Ann Ist Super Sanità 2023 | Vol. 59, No. 4: 267-270
DOI: 10.4415/ANN_23_04_05

Exploring the potential of ChatGPT for clinical reasoning and decision-making: a cross-sectional study on the Italian Medical Residency Exam

Giacomo Scaioli, Giuseppina Lo Moro, Francesco Conrado, Lorenzo Rosset, Fabrizio Bert and Roberta Siliquini

Dipartimento di Scienze della Sanità Pubblica e Pediatriche, Università degli Studi di Torino, Turin, Italy

Abstract

Background. This study aimed to assess the performance of ChatGPT, a large language model (LLM), on the Italian State Exam for Medical Residency (SSM) test to determine its potential as a tool for medical education and clinical decision-making support.

Materials and methods. A total of 136 questions were obtained from the official SSM test. ChatGPT responses were analyzed and compared to the performance of medical doctors who took the test in 2022. Questions were classified into clinical cases (CC) and notional questions (NQ).

Results. ChatGPT achieved an overall accuracy of 90.44%, with higher performance on clinical cases (92.45%) than on notional questions (89.15%). Compared to medical doctors' scores, ChatGPT performance was higher than 99.6% of the participants.

Conclusions. These results suggest that ChatGPT holds promise as a valuable tool in clinical decision-making, particularly in the context of clinical reasoning. Further research is needed to explore the potential applications and implementation of large language models (LLMs) in medical education and medical practice.

Key words

- ChatGPT
- medical education
- clinical reasoning
- artificial intelligence
- examination performance

INTRODUCTION

In recent years, there has been a growing interest in the application of artificial intelligence (AI) in healthcare. AI technologies have been applied in different fields of medicine, showing promising results [1-3]. AI has the potential to overcome errors made by doctors in decision-making, which are due to lack of knowledge, the so-called "salient distracting clinical features", and irrelevant factors, such as current mood, time since the last meal, or the weather [3, 4]. One promising development in this area is the use of large language models (LLMs), such as ChatGPT, to assist in clinical reasoning and decision-making [5].

ChatGPT is a general LLM developed by OpenAI, an organization founded in December 2015, with the primary objective of promoting responsible and beneficial applications of artificial general intelligence for society, that has been trained on a massive corpus of text data from the internet via reinforcement and supervised learning methods. Unlike traditional rule-based systems, LLMs can process natural language input and generate output similar to human-generated text (https://openai.com/about). ChatGPT, specifically, has garnered significant attention due to its ability to perform a diverse array of natural language tasks, and exhibit evidence of deductive reasoning, the chain of thought, and long-term dependency skills (https://openai.com/).

Although initially it seemed that research on ChatGPT's role in clinical settings was sparse, recent literature indicates a burgeoning interest in this area. Studies like Liu *et al.* [6] highlighted the utility of ChatGPT in routine clinical practice. At the same time, Ferdush *et al*. delved into the broader implications, applications, and limitations of ChatGPT in clinical decision support [6, 7]. Particularly notable is the work by Alessandri Bonetti *et al.*, which specifically investigated ChatGPT's performance on the Italian Residency Admission National Exam, drawing comparisons to a vast cohort of medical graduates [8].

The purpose of this study is to determine its potential as a tool for clinical decision-making support, by assessing the performance of ChatGPT on a test performed by graduated medical students to become medical resi-

Address for correspondence: Francesco Conrado, Dipartimento di Scienze della Sanità Pubblica e Pediatriche, Università degli Studi di Torino, Via Santena 5 bis, 10126 Turin, Italy. E-mail: francesco.conrado@unito.it.

dents (the Italian State Exam for Medical Residency (SSM) test). Specifically, we aimed to evaluate the accuracy of ChatGPT responses to questions from the SSM test, and compare its performance to that of medical doctors who had taken the test in 2022.

METHODS

Artificial intelligence

ChatGPT is a language model developed by OpenAI (https://openai.com/blog/chatgpt). It uses advanced self-attention mechanisms and a vast corpus of training data to generate natural language responses in a conversational context. ChatGPT is especially adept at handling complex dependencies over long distances and can produce coherent and contextually appropriate responses. While it has been trained on vast amounts of text data from the internet, it operates in a standalone mode once trained. This means that ChatGPT cannot actively access or browse the internet post-training. All responses it generates come from its internal knowledge, based on the data it was initially trained on, until its last update in 2021. Thus, any insights or information it provides reflects its training data and not from real-time online searches (https://openai.com/blog/chatgpt). Consequently, all responses are generated internally, based on the abstract relationships between input words in the neural network.

Dataset

The Italian State Exam for Medical Residency (SSM) is a comprehensive standardized testing program covering all topics in physicians' fund of knowledge (https://www.universitaly.it/). The difficulty and complexity of questions in the SSM test are highly standardized and regulated, making it an ideal input substrate for AI testing. A total of 140 publicly-available multiple-choice practice questions were obtained from the official website of the SSM released in July 2022, which ensured that all inputs represented accurate out-of-training samples for the GPT-3.5 model. To verify this, a random sample of questions was checked to ensure that none of the answers, explanations, or related content were available on Google before January 1, 2022, representing the last date accessible to the ChatGPT training dataset. Any questions containing visual assets such as clinical images, medical photography, and graphs were removed (questions 13, 83, 84, 85), resulting in 136 items available for encoding. Questions were classified into two categories: clinical case (CC) and notional question (NQ). Two research operators blindly assigned the test questions to one of these categories and a third one resolved discordant assignments. All the researchers involved in this task are licensed physicians. A total of 17 items (12.5% of the dataset) required arbitration. The final dataset consisted of 136 questions, of which 83 NQ questions and 53 CC questions.

Input and output

Questions were formatted into single multiple-choice answers. A new chat session was started in ChatGPT for each entry to reduce memory retention bias. In case of elusive, unclear answers, a single attempt was made to force the AI to answer with one of the options available. The input phrase "Answer with the correct option only" was used to do so. It was coded as incorrect if the answer was still elusive or unclear. Given that the SSM test is written in Italian, all the inputs were submitted in Italian on March 6, 2023.

Statistical analysis

Firstly, AI outputs were dichotomized (1=correct; 0=incorrect). Then, the overall score, clinical case score and notional question score were calculated. Adjusted score on a 140-point scale to compare overall results was calculated with the criteria of the official SSM test, awarding 1 point for each correct answer and -0.25 for the incorrect ones. The overall scores of the medical doctors (MDs) were anonymously retrieved from the official ministerial website. A descriptive analysis of the data was performed. AI score was compared to the mean and median scores of medical doctors who took the same test in 2022. Percentile distribution was calculated to locate the AI-adjusted overall score.

RESULTS

ChatGPT answered 90.44% of the questions correctly. It scored slightly higher on clinical cases compared to notional questions (92.45% *vs* 89.15%). The score adjusted on a 140-point scale was 123.27 following the SSM test criteria of evaluation. A detailed description of the AI answer scores is provided in *Table 1*.

Regarding the distribution of MDs scores, the mean value was 79.42/140, and the median value was 80.75/140 out of a total population of 15,869 participants. The quartile distribution is provided in *Figure 1*.

Analyzing percentile distribution, ChatGPT scored higher than 99.6% of the MDs who took the SSM test in 2022.

DISCUSSION

The present study investigated the feasibility of using ChatGPT as a clinical reasoning and decision-making tool. ChatGPT performance on the SSM test, a standardized assessment for evaluating clinical reasoning skills and medical knowledge in medical doctors entering residency programs, was assessed.

Table 1
Descriptive statistics of ChatGPT answers to the SSM test

	Overall score	NQ score	CC score	Adj score by SSM criteria	Adj score by SSM test criteria on a 140 scale
ChatGPT answers	123/136 (90.44%)	74/83 (89.15%)	49/53 (92.45%)	119.75/136 (88.05%)	123.27/140 (88.05%)

NQ: notional question; CC: clinical cases; SSM: Italian State Exam for Medical Residency.

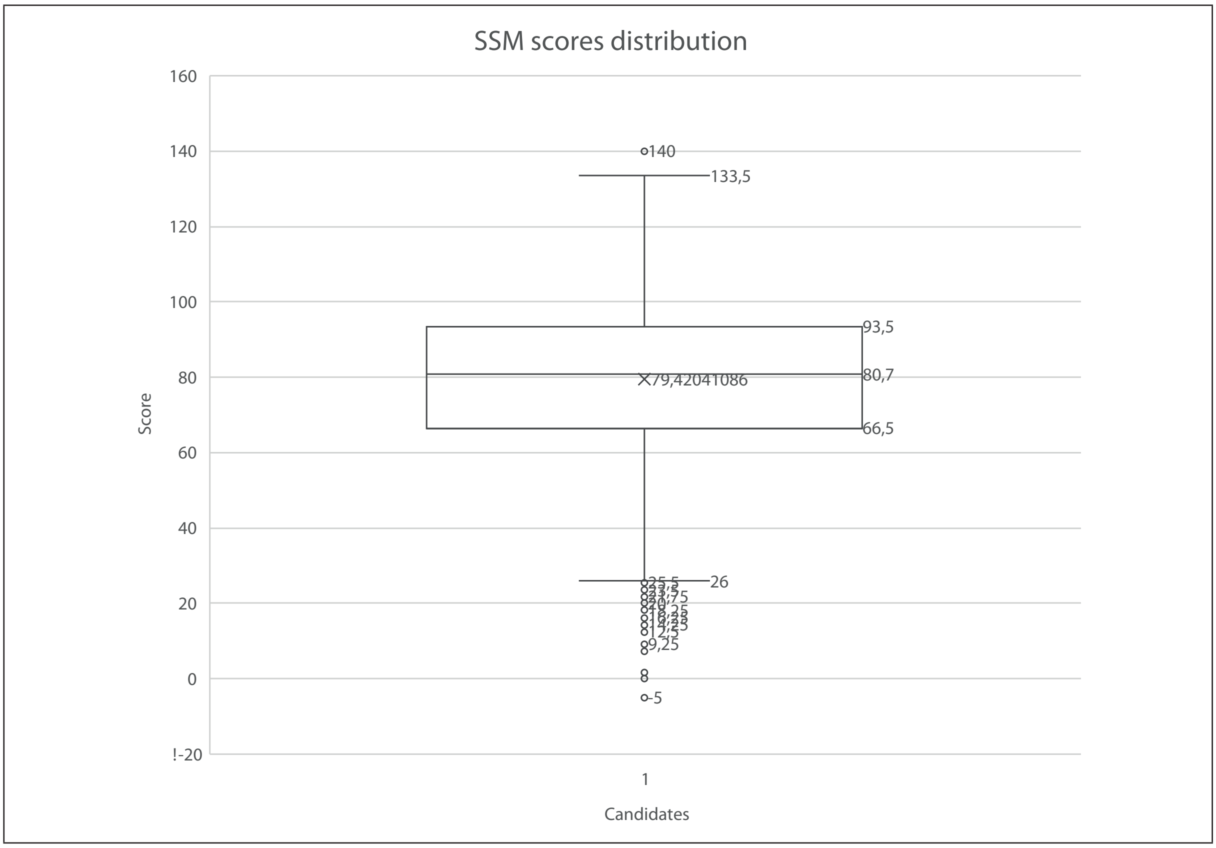

Figure 1
Quartile distribution of SSM test score in 2022. SSM: Italian State Exam for Medical Residency.

Our results demonstrated that ChatGPT achieved an overall score of 90.44% on the SSM test, higher than other studies on different datasets. A study conducted by Gilson *et al*. examined ChatGPT's abilities in the medical field by evaluating its performance on the NBME-Free-Step-1 dataset, which is a part of the USMLE (United States Medical Licensing Examination) [9]. According to their report, ChatGPT exceeded the 60% threshold, which is equivalent to a passing score for a third-year medical student. Additionally, the study highlighted ChatGPT's proficiency in providing logical and informative context for most of its responses. These findings suggest that ChatGPT has the potential to be a valuable medical education tool, capable of enhancing and potentially transforming the learning process. Compared to our findings, the improved performance on the SSM test can be determined by the type of questions, the language in which the test was administered, variations in exam structure, such as the balance between multiple-choice and descriptive questions, and the different region-specific fine-tuning of ChatGPT.

This high level of accuracy is particularly encouraging given the complexity of the SSM test, which requires to integrate knowledge from multiple sources, making diagnostic decisions, and prioritizing patient care. Compared to the doctors' results of the 2022 test, ChatGPT achieved a higher score than 99.6% of the participants.

In a recent study by Bonetti *et al*. [8], ChatGPT was found to correctly answer 122 out of 140 questions on the Italian Residency Admission National Test, positioning it in the top 98.8th percentile among 15,869 medical graduates. Notably, they observed errors in ten questions evaluating direct basic science medical knowledge and in eight questions gauging applied clinical knowledge. Logical errors appeared in two instances, while informational errors were more frequent, noted in 16 instances. These comparative insights from the two studies, both based on the Italian Residency Admission National Test, underscore the remarkable potential of ChatGPT in medical examinations and clinical decision-making.

Given the performance results of ChatGPT on clinical case scenarios, and the emerging evidence of AI applications [10, 11], LLMs have been proven to have the potential to be an effective support for physicians in clinical decisions. However, to become valuable and helpful tools in support of healthcare professionals, LLMs should be tested and validated, and the sources with which ChatGPT constructs the answers must be clear and evident. Moreover, ChatGPT has limited knowledge of the world and events after 2021, with the risk of a lack of information about innovation in clinical practice and diagnosis.

ChatGPT is a useful tool for quickly summarizing the latest medical knowledge, echoing the philosophy of Evidence-Based Medicine but with greater immediacy. While its prowess in quickly synthesizing information is undeniable, over-reliance on these tools may lead to the solidification of certain medical practices, especially in contexts that foster defensive medicine. As underscored by Beaulieu-Jones *et al*., while AI aids in decision-making, the physician's adaptive judgment and experience remain irreplaceable, ensuring nuanced and dynamic clinical decisions [12].

This study has some limitations. Firstly, the sample size is relatively small, as the AI was tested only on the SSM test, including 136 multiple-choice questions. Secondly, the test was written in Italian, and the results may not be generalizable to other languages. Finally, the study only evaluated ChatGPT performance on the SSM test and did not evaluate its potential in every area of medical knowledge.

CONCLUSIONS

This study provides preliminary evidence that ChatGPT has the potential to support clinical decisions, particularly in the context of clinical reasoning and decision-making. The results show that ChatGPT achieved an overall accuracy of 90.44% on the SSM test, which is a promising indication of its ability to handle complex medical concepts and generate contextually appropriate responses. Future studies should focus on understanding the modalities in which LLMs, such as ChatGPT can be implemented in real clinical decision-making scenarios.

Conflict of interest statement

There are no potential conflicts of interest or any financial or personal relationships with other people or organizations that could inappropriately bias conduct and findings of this study.

Received on 28 June 2023.
Accepted on 22 September 2023.

REFERENCES

1. Rigla M, García-Sáez G, Pons B, Hernando ME. Artificial intelligence methodologies and their application to diabetes. J Diabetes Sci Technol. 2018;12(2):303-10. doi: 10.1177/1932296817710475
2. Le Berre C, Sandborn WJ, Aridhi S, Devignes MD, Fournier L, Smaïl-Tabbone M, Danese S, Peyrin-Biroulet L. Application of artificial intelligence to gastroenterology and hepatology. Gastroenterology. 2020;158(1):76-94.e2. doi: 10.1053/j.gastro.2019.08.058
3. Bonderman D. Artificial intelligence in cardiology. Wien Klin Wochenschr. 2017;129(23-24):866-8. doi: 10.1007/s00508-017-1275-y
4. Mamede S, van Gog T, van den Berge K, van Saase JL, Schmidt HG. Why do doctors make mistakes? A study of the role of salient distracting clinical features. Acad Med. 2014;89(1):114-20. doi: 10.1097/ACM.0000000000000077
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 Digit Health. 2023;2(2):e0000198. doi: 10.1371/journal.pdig.0000198
6. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res. 2023;25:e48568. doi: 10.2196/48568
7. Ferdush J, Begum M, Hossain ST. ChatGPT and clinical decision support: Scope, application, and limitations. Ann Biomed Eng. 2023 Jul 29. doi: 10.1007/s10439-023-03329-4
8. Alessandri Bonetti M, Giorgino R, Gallo Afflitto G, De Lorenzi F, Egro FM. How does ChatGPT perform on the Italian Residency Admission National Exam Compared to 15,869 Medical Graduates? Ann Biomed Eng. 2023 Jul 25. doi: 10.1007/s10439-023-03318-7
9. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. 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
10. Tanguay-Sela M, Rollins C, Perez T, Qiang V, Golden G, Tunteng JF, Perlman K, Simard J, Benrimoh D, Margolese HC. A systematic meta-review of patient-level predictors of psychological therapy outcome in major depressive disorder. J Affect Disord. 2022;317:307-18. doi: 10.1016/j.jad.2022.08.041
11. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, Marsh J, DeVylder J, Walter M, Berrouiguet S, Lemey C. Machine learning and natural language processing in mental health: Systematic review. J Med Internet Res. 2021;23(5):e15708. doi: 10.2196/15708
12. Beaulieu-Jones BK. Yuan W, Brat GA, Beam AL, Weber G, Ruffin M, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ Digit Med. 2021;4(1):62. doi: 10.1038/s41746-021-00426-3