How Does ChatGPT Perform on the Italian National Residency Program Admission Test?

Gianmaria Barone *, Filippo Confalonieri, Alessandro Gaeta, Vanessa Ferraro, Paolo Vinciguerra, Alessandra Di Maria

Posted Date: 27 March 2024

doi: 10.20944/preprints202403.1684.v1

Keywords: ChatGPT; Chat Generative Pre-trained Transformer; GPT-3.5; GPT-4; artificial intelligence (AI); chatbot; natural language processing (NLP); medical education; Italian National Residency Program Admission Test; Scuole di Specializzazione in Medicina (SSM)

Article

How Does ChatGPT Perform on the Italian National Residency Program Admission Test?

Gianmaria Barone [1,2,*,†], Filippo Confalonieri [1,2,†], Alessandro Gaeta [3], Vanessa Ferraro [1,2], Paolo Vinciguerra [1,2] and Alessandra Di Maria [1,2]

[1] Department of Ophthalmology, IRCCS Humanitas Research Hospital, Milan, Italy
[2] Department of Biomedical Sciences, Humanitas University, Milan, Italy
[3] Department of Internal Medicine and Medical Specialties (DIMI), Università di Genova, Viale Benedetto XV, 6-16132 Genova, Italy
* Correspondence: gianmaria.barone@humanitas.it
† Equal contribution.

Abstract: Background: Open AI developed ChatGPT, a language model based on the GPT architecture, designed for text-based communication. Trained on diverse internet texts, ChatGPT generates contextually appropriate responses using machine learning. It understands input through analysis and context interpretation, generating coherent and contextually relevant responses. Interaction is possible through messaging platforms. **Materials and Methods**: In November 2023, we utilized ChatGPT 3.5, the default version at that time, to answer questions from the Italian National Residency Program Admission Tests (SSMs) of 2021, 2022, and 2023. The questions cover clinical, diagnostic, analytical, therapeutic, and epidemiological scenarios, sometimes accompanied by images. The study compared ChatGPT's answers to the official corrections on the Italian Ministry of University and Research (MUR) website. The scoring method used for evaluation was 1 point for correct answers, 0 points for unanswered questions, and -0.25 points for incorrect answers, reflecting the SSM test scoring system. **Results**: In summary, ChatGPT was tested with a total of 420 questions, 140 for each test. It achieved an overall accuracy of 80.48%, providing correct answers for 338 questions and incorrect answers for 82 questions. When faced with questions containing both text and images, it answered 55% correctly and 45% incorrectly. The model's performance varied over time, with an 82.14% accuracy rate in 2021 and 2022 (115 correct out of 140) and a 77.14% accuracy rate in 2023 (108 correct out of 140). Applying thisscoring method to the SSM test, ChatGPT would have scored 105 points in 2021 and 2022, and 100 points in 2023. **Conclusions**: ChatGPT has exhibited above-average performance in the last three SSM tests, highlighting its robust capability to interpret clinical scenarios and offer precise diagnostic and therapeutic guidance. Despite this, some limitations persist, notably the software's inability to interpret non-textual information.

Keywords: ChatGPT; chat generative pre-trained transformer; GPT-3.5; GPT-4; artificial intelligence (AI); chatbot; natural language processing (NLP); medical education; Italian National Residency Program Admission Test; Scuole di Specializzazione in Medicina (SSM)

1. Introduction

OpenAI is a non-profit parent corporation of OpenAI Inc. and a for-profit firm called OpenAI LP that works together to conduct artificial intelligence research. The main objective of OpenAI, which was founded in December 2015 is to make sure that the advancement of artificial general intelligence (AGI) benefits all people. AGI describes extremely self-sufficient systems that outperform humans in the majority of economically significant tasks. [1]

Acknowledged for its innovative efforts, OpenAI has been a frontrunner in artificial intelligence research and development, prioritizing both scientific progress and thoughtful analysis of the wider societal effects of AI. Among its noteworthy accomplishments is the development of the GPT

(Generative Pre-trained Transformer) family of language models, of which GPT-3 is one of the biggest and most potent models to date, exhibiting remarkable powers for natural language generation and comprehension.

OpenAI created the language model known as ChatGPT. It is a version built to communicate with people through text chats and is based on the GPT architecture. ChatGPT started in 2019 with the publication of GPT-2, a large-scale language model that attracted notice for its capacity to produce text that was both coherent and contextually appropriate. Initially, OpenAI only published smaller copies of the model and refrained from sharing the whole version due to concerns about potential exploitation of such technology. [2]

ChatGPT is able to produce coherent and appropriately contextualized responses in a variety of scenarios because it has been trained on a wide range of texts from the internet. [3,4] The model uses machine learning to predict and generate text based on the provided context. It can answer questions, provide explanations, write creative texts, and perform a range of other language-related tasks. ChatGPT has been trained on a vast body of literature sourced from the internet, including articles, web pages, books, and more. This training has enabled it to learn linguistic structures, context, and connections between words in different contexts. The GPT model, on which ChatGPT is based, is "pre-trained" through this process, meaning it learns to generate coherent text without being directly programmed on what to say. [5]

When users interact with ChatGPT by providing input, the model uses the context of the input itself to generate a response. Specifically, ChatGPT The assistant responds to questions by examining the provided textual input and generating responses based on the information and language patterns learned during its training. The process [6] involves several stages:

- Input Analysis: Initially, the assistant reads the given input, which can be a question, a request, or a sentence.
- Context Understanding: The information within the input is used to understand the context of the situation. This helps the assistant identify the subject of the conversation and interpret possible nuances.
- Response Generation: Drawing upon linguistic, grammatical, and semantic knowledge learned during training, the assistant generates an appropriate response. If the input consists of specific questions, the aim is to provide clear and relevant answers. If the input requires explanations or details, the assistant generates explanatory text.
- Coherence and Fluency: During response generation, the assistant strives to maintain coherence with the context and produce smooth, well-structured text, drawing inspiration from examples of correct text learned during training.
- Sending the Response: Once the response is generated, the assistant presents it as output to the user. The length of the response can vary based on the complexity of the initial input.
- Iteration: If the user interacts further, the assistant repeats the process, assessing the new input and generating coherent responses accordingly. Interaction can continue in this manner based on user requests.

It should be emphasized that the assistant's responses are based on statistical and probabilistic models, without possessing a conceptual understanding like that of a human being. The quality of responses depends on the accuracy of the provided input and the effectiveness of the language patterns learned during training. [7]

It is possible to interact with ChatGPT through messaging platforms or text-based applications. [8] ChatGPT can be used to manage common questions and provide quick and accurate answers to improve customer experience and reduce operator workload. Additionally, it can be used to enhance virtual personal assistants, giving them a more conversational tone and enhancing their capacity to help users with a variety of activities like information provision, scheduling, and recommendation making. It can facilitate text translation from one language to another in language translation, facilitating more seamless language-to-language communication. ChatGPT can be used to create human-like text in a range of creative writing types and styles, including articles, conversations, and stories. ChatGPT can assist with creating headlines, summaries, product descriptions, social media postings, and website copy, all at the content level.

In the healthcare field, the use of software as an aid in routine clinical practice is now well-established. This includes tasks such as data interpretation, anatomical study, and analysis of instrumental examinations. However, there are still no software solutions capable of solving complex clinical cases by identifying the correct diagnosis and proposing a therapeutic pathway. Nevertheless, some recent publications have tested ChatGPT's performance in answering various medical questions. [9–12] Since no one has ever tested ChatGPT's ability to solve clinical cases in the Italian language, we decided to present the software with clinical questions taken from the Italian National Residency Program Admission Test over the last three years.

The Italian National Residency Program Admission Test, better known in Italian as 'Test di Accesso alle Scuole di Specializzazione in Medicina (SSM Test)', is a selection exam that is required of medical graduates who wish to enter a medical specialization program in Italy. Medical graduate schools offer advanced and specialized training in several medical disciplines, enabling medical graduates to gain in-depth skills in a specific area of medical practice.

This admissions test is designed to select the most suitable and motivated candidates for available positions in graduate schools. Test content may vary by specialty area and university, but typically includes questions on medical, scientific, and clinical topics relevant to the specialty in question. [13]

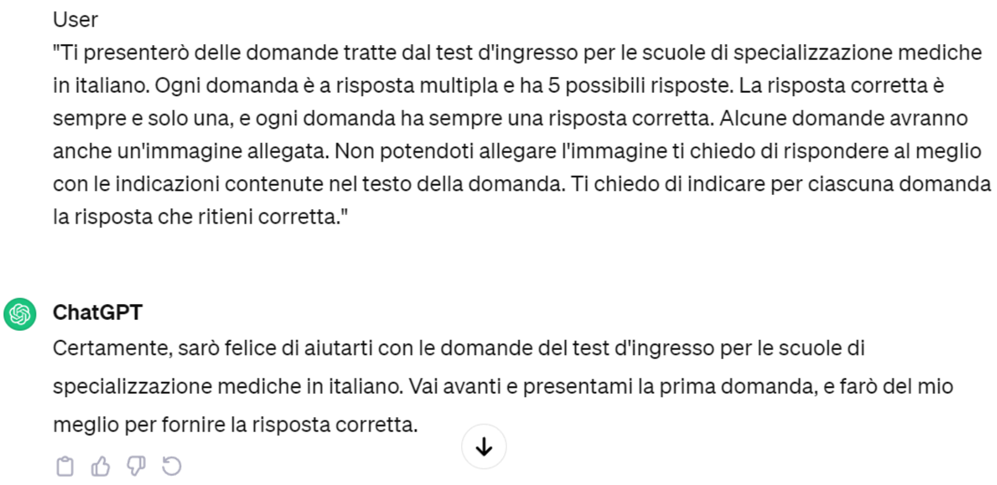
In this paper we examined ChatGPT's capability of answering to SSM's questions, analyzing its score and performance as if it was a candidate.

2. Materials and Methods

We used the ChatGPT 3.5 version, the default version of the software at the time of writing this article (November 2023). The questions presented to ChatGPT were those from the SSM Tests from 2021, 2022, and 2023. The SSM test is held once a year, usually in July, and consists of 140 multiple-choice questions, each with 5 possible answers to be completed within a total of 210 minutes. The questions in the SSM test pertain to the evaluation of clinical, diagnostic, analytical, therapeutic, and epidemiological data in predefined mono- or interdisciplinary scenarios. Some questions are based on attached images, such as electrocardiographic tracings, X-ray images, magnetic resonance imaging (MRI), or computed tomography (CT) scans. The questions are public, and the questions along with the correct answers are available on the website of the Italian Ministry of University and Research (MUR). We used ChatGPT from November 18th to November 20nd, 2023. After logging into the website, we formulated the following instructions in Italian:

> *"I will present you with questions taken from the entrance test for medical specialization schools in Italian. Each question is multiple-choice and has 5 possible answers. The correct answer is always and only one, and each question always has one correct answer. Some questions will also have an attached image. Unable to attach the image, I ask you to respond as best as you can based on the information provided in the question text. Please indicate for each question the answer you believe is correct."* (**Figure 1**)

Subsequently, we inputted all 140 questions from each SSM test and compared the correct answers provided by ChatGPT with the official correction provided by the Italian MUR. After calculating the percentage of correct and incorrect answers, we computed the score that ChatGPT would have achieved had it truly participated in the SSM. In the test, the score is calculated as follows: 1 point for each correct answer, 0 points for each unanswered question, -0.25 points for each incorrect answer.



User

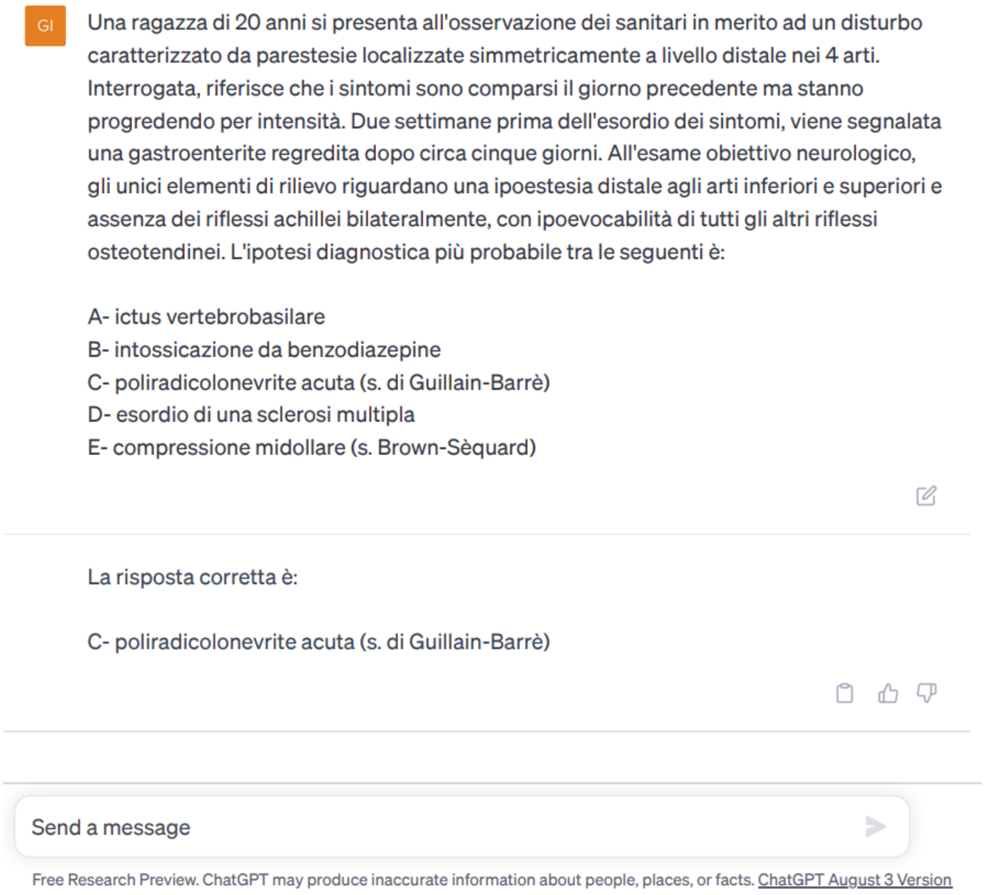
"Ti presenterò delle domande tratte dal test d'ingresso per le scuole di specializzazione mediche in italiano. Ogni domanda è a risposta multipla e ha 5 possibili risposte. La risposta corretta è sempre e solo una, e ogni domanda ha sempre una risposta corretta. Alcune domande avranno anche un'immagine allegata. Non potendoti allegare l'immagine ti chiedo di rispondere al meglio con le indicazioni contenute nel testo della domanda. Ti chiedo di indicare per ciascuna domanda la risposta che ritieni corretta."

ChatGPT

Certamente, sarò felice di aiutarti con le domande del test d'ingresso per le scuole di specializzazione mediche in italiano. Vai avanti e presentami la prima domanda, e farò del mio meglio per fornire la risposta corretta.

Figure 1. Screenshot of the instructions provided to ChatGPT for interpreting the questions.

Standard descriptive statistics were used to calculate the numbers, proportions, and means for each data set.



Una ragazza di 20 anni si presenta all'osservazione dei sanitari in merito ad un disturbo caratterizzato da parestesie localizzate simmetricamente a livello distale nei 4 arti. Interrogata, riferisce che i sintomi sono comparsi il giorno precedente ma stanno progredendo per intensità. Due settimane prima dell'esordio dei sintomi, viene segnalata una gastroenterite regredita dopo circa cinque giorni. All'esame obiettivo neurologico, gli unici elementi di rilievo riguardano una ipoestesia distale agli arti inferiori e superiori e assenza dei riflessi achillei bilateralmente, con ipoevocabilità di tutti gli altri riflessi osteotendinei. L'ipotesi diagnostica più probabile tra le seguenti è:

A- ictus vertebrobasilare
B- intossicazione da benzodiazepine
C- poliradicolonevrite acuta (s. di Guillain-Barrè)
D- esordio di una sclerosi multipla
E- compressione midollare (s. Brown-Sèquard)

La risposta corretta è:

C- poliradicolonevrite acuta (s. di Guillain-Barrè)

Send a message
Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT August 3 Version

Figure 2. Example of a question posed to ChatGPT and response with the correct solution.

3. Results

ChatGPT was queried a total of 420 questions (140 for each test). It found the correct answer for 338 questions, resulting in a total of 80,48% correct responses. It provided incorrect answers for 82 questions, accounting for 19,52% incorrect responses. Out of the 20 questions that included both text and an accompanying image for interpretation, the software correctly answered 11 questions (55%), while it provided an incorrect answer for 9 questions (45%). Specifically, the correct answers were

115 out of 140 in 2021 and 2022 (82.14% correct answers), and 108 in 2023 (77.14% correct answers). Considering the scoring methodology, if ChatGPT had participated in the SSM test, it would have scored a total of 105 points in 2021 and 2022, and 100 points in 2023.

In detail, the subjects to which ChatGPT was able to answer all questions correctly (100% correct answers overall) were Dermatology, Rheumatology, Immunology, Clinical Biochemistry, Pharmacology, Neurology, Psychiatry, Ophthalmology, Otolaryngology, Anesthesia, Statistics, Neurosurgery, Geriatrics, and Senology. On the contrary, the subjects where ChatGPT made the most mistakes were Radiology (66.6% correct answers), Nephrology (57.1% correct answers), Pathological Anatomy (50% correct answers), and Genetics (50% correct answers).

Table 1 summarizes the percentages of correct answers overall and for each individual subject for each of the 3 tests and the overall sum combining the 3 tests.

Table 1. Percentage of correct answers overall and for each individual subject for each of the 3 tests and the overall sum combining the 3 tests. Where there were no questions related to the individual subject in a test, 'no' has been marked in the corresponding cell.

Subject	% correct answer 2021	% correct answer 2022	% correct answer 2023	% correct answer (total)
Total	84.12%	84.12%	77.14%	80,48%
Internal Medicine	86.6%	80%	93.3%	85.5%
Cardiology	73.7%	87.5%	76.2%	78.6%
Gastroenterology	66.6%	77.7%	90.9%	81.5%
Pneumology	100%	77.7%	72.7%	81.5%
Infectious Diseases	80%	80%	100%	84.6%
Nephrology	83.3%	100%	28.6%	57.1%
Dermatology	100%	100%	100%	100%
Orthopedics	87.5%	100%	100%	93.3%
Rheumatology	100%	no	no	100%
Oncology	81.8%	83.3%	92.8	86.5%
Endocrinology	83.3%	100%	80%	85.7%
Immunology	100%	100%	no	100%
Clinical Biochemistry	100%	no	no	100%
Pharmacology	100%	100%	no	100%
Neurology	71.4%	100%	100%	100%
Psichiatry	100%	100%	100%	100%
Emathology	0%	100%	80%	75%
Pediatrics	75%	100%	77.7%	73.3%
Emergency medicine	75%	no	no	75%
Gynecology	80%	90%	80%	83%
Ophtalmology	100%	100%	100%	100%
Otolaringology	100%	no	100%	100%
Urology	100%	100%	no	83.3%
Pathological Anatomy	50%	no	no	50%
Anesthesia	100%	no	no	100%
Forensic medicine	50%	100%	100%	85.7%
Statistics	no	100%	100%	100%
Genetics	no	100%	0%	50%
Radiology	no	66.6%	66.6%	66.6%

Neurosurgery	no	100%	no	100%
Geriatrics	no	100%	no	100%
Senology	no	no	100%	100%

4. Discussion

The creation of computer systems that are capable of carrying out activities that normally require human intelligence is known as artificial intelligence, or AI. Learning, thinking, problem-solving, perception, comprehending spoken language, and even interacting with the surroundings are some of these skills. The goal of AI is to build devices or software that can mimic human intelligence, allowing them to learn from experience and adapt and perform better.

AI comes in two primary varieties. Weak AI, also known as narrow AI, is made to accomplish a limited number of tasks or a single task. It performs exceptionally well in the predetermined set of operations. Recommendation algorithms, image and audio recognition software, and virtual personal assistants such as Siri are some examples. General AI, also known as strong AI, is a more sophisticated type of AI that is comparable to human intellect in that it can comprehend, learn, and apply knowledge to a variety of activities. As of right now, general artificial intelligence is still only a theoretical concept.

AI comprises several subfields, including robotics, computer vision, natural language processing, and machine learning. Specifically, machine learning is a branch of that focuses on creating statistical models and algorithms that let computers carry out tasks without the need for explicit programming. The model is trained using data, which enables it to perform better over time.

Applications of AI are numerous and span many industries, such as healthcare, banking, education, transportation, and more. Its capabilities and social influence are growing at a quick pace due to continuous research and advances.

ChatGPT is clearly an innovative and potentially revolutionary element in every field of human knowledge. This is due to a significant advancement in natural language processing and artificial intelligence. Its advanced features allow for natural and conversational text interaction, paving the way for a range of new possibilities, even in the medical field. [14,15]

Many authors have raised ethical questions about ChatGPT's actual ability to replace humans in the scientific field, with the almost unanimous conclusion that AI is a tool with enormous potential but cannot replace the human brain. [16] Many discussions have arisen regarding the possibility of using ChatGPT for the writing of scientific articles. Salvagno et al. [17] have recently analyzed ChatGPT’s ability to write scientific articles instead of humans, concluding that there is a consensus that regulations on the use of chatbots in scientific papers will soon be required. Thorp's opionion appears much more definitive, as he has explicitly written “*ChatGPT is fun, but not an author*”. [16] Many other authors, on the other hand, have focused on ChatGPT's ability to address clinical scenarios. Alkaissi et al. [18] subjected ChatGPT to two cases of rare metabolic disorders, observing a good capacity for case analysis but various limitations, particularly errors in source retrieval, inadequacy of sources, and different 'hallucinations'. ‘Hallucinations’ are one of the most well-known issues of AI. These are defined as "phenomenon of a machine, such as a chatbot, generating seemingly realistic sensory experiences that do not correspond to any real-world input" [18]. The analysis of this phenomenon is of fundamental importance and highlights how to-day's AI is capable of organizing already known information, with errors that can stem from the algorithm itself, but without being able to create new knowledge that is not already present in the pool of resources available to the software. [19]

One of the key features touted by the advancement of ChatGPT is its ability to understand context and maintain a coherent and relevant conversation on the topic at hand. In this article, we have demonstrated that this extends to the medical field by evaluating ChatGPT on a series of 320 questions related to various areas of medicine, structuring the conversation with the software using questions and answers.

We found that the model is capable of correctly answering over 77% of the questions from the SSM 2023, achieving an overall score of 100 points. Considering that the average score among the

14,036 participants in the SSM 2023 was 83.09 points, with a median of 84.5 points, ChatGPT has demonstrated a clearly above-average response capability. If ChatGPT had participated in the test, it would have been ranked between positions 2858 and 2930 out of 14,036 participants. Similarly, for the tests of previous years, ChatGPT has achieved a score above average, even obtaining 115 points, potentially placing it in the top 1000 positions among the 19,449 participants in the 2021 test and the 15,873 participants in the 2022 test.

Our results also emphasize that in many cases the incorrect answers can be attributed to ChatGPT's inability to analyze the images attached to the questions. Among the questions with an attached image, ChatGPT was able to answer correctly in just over half of the cases. Specifically, the correct answer was found when the text contained sufficient information to understand the question, while the software was unable to provide an answer when asked to simply analyze an image. Furthermore, our findings demonstrate that even in the case of incorrect answers, the responses provided by the model always contained a logical explanation for the answer selection, and more than 90% of the time, this response directly included information from the context of the question.

As indicated by the results of this study and several previous studies, language models based on machine learning, such as ChatGPT, have made significant progress [20,21]. Particularly, the above-average rate of correct responses among test candidates highlights how linguistic models are capable of competing with humans in highly specialized fields. However, it's also important to consider that ChatGPT relies on a multitude of texts and references at its disposal, yet it may not always rigorously assess the reliability of the sources it draws from. [17] Therefore, it's not always clear how AI would perform in situations where it draws from two sources that provide different or opposing information on the same topic. Additionally, Thirunavukarasu et al. [22] recently observed that the use of ChatGPT serves as a useful support for decision-making in clinical practice, but it cannot replace the central role of the human figure in the diagnostic and therapeutic management of the patient.

It is crucial to note that this is the first study to evaluate ChatGPT's capability of answering to a national board examination entirely provided in a language other than English and, specifically, in Italian.

The main limitations of this study are related to the deliberate use of the default version of ChatGPT (3.5) rather than version 4.0, which, at the time of paper writing, is a paid version. This choice was made to utilize the most accessible version of the software currently available. However, it is known that the premium version (ChatGPT 4.0) provides slightly better performance. [23] Therefore, it is reasonable to hypothesize that the future spread of more updated versions of the software will allow for even greater reliability. Furthermore, some answers were affected by ChatGPT's inability to recognize attached images. Improving the integration of non-textual attachments in the chat will lead to a better experience and even more relevant assistance from ChatGPT for doctors during their clinical practice.

5. Conclusions

ChatGPT has demonstrated the ability to achieve a score above average in the 2021, 2022 and 2023 SSM test, showcasing a strong capacity to interpret clinical scenarios and provide accurate diagnostic and therapeutic guidance. However, limitations still exist, particularly concerning the software's limited ability to interpret non-textual information. Nonetheless, it's evident that this tool must be strongly considered, as it will become a valuable support in common clinical practice in the coming years.

Author Contributions: Conceptualization, G.B., A.G., F.C., V.F., A.D.M.; methodology, F.C., V.F., A.D.M., G.B, A.G.; software, A.G. and G.B.; validation, G.B., F.C., V.F., A.D.M. and G.A; formal analysis, G.B., A.G., F.C., V.F., A.D.M.; investigation, F.C., V.F., A.D.M., G.B., and G.A.; resources, F.C., A.D.M. and G.B. ; data curation, F.C., V.F., A.D.M., G.B. and G.A.; writing—original draft preparation, G.B.; writing—review and editing, all authors., supervision, A.D.M., G.B. and F.C.; project administration, A.D.M., G.B and F.C.; funding acquisition, F.C., V.F., A.D.M., G.B. and G.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available on reasonable request by the corresponding authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mintz Y, Brodie R. Introduction to Artificial Intelligence in Medicine. Minim Invasive Ther Allied Technol. 2019 Apr;28(2):73-81. Doi: 10.1080/13645706.2019.1575882. Epub 2019 Feb 27. PMID: 30810430.
2. Hamet P, Tremblay J. Artificial Intelligence in Medicine. Metabolism. 2017 Apr;69S:S36-S40. Doi: 10.1016/j.Metabol.2017.01.011. Epub 2017 Jan 11. PMID: 28126242.
3. Soman S, Ranjani HG. Observations on LLMs for Telecom Domain: Capabilities and Limitations. arXiv. Doi: 10.48550/arXiv.2305.13102.
4. Ramesh AN, Kambhampati C, Monson JR, Drew PJ. Artificial Intelligence in Medicine. Ann R Coll Surg Engl. 2004 Sep;86(5):334-8. Doi: 10.1308/147870804290. PMID: 15333167; PMCID: PMC1964229.
5. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023 Mar 19;11(6):887. Doi: 10.3390/Healthcare11060887. PMID: 36981544; PMCID: PMC10048148.
6. Eysenbach G. The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. JMIR Med Educ. 2023 Mar 6;9:E46885. Doi: 10.2196/46885. PMID: 36863937; PMCID: PMC10028514.
7. Welsby P, Cheung BMY. ChatGPT. Postgrad Med J. 2023 Sep 21;99(1176):1047-1048. Doi: 10.1093/Postmj/Qgad056. PMID: 37462242.
8. Gordijn B, Have HT. ChatGPT: Evolution or Revolution? Med Health Care Philos. 2023 Mar;26(1):1-2. Doi: 10.1007/S11019-023-10136-0. PMID: 36656495.
9. 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 Feb 9;2(2):E0000198. Doi: 10.1371/Journal.Pdig.0000198. PMID: 36812645; PMCID: PMC9931230.
10. Friederichs H, Friederichs WJ, März M. ChatGPT in Medical School: How Successful Is AI in Progress Testing? Med Educ Online. 2023 Dec;28(1):2220920. Doi: 10.1080/10872981.2023.2220920. PMID: 37307503; PMCID: PMC10262795.
11. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions. Cureus. 2023 Jun 22;15(6):E40822. Doi: 10.7759/Cureus.40822. PMID: 37485215; PMCID: PMC10362981.
12. Grewal H, Dhillon G, Monga V, Sharma P, Buddhavarapu VS, Sidhu G, Kashyap R. Radiology Gets Chatty: The ChatGPT Saga Unfolds. Cureus. 2023 Jun 8;15(6):E40135. Doi: 10.7759/Cureus.40135. PMID: 37425598; PMCID: PMC10329466.
13. Decreto Direttoriale n. 645 Del 15-05-2023, “Bando Di Ammissione Dei Medici Alle Scuole Di Specializzazione Di Area Sanitaria per l’a.a. 2022/2023.”, Ministero Dell’Università e Della Ricerca (MUR) Della Repubblica Italiana.
14. Dave T, Athaluri SA, Singh S. ChatGPT in Medicine: An Overview of Its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations. Front Artif Intell. 2023 May 4;6:1169595. Doi: 10.3389/Frai.2023.1169595. PMID: 37215063; PMCID: PMC10192861.
15. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst. 2023 Mar 4;47(1):33. Doi: 10.1007/S10916-023-01925-4. PMID: 36869927; PMCID: PMC9985086.
16. Thorp HH. ChatGPT Is Fun, but Not an Author. Science. 2023 Jan 27;379(6630):313. Doi: 10.1126/Science.Adg7879. Epub 2023 Jan 26. PMID: 36701446.
17. Salvagno M, Taccone FS, Gerli AG. Can Artificial Intelligence Help for Scientific Writing? Crit Care. 2023 Feb 25;27(1):75. Doi: 10.1186/S13054-023-04380-2. Erratum in: Crit Care. 2023 Mar 8;27(1):99. PMID: 36841840; PMCID: PMC9960412.
18. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus. 2023 Feb 19;15(2):E35179. Doi: 10.7759/Cureus.35179. PMID: 36811129; PMCID: PMC9939079.
19. Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. ACM Comput. Surv. 55, 12, Article 248 (December 2023), 38 Pages. Https://Doi.Org/10.1145/3571730.
20. Harsha N, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. arXiv. Doi: 10.48550/arXiv.2303.13375. Preprint Posted Online on March 20, 2023.

21. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023 Mar 30;388(13):1233-1239. Doi: 10.1056/NEJMsr2214184. PMID: 36988602.
22. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, Shah S. Trialling a Large Language Model (ChatGPT) in General Practice With the Applied Knowledge Test: Observational Study Demonstrating Opportunities and Limitations in Primary Care. JMIR Med Educ. 2023 Apr 21;9:E46599. Doi: 10.2196/46599. PMID: 37083633; PMCID: PMC10163403.
23. 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 29;9:E48002. Doi: 10.2196/48002. PMID: 37384388; PMCID: PMC10365615.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.