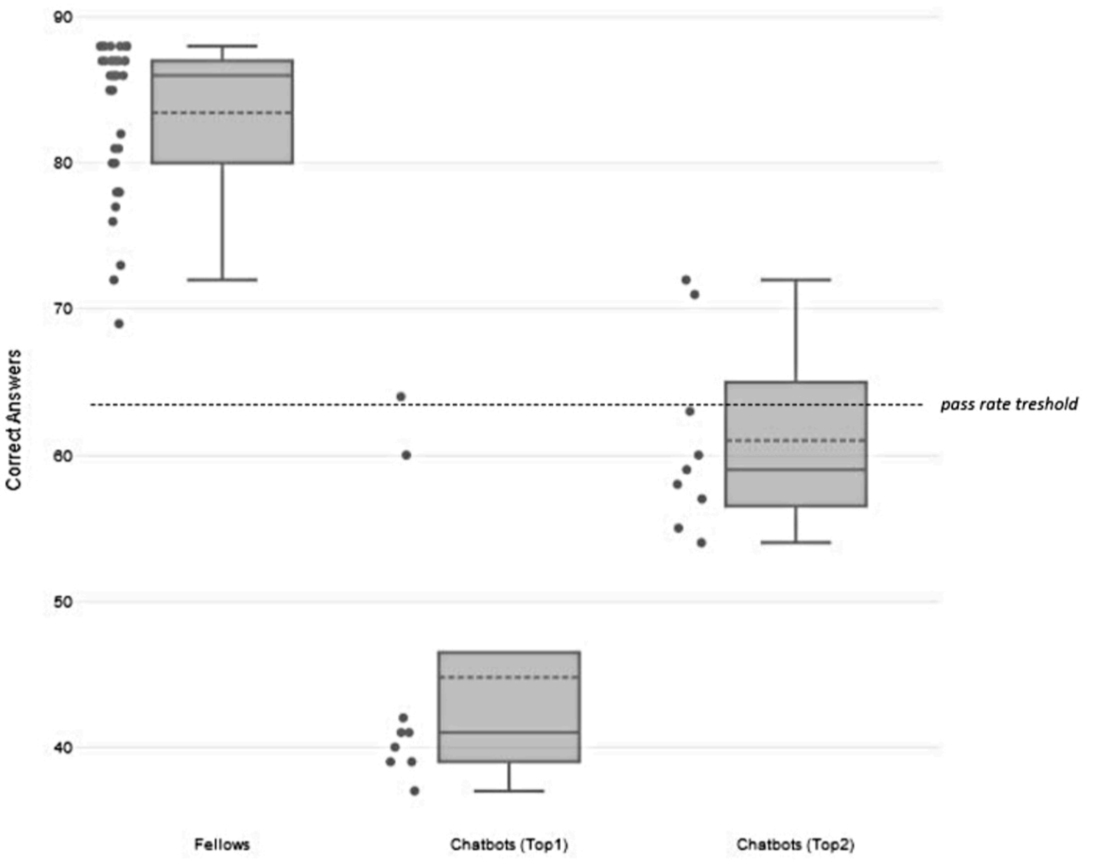
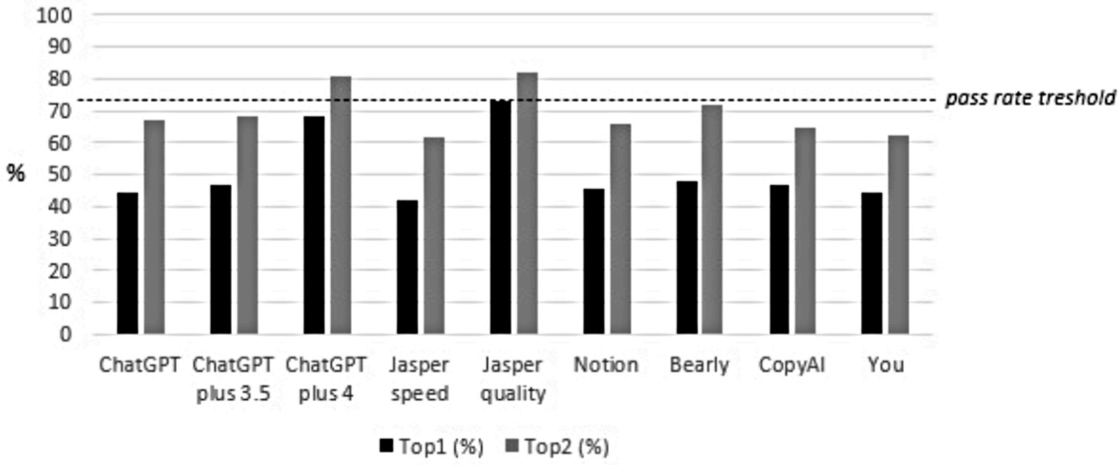


Figure 2: Accuracy of different chatbots (Top-1 and Top-2 accuracy). The dashed line is set at a pass rate of 73% correct answers (minimal requirement/pass rate threshold).



Discussion

The present study aimed to compare the performance of different chatbots with cardiology fellows in solving a theoretical cardiology board exam. All 36 cardiology fellows who participated in the study passed the theoretical exam, resulting in a 100% pass rate. The median percentage of correct answers among the fellows was 98%, indicating a high level of preparedness and competence in cardiology. Out of the 9 chatbots tested, only one (Jasper quality) was able to achieve the minimal pass rate threshold of 73% correct answers. When the dataset was reduced to 76 questions by excluding the 12 image-based questions, two chatbots exceeded the pass threshold (Jasper Quality 80%, Chat GPT plus 4.0 76%). The other chatbots performed poorly, with a median Top-1 Accuracy of 47%. This indicates that most chatbots struggled to accurately select the correct answer choice for the multiple-choice questions. When evaluating the Top-2 Accuracy metric, which measures whether the correct answer was among the top two chosen answer choices, the chatbots showed some improvement, with an average increase of 18% in the probability of selecting the correct answer. However, even with this advantage, again only two chatbots (Jasper quality and ChatGPT plus 4.0) would have passed the theoretical exam.

Chatbots appear to have transformative potential, since the ability to pass reputable exams has been previously published in other countries and subspecialties: ChatGPT provided highly accurate answers to the US Certified Public Accountant exam and the US bar exam [10, 11] and achieved the passing criteria for the US Medical Licensing Examination (USMLE) [8, 12]. In an ophthalmology examination, Antaki et al. showed that ChatGPT currently performed at the level of an average first-year resident [13]. A recent study [14] compared the performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination (JMLE) and evaluated the reliability of these models for clinical reasoning and medical knowledge in non-English languages: in terms of the correct response rate for individual questions, the examinees' rate for essential knowledge questions was 89.2% compared to 87.2% for GPT-4 (55.1% correct response rate of GPT-3.5). In all cases, GPT-4 achieved the passing rates for the JMLE. However, none of these rates exceeded the total percentage of correct answers by examinees. In cardiology, a recent study by Skalidis et al. [7] evaluated ChatGPT's ability to answer European Examination in Core Cardiology questions sourced from various official materials (questions containing audio or visual elements such as clinical images, charts, tables, and videos were excluded). The model accurately answered 340 out of 362 questions, achieving an overall accuracy of 58.8%. Notably, its accuracy varied across different sources, with performance ranging from 52.6% to 63.8%. The criteria for passing the EECC are established based on candidates' performance in the administered exam [15]. Over recent years, the pass mark for the

Figure 3: Accuracy of the different chatbots (Top-1 and Top-2 accuracy) in the reduced dataset of 76 questions (exclusion of 12 questions with images). The dashed line is set at a pass rate of 73% correct answers (minimal requirement/pass rate threshold).

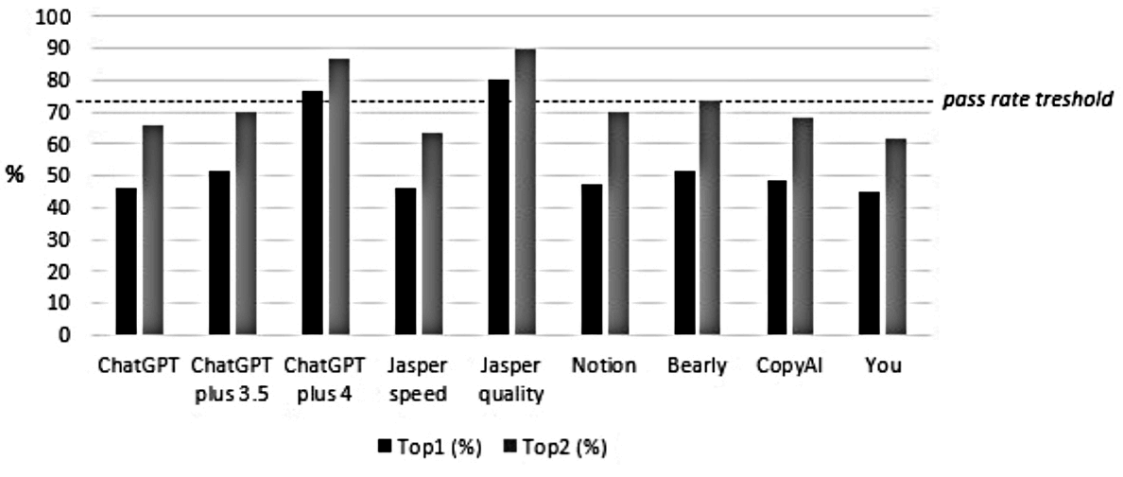


Table 1:
Comparative accuracy of the chatbots.

	Full dataset (n = 88)				Reduced dataset (n = 76)			
	Top-1 accuracy		Top-2 accuracy		Top-1 accuracy		Top-2 accuracy	
	n	%	n	%	n	%	n	%
ChatGPT	39	44%	59	67%	35	46%	50	66%
ChatGPT plus 3.5	41	47%	60	68%	39	51%	53	70%
ChatGPT plus 4.0	60	68%	71	81%	58	76%	66	87%
Jasper speed	37	42%	54	61%	35	46%	48	63%
Jasper quality	64	73%	72	82%	61	80%	68	89%
Notion	40	45%	58	66%	36	47%	53	70%
Bearly	42	48%	63	72%	39	51%	56	74%
CopyAI	41	47%	57	65%	37	49%	52	68%
You	39	44%	55	63%	34	45%	47	62%

EECC has ranged from 65 (54%) to 70 (58%) correct responses out of 120 questions. Plummer et al. [16] questioned whether ChatGPT would be capable of meeting the requirements of the current EECC: some of the questions analysed by Skalidis et al. [7] differ significantly from those in the EECC (structure, format, editorial process, etc.). Moreover, ChatGPT would likely be unable to tackle 36 (30%) of the 120 questions in the EECC that incorporate visual elements like images or video clips. With an overall accuracy of 58.8% in text-based questions, it is anticipated that ChatGPT would achieve 49 correct responses (41%), falling below any previous passing threshold. As the actual EECC datasets are confidential and inaccessible, we used in the present study a historical Swiss theoretical board exam setting from 2017 (previously predefined questions, higher pass rate threshold).

Despite the possibility of misuse resulting in academic dishonesty [17] and the limitations discussed below, there is clear positive potential of chatbots in medical education [3]: Artificial intelligence can identify flaws in medical education [18], can contribute to tailoring education based on the needs of the student with immediate feedback [19] and can rapidly craft consistent realistic clinical vignettes of variable complexities as a valuable educational source with lower costs [20]. Furthermore, chatbot use can be considered as a motivation in healthcare education based on personalised interaction, either as a self-learning tool or as an adjunct to group learning [8, 21, 22].

Limitations

It's important to consider the limitations of the studied chatbots, such as their knowledge cut-off of September 2021 and their reliance on the data they were trained on. The knowledge cut-off should be taken into consideration when interpreting the responses generated by the studied chatbots, as they may not reflect the most recent developments or advancements in the field. Until 2017, the theoretical multiple-choice Cardiological Board Exam in Switzerland was conducted using a selected set of multiple-choice questions covering various cardiology topics and subfields from the 10th edition of Braunwald's Heart Disease Review and Assessment [9]. Although it was published in 2015, the chatbots seemed not to have been trained on this specific text regarding the poor performance. This reflects a clear data selection bias. Nevertheless, since for exam preparation the whole question pool was previously accessible to all fellows, they all had – by definition – a huge selection bias too. Overall, this may lead to an inherently biased comparison.

Besides selection bias, another limitation is a time bias: evidence has evolved since 2015, updated guidelines have been published since then and what could be considered the most accurate answer at that time may no longer have been the first choice in 2021 (the knowledge cutoff of the Chatbots is September 2021). Furthermore, didactic approaches and medical education assessment are under constant evolution – since chatbots have been trained with supervised learning and reinforcement learning from human feedback, outdated linguistic patterns may influence the answering behaviour. We therefore reassessed all 88 questions if "best option/correct answer" would still be applicable in 2023: no incorrect answer was identified under application of latest guidelines (e.g. infective endocarditis, cardiomyopathies). Accordingly, performance itself is not affected. Nevertheless, 6 questions (<10%) appeared didactically outdated and may be linguistically confusing.

As for every artificial intelligence model, the answers are only as accurate as the initial question (prompt question). The chatbot can only provide the user with the answers based on the given information and question, meaning it is important to understand that the initial question must be accurate and specific to get the best results. In the present study, we did not evaluate the effects of different initial questions (no prompt engineering).

The findings of the present study are not applicable to the current Cardiological Board Exam in Switzerland. Since 2018, the theoretical board exam is an online-proctored examination (European Examination in Core Cardiology), endorsed by the European Society of Cardiology. These datasets are confidential and inaccessible. In the present study, we used a historical Swiss theoretical board exam context from 2017 (previously predefined questions, higher pass threshold), which no longer fits with modern medical education assessment.

Conclusion

Overall, the study suggests that most current language-based chatbots have limitations in accurately solving theoretical medical board exams, especially when compared with well-prepared human experts such as cardiology fellows. An unselected application of currently widely available chatbots fell short of achieving a passing score in a theoretical cardiology board exam. Nevertheless, selected and more advanced chatbots/language models showed promising results. Further research and improvements in artificial intelligence language models may lead to better performance in medical knowledge application in the future. However, it remains so far essential to rely on human expertise and judgement for critical tasks.

Acknowledgments

We would like to thank Karin Guldenfels and Marjam Rüdiger (Swiss Society of Cardiology [SSC]) for assistance in data management and administration.

Author contributions: JH and CW contributed to the conception of the work and JH, MR, KG to the acquisition of data. JH and CW contributed to the analysis and interpretation of data for the work and drafted the manuscript. FT, PB and LO critically revised the manuscript. All gave final approval and agreed to be accountable for all aspects of work ensuring integrity and accuracy.

This manuscript incorporates segments of text enhanced with the assistance of the ChatGPT language model developed by OpenAI. The purpose of incorporating ChatGPT in this manuscript was to facilitate various aspects of writing, including drafting, brainstorming and generating textual content. ChatGPT contributed to syntactical improvements and language enhancements in certain sections, guided by the authors' review to ensure congruence with the research objectives. The research findings and conclusions remain unaffected. This disclosure is consistent with the ICMJE recommendation of acknowledging writing assistance [23].

Potential competing interests

All authors have completed and submitted the International Committee of Medical Journal Editors form for disclosure of potential conflicts of interest. No potential conflict of interest related to the content of this manuscript was disclosed.

References

1. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM. 1966;9(1):36–45. http://dx.doi.org/10.1145/365153.365168.
2. Ferrucci D, Brown E, Chu-Carroll J, Fan J, Gondek D, Kalyanpur AA, et al. Building Watson: An Overview of the DeepQA Project. AI Mag. 2010 Jul;31(3):59–79. http://dx.doi.org/10.1609/aimag.v31i3.2303.
3. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023 Mar;11(6):887. http://dx.doi.org/10.3390/healthcare11060887.
4. Patra BG, Sharma MM, Vekaria V, Adekkanattu P, Patterson OV, Glicksberg B, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. J Am Med Inform Assoc. 2021 Nov;28(12):2716–27. http://dx.doi.org/10.1093/jamia/ocab170.
5. Byrd RJ, Steinhubl SR, Sun J, Ebadollahi S, Stewart WF. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. Int J Med Inform. 2014 Dec;83(12):983–92. http://dx.doi.org/10.1016/j.ijmedinf.2012.12.005.
6. Li C, Zhang Y, Weng Y, Wang B, Li Z. Natural Language Processing Applications for Computer-Aided Diagnosis in Oncology. Diagnostics (Basel). 2023 Jan;13(2):286. http://dx.doi.org/10.3390/diagnostics13020286.
7. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health. 2023 Apr;4(3):279–81. http://dx.doi.org/10.1093/ehjdh/ztad029.
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023 Feb;2(2):e0000198. http://dx.doi.org/10.1371/journal.pdig.0000198.
9. Lilly LS. Braunwald's Heart Disease Review and Assessment E-Book. Elsevier Health Sciences; 2015.
10. Bommarito J, Bommarito M, Katz DM, Katz J. "Gpt as knowledge worker: A zero-shot evaluation of (ai) cpa capabilities," arXiv preprint arXiv:2301.04408, 2023.
11. Bommarito M 2nd, Katz DM. "GPT takes the bar exam," arXiv preprint arXiv:2212.14402, 2022.
12. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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," (in English), *JMIR Med Educ*. JMIR Med Educ. 2023;9:e45312. http://dx.doi.org/10.2196/45312.
13. Fares A, Samir T, Daniel M, Jonathan EK, Renaud D. "Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of its Successes and Shortcomings," medRxiv, p. 2023.01.22.23284882, 2023, doi: http://dx.doi.org/10.1101/2023.01.22.23284882.
14. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison Study. JMIR Med Educ. 2023 Jun;9:e48002. http://dx.doi.org/10.2196/48002.
15. Plummer C, et al. "Behind the scenes of the European Examination in General Cardiology," Heart, vol. 105, pp. heartjnl-2018, 02/02 2019, doi: http://dx.doi.org/10.1136/heartjnl-2018-314495.
16. Plummer C, Mathysen D, Lawson C. Does ChatGPT succeed in the European Exam in Core Cardiology? Eur Heart J Digit Health. 2023 Jul;4(5):362–3. http://dx.doi.org/10.1093/ehjdh/ztad040.
17. Cotton DR, Cotton PA, Shipway JR. Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innov Educ Teach Int. 2023;61(2):228–39. http://dx.doi.org/10.1080/14703297.2023.2190148.
18. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health. 2023 Feb;2(2):e0000205. http://dx.doi.org/10.1371/journal.pdig.0000205.
19. van Dis EA, Bollen J, Zuidema W, van Rooij R, Bockting CL. "ChatGPT: five priorities for research," in Nature, vol. 614, no. 7947). England, 2023, pp. 224-226. http://dx.doi.org/10.1038/d41586-023-00288-7.
20. James RA. "ChatGPT for Clinical Vignette Generation, Revision, and Evaluation," medRxiv, p. 2023.02.04.23285478, 2023, doi: http://dx.doi.org/10.1101/2023.02.04.23285478.
21. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American heart association course? Resuscitation. 2023 Apr;185:109732. http://dx.doi.org/10.1016/j.resuscitation.2023.109732.
22. Maurer S. "ChatGPT, schreib meine Zusammenfassung," ed: Schweiz Ärzteztg. 2023;104(33):16-20. http://dx.doi.org/10.4414/saez.2023.21992.
23. "Non-Author Contributors, Defining the Role of Authors and Contributors. ICMJE. Available from: https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html [Last accessed on 2023 Aug 17].", ed, 2023.

Appendix: comparative statistics

A Kruskal-Wallis test showed that there was no significant difference between the categories of the independent variable with respect to the dependent variable *performance*, p = 0.433.

Groups	n	Median	Mean rank
ChatGPT	1	39	2.5
ChatGPT plus 3.5	1	41	5.5
ChatGPT plus 4.0	1	60	8
Jasper speed	1	37	1
Jasper quality	1	64	9
Notion	1	40	4
Bearly	1	42	7
CopyAI	1	41	5.5
You	1	39	2.5
Total	9	41	

AI: Artificial intelligence.

	Chi^2	df	p
Perfomance	8	8	0.433

Post hoc test.

	Test statistic	Std. error	Std. test statistic	p	Adj. p
ChatGPT – ChatGPT plus 3.5	–3	3.84	–0.78	0.435	1
ChatGPT – ChatGPT plus 4.0	–5.5	3.84	–1.43	0.152	1
ChatGPT – Jasper speed	1.5	3.84	0.39	0.696	1
ChatGPT – Jasper quality	–6.5	3.84	–1.69	0.091	1
ChatGPT – Notion	–1.5	3.84	–0.39	0.696	1
ChatGPT – Bearly	–4.5	3.84	–1.17	0.241	1
ChatGPT – CopyAI	–3	3.84	–0.78	0.435	1
ChatGPT – You	0	3.84	0	1	1
ChatGPT plus 3.5 – ChatGPT plus 4.0	–2.5	3.84	–0.65	0.515	1
ChatGPT plus 3.5 – Jasper speed	4.5	3.84	1.17	0.241	1
ChatGPT plus 3.5 – Jasper quality	–3.5	3.84	–0.91	0.362	1
ChatGPT plus 3.5 – Notion	10.5	30.84	00.39	0.696	1
ChatGPT plus 3.5 – Bearly	–1.5	3.84	–0.39	0.696	1
ChatGPT plus 3.5 – CopyAI	0	3.84	0	1	1
ChatGPT plus 3.5 – You	3	3.84	0.78	0.435	1
ChatGPT plus 4.0 – Jasper speed	7	3.84	1.82	0.068	1
ChatGPT plus 4.0 – Jasper quality	–1	3.84	–0.26	0.795	1
ChatGPT plus 4.0 – Notion	4	3.84	1.04	0.298	1
ChatGPT plus 4.0 – Bearly	1	3.84	0.26	0.795	1
ChatGPT plus 4.0 – CopyAI	2.5	3.84	0.65	0.515	1
ChatGPT plus 4.0 – You	5.5	3.84	1.43	0.152	1
Jasper speed – Jasper quality	–8	3.84	–2.08		1
Jasper speed – Notion	–3	3.84	–0.78	0.435	1
Jasper speed – Bearly	–6	3.84	–1.56	0.118	1
Jasper speed – CopyAI	–4.5	3.84	–1.17	0.241	1
Jasper speed – You	–1.5	3.84	–0.39	0.696	1
Jasper quality – Notion	5	3.84	1.3	0.193	1
Jasper quality – Bearly	2	3.84	0.52	0.603	1
Jasper quality – CopyAI	3.5	3.84	0.91	0.362	1
Jasper quality – You	6.5	3.84	1.69	0.091	1
Notion – Bearly	–3	3.84	–0.78	0.435	1
Notion – CopyAI	–1.5	3.84	–0.39	0.696	1
Notion – You	1.5	3.84	0.39	0.696	1
Bearly – CopyAI	1.5	3.84	0.39	0.696	1
Bearly – You	4.5	3.84	1.17	0.241	1
CopyAI – You	3	3.84	0.78	0.435	1

Adj. p: Values adjusted with Bonferroni correction; AI: Artificial intelligence.