ESC European Society of Cardiology

European Heart Journal - Digital Health (2023) **4**, 279–281
https://doi.org/10.1093/ehjdh/ztad029

SHORT REPORT

ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story?

Ioannis Skalidis [1], Aurelien Cagnina[1], Wongsakorn Luangphiphat[1], Thabo Mahendiran[1,2], Olivier Muller [1], Emmanuel Abbe[2], and Stephane Fournier [1,*]

[1]Cardiology Department, University Hospital of Lausanne, Rue du Bugnon 46, 1011 Lausanne, Switzerland; and [2]Institute of Mathematics and School of Computer and Communication Sciences, EPFL, EPFL FSB SMA, Station 8,1015 Lausanne, Switzerland

Received 18 February 2023; revised 12 April 2023; accepted 21 April 2023; online publish-ahead-of-print 24 April 2023

Chat Generative Pre-trained Transformer (ChatGPT) is currently a trending topic worldwide triggering extensive debate about its predictive power, its potential uses, and its wider implications. Recent publications have demonstrated that ChatGPT can correctly answer questions from undergraduate exams such as the United States Medical Licensing Examination. We challenged it to answer questions from a more demanding, post-graduate exam—the European Exam in Core Cardiology (EECC), the final exam for the completion of specialty training in Cardiology in many countries. Our results demonstrate that ChatGPT succeeds in the EECC.

Keywords Artificial Intelligence • ChatGPT • Medical education • Machine learning • Large language models

Recently, artificial intelligence (AI) has flourished and established its role in various industries, encompassing a broad variety of subfields in natural language processing and computer vision, and proposing a revolutionary way to approach various tasks and problems.[1] Over the past few months, a novel AI tool known as ChatGPT (Chat Generative Pre-trained Transformer) has seen a surge of interest and has been subject to endless speculation about its predictive capabilities, with at least 10 authorships in peer-reviewed scientific journals at the time of writing.[2,3] It represents the most advanced version of a large language model (LLM) launched by OpenAI in November 2022, and it operates as a text-based chatbot interface. Chat Generative Pre-trained Transformer is based on a large artificial neural network model called a transformer that is trained to predict the most likely text outputs on prompts from a massive amount of text available on the internet. It distinguishes itself from the previous LLMs through its strong interactive capabilities and its large scale.

Chat Generative Pre-trained Transformer performance on exams and the European Exam in Core Cardiology

Recent preliminary studies have shown that ChatGPT can achieve a passing grade in exams such as the United States Medical Licensing Examination and the final exam in an MBA course at the University of Pennsylvania.[4,5] However, its ability to succeed in a more challenging post-graduate exam such as the European Exam in Core Cardiology (EECC), the final exam for the completion of specialty training in Cardiology in many countries, is not known.[6] The EECC is a knowledge-based assessment designed to provide a broad, balanced and up-to-date test of the core cardiology knowledge [detailed in the European Society of Cardiology (ESC) Core Curriculum for the Cardiologist] that is required by cardiology specialty trainees for independent practice. It consists of 120 multiple choice questions (MCQs) covering the whole spectrum of cardiology, testing knowledge of pathophysiology, clinical reasoning, and guideline-recommended medical management. It is considered the final theoretical exam for the completion of cardiology speciality training in numerous countries. The pass mark, while varying by year, is approximately 60%.

We evaluated the performance of ChatGPT on the EECC in order to assess its predictive power on a more challenging, high-level, post-graduate exam.

Input source

As a source of MCQs, we used sample exam questions released since 2018 from the official ESC website, as well as the 2022 edition of StudyPRN and Braunwald's Heart Disease Review and Assessment (BHDRA). All of these sources represent the traditionally used preparation material for the EECC.

* Corresponding author. Tel: +41 79 556 82 05, Email: stephane.fournier@chuv.ch

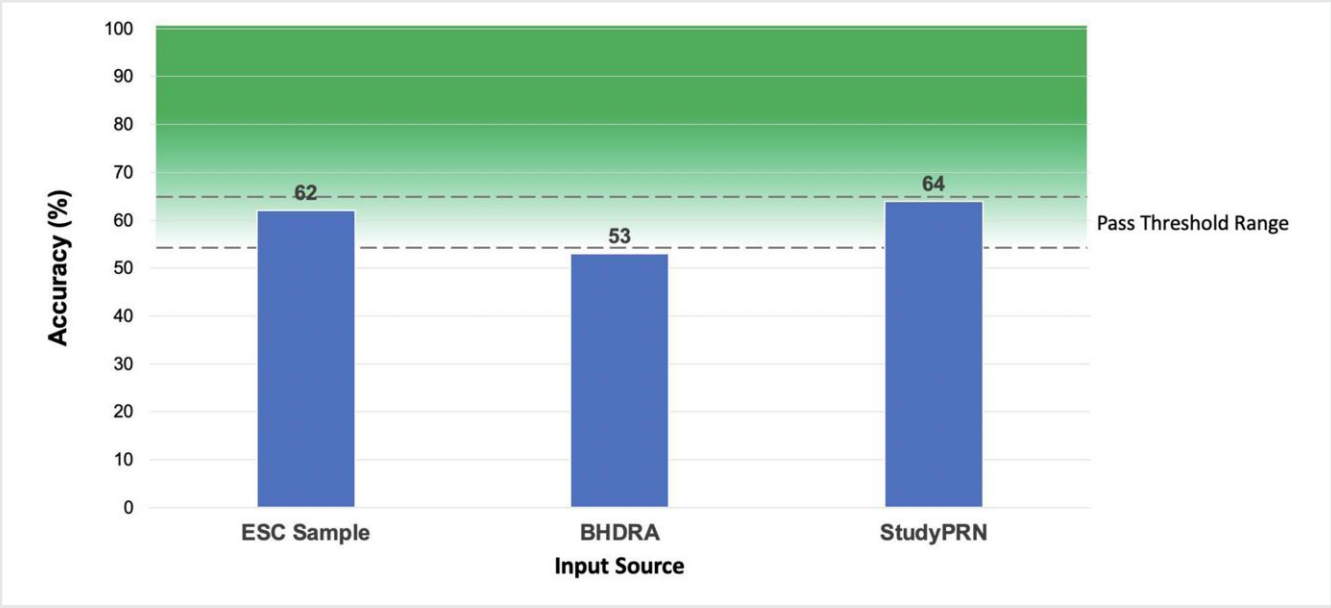

Figure 1 Performance of Chat Generative Pre-trained Transformer across the different multiple choice questions sources. ESC sample, questions released as sample exam from ESC official website; ESC, European Society of Cardiology; BHDRA, Braunwald's Heart Disease Review and Assessment; ChatGPT, Chat Generative Pre-trained Transformer; MCQ, multiple choice questions.

A total of 488 publicly-available single-answer MCQs were randomly obtained, consisting of 88 from the ESC samples, 200 from the StudyPRN test questions, and 200 from the BHDRA. All MCQs were screened, and questions containing audio or visual elements such as clinical images, charts, tables, and videos were excluded. After filtering, 362 MCQ items (ESC sample: 68, BHDRA: 150, and StudyPRN: 144) were included in the final question bank to be used as the input for ChatGPT. With regard to the output from ChatGPT, responses that were clearly incorrect or indeterminate were both considered incorrect for the present study.

In order to evaluate ChatGPT's performance, we submitted each question from the question bank to the AI model in the form of a text prompt. To facilitate interaction with ChatGPT, we utilized OpenAI's online platform, where the questions were inserted directly into the provided interface. Upon receiving the responses generated by ChatGPT, we rigorously compared them with the correct answers from the source material to ascertain the model's overall accuracy in answering the EECC questions.

Chat Generative Pre-trained Transformer performance

Chat Generative Pre-trained Transformer answered 340 questions out of 362, with 22 indeterminate answers in total. The overall accuracy was 58.8% across all the question sources. More specifically, it demonstrated an accuracy with the ESC sample, BHDRA, and StudyPRN of 61.7%, 52.6%, and 63.8%, respectively. It correctly answered 42/68 (4 indeterminate) of the ESC sample questions, 79/150 (11 indeterminate) of the BHDRA questions, and 92/144 (7 indeterminate) of the StudyPRN questions (*Figure 1*).

Is passing a medical exam enough?

Chat Generative Pre-trained Transformer correctly answered the majority of questions and shows consistency across all different MCQs sources exceeding 60% in most analyses. Although the EECC pass mark depends on the overall performance of the candidates, it is typically around 60%. Consequently, based on these results, ChatGPT achieves a score above or near the pass mark.

The EECC is a challenging exam, and the average candidate has completed approximately 6 years of medical school, 1–2 years of general medical training, and 3–4 years of cardiology speciality training. In addition, candidates typically spend months revising for the exam. However, our results demonstrate that ChatGPT, which currently represents a beta version and not yet the final product, was able to achieve a score in the range of the exam's historical pass mark. European Exam in Core Cardiology questions consist of text vignettes with nuanced scenarios that require deductive reasoning. As a result, a rational approach and a significant amount of knowledge are required to successfully answer the questions, explaining the success of ChatGPT with this particular task.

Importantly, ChatGPT is designed for natural language processing tasks and thus currently only accepts text-based inputs, resulting in the exclusion of all questions with image content. This represents a major limitation as such questions constitute approximately 25% of the total QCMs. It is possible that future AI-based tools could manage more diverse inputs such as images and videos, increasing their utility. With regard to future work, an analysis that stratifies the accuracy of ChatGPT by subject (e.g. coronary disease, arrhythmias, vascular diseases, haemodynamics, etc.) would be of great interest.

While the results are surely intriguing and impressive, ChatGPT still has to be thoroughly explored in order to find its role in medical standardized examinations that have emerged as an indispensable part of medical training.[7,8] The major challenge for the future is to investigate the possible ways AI programs, and LLMs in general, could serve as useful tools in medical education, research, and even clinical decision making. As the models evolve in the near future, it is beyond doubt that the scientific world will aim to explore its full potential.[9,10]

Conclusion

We are entering an era where AI and LLMs, such as ChatGPT, are reaching a maturity level that will gradually have an impact in healthcare.

The growing body of evidence demonstrating that ChatGPT succeeds in medical licensing exams could lead to an eventual transformation in medical training. However, it is crucial to highlight ChatGPT's strengths without overlooking its limitations: We have shown that it is able to effectively process medical information and provide appropriate answers to questions, however, it is currently not a substitute for critical thinking, innovation, and creativity, some of the key attributes that doctors are expected to showcase.

Funding

None declared.

Conflict of interest: None declared.

Data availability

Data will be made available by the corresponding author for reasonable requests.

References

1. Liu PR, Lu L, Zhang JY, Huo TT, Liu SX, Ye ZW. Application of artificial intelligence in medicine: an overview. *Curr Med Sci* 2021;**41**:1105–1115. Epub 2021 Dec 6. PMID: 34874486.
2. Castelvecchi D. Are ChatGPT and AlphaCode going to replace programmers? *Nature* 2022. doi:10.1038/d41586- 022-04383-z. Epub ahead of print. PMID: 36481949.
3. Chatterjee J, Dethlefs N. This new conversational AI model can be your friend, philosopher, and guide … and even your worst enemy. *Patterns (N Y)* 2023;**4**: 100676.
4. 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;**2**:e0000198. doi:10.1371/journal.pdig.0000198. PMID: 36812645; PMCID: PMC9931230.
5. Terwiesch Christian. Would Chat GPT get a Wharton MBA? A prediction based on its performance in the Operations Management course. Mack Institute for Innovation Management, Wharton School, University of Pennsylvania. https://mackinstitute.wharton.upenn.edu/2023/would-chat-gpt3-get-a-wharton-mba-new-white-paper-by-christian-terwiesch/
6. About the European Exam in Core Cardiology (EECC). https://www.escardio.org/Education/Career-Development/European-Exam-in-Core-Cardiology-(EECC)
7. Thorp HH. ChatGPT is fun, but not an author. *Science* 2023;**379**:313. Epub 2023 Jan 26. PMID: 36701446.
8. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, *et al.* ChatGPT and other large language models are double-edged swords. *Radiology* 2023;230163.
9. Gordijn B, Have HT. ChatGPT: evolution or revolution? *Med Health Care Philos* 2023; **26**:1–2. doi:10.1007/s11019-023-10136-0. Epub ahead of print. PMID: 36656495
10. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. *Nature* 2023;**613**:620–621. PMID: 36653617.