ChatGPT Performs on the Chinese National Medical Licensing Examination

Xinyi Wang
Zhenye Gong
Guoxin Wang
Jingdan Jia
Ying Xu
Jialu Zhao
Qingye Fan
Shaun Wu
Weiguo Hu
Xiaoyang Li (woodslee429@126.com)

Research Article

Keywords: ChatGPT, Chinese National Medical Licensing Examination, medical student

Posted Date: February 16th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2584079/v1

Abstract

INTRODUCTION:

ChatGPT, a language model developed by OpenAI, uses a 175 billion parameter Transformer architecture for natural language processing tasks. This study aimed to compare the knowledge and interpretation ability of ChatGPT with those of medical students in China by administering the Chinese National Medical Licensing Examination (NMLE) to both ChatGPT and medical students.

METHODS

We evaluated the performance of ChatGPT in two years' worth of the NMLE, which consists of four units. At the same time, the exam results were compared to those of medical students who had studied for five years at medical colleges.

RESULTS

ChatGPT’s performance was lower than that of the medical students, and ChatGPT’s correct answer rate was related to the year in which the exam questions were released.

CONCLUSION

ChatGPT’s knowledge and interpretation ability for the NMLE were not yet comparable to those of medical students in China. It is probable that these abilities will improve through deep learning.

Introduction

ChatGPT is a language model developed by OpenAI. It is based on the transformer architecture and uses deep learning algorithms to generate text. [1] The model has been trained on a large corpus of text data, allowing it to generate coherent and contextually appropriate responses to natural language inputs. [2] As a response generation model for conversational AI applications, ChatGPT has been fine-tuned and improved. It is considered to be the most expressive text generation model when it comes to writing, achieving state-of-the-art effectiveness in natural language processing tasks. [3] ChatGPT has seen significant improvements in its interaction with humans, from chatbots and customer service to content creation.[4] Its text is derived from a vast language corpus network and has demonstrated state-of-the-art performance in a wide range of natural language tasks. ChatGPT holds great potential as an assistant for professional work. [5]

Large language models (LLMs) have been explored in the medical field as a means of providing personalized patient interaction and educating consumers about their health. [6] Workers in the medical-

related fields are starting to use AI such as ChatGPT in various ways to improve efficiency in various aspects of the healthcare system, such as diagnosis, medical imaging analysis, predictive models, and personalized medicine. ChatGPT has the potential to enhance this application by providing a combination of medical knowledge and conversational interaction. [7]

ChatGPT performs well in Certified Public Accountant (CPA) exams [8]and Uniform Bar Examination (UBE) [9], demonstrating its ability to generate accurate text responses to complex input standards. ChatGPT has been applied in the United States medical licensing exams and has demonstrated a significant advancement in natural language processing, specifically in its ability to answer medical questions. Additionally, the model's demonstration of logical and contextual information in the majority of answers highlights its potential as a valuable tool for medical education and learning support. [10] In South Korea, ChatGPT has also shown remarkable performance in parasitology exams, despite some differences with the scores of medical students. [11] This suggests that ChatGPT holds promise as a tool in medical education and assessment in the future.

Passing the National Medical Licensing Examination (NMLE) is a necessary requirement in China to become a doctor. The NMLE is a legally mandated qualification examination for individuals seeking to practice medicine in the country. [12] The examination is comprised of four units. The first unit focuses on assessing students' medical knowledge, medical policies and regulations, and preventive medicine. The second and third units primarily assess clinical knowledge primarily in internal and surgical medicine. The fourth unit primarily assesses knowledge in the female reproductive system, pediatric diseases, and the mental and nervous system. Those who successfully pass the NMLE will receive the designation of medical practitioner and can then register with the local health department at the county level or above.

Medical students will be able to evaluate the accuracy of medical information generated by AI and have the abilities to create reliable, validated information for patients and the public. Therefore, it is necessary to determine how accurately ChatGPT, a recently developed AI chatbot, can solve questions on medical examinations. [1] But, there have been no reported studies in the literature databases, including PubMed, Scopus, and Web of Science, on the comparability of ChatGPT’s performance to that of students on Chinese medical examinations. This study aims to evaluate the performance of resident physicians and ChatGPT in the Chinese NMLE and provide insights into the training and education of medical professionals in China.

Method

Medical Examination Data Sets

We utilized the original questions from the Chinese National Medical Licensing Examination in 2021 and 2022. Each set of questions consists of 4 units, with 150 questions per unit. Based on the requirements to pass the NMLE, a total score of 360 or above is considered qualified.

We analyzed the performance of the resident physicians who underwent standardized training at Ruijin Hospital during the years 2021 and 2022 in the NMLE. All participants received 5 years of undergraduate medical education from 18 medical schools across the country. Additionally, we utilized ChatGPT to answer the questions from the China Medical Licensing Examination in 2021 and 2022 in order to compare its performance with that of Chinese medical students.

In this study, a total of 49 medical students participated in the examination in 2021, and 65 students took part in the examination in 2022. There were no exclusion criteria. There was no bias in the selection of examinees. All students in Ruijin Hospital who attended the NMLE were included.

This is a descriptive study to compare the ability of ChatGPT to answer questions with that of medical students. Sample size estimation was not required because all target students were included, and one AI platform was added. This was not a study of human subjects, but an analysis of the results of an educational examination routinely conducted. Therefore, neither receiving approval from the institutional review board nor obtaining informed consent was required.

Prompt Engineering

To improve the accuracy of the generative language model (ChatGPT) output, we standardized the input formats of the NMLE data sets by following these steps:

1. In order to enhance ChatGPT's understanding and eliminate language barriers, we input both English and Chinese versions of the questions in a comparative manner.
2. We revised the format of the questions by placing the question text at the top, followed by the direct question on the next line, in order to facilitate better understanding and avoid duplicates.

Data Analysis

All statistical analyses were performed using SPSS 22.0 statistics software (SPSS Inc., Chicago, IL, USA). Unpaired chi-square tests were applied to determine if the examination year significantly impacts ChatGPT's performance in the exam.

Result

According to data, ChatGPT correctly answered 275 out of 600 items (45.8%) in 2021 NLME. (Fig. 1A) This score was lower than the average score of 49 medical students, which was 407.5 out of 600 (67.9%), with a minimum score of 298 (49.7%) and a maximum score of 498 (83.0%). In 2022 NLME, ChatGPT correctly answered 219 out of 600 items (36.5%). (Fig. 1B) Meanwhile, the average score of 65 medical students was 412.7 (68.7%), with a minimum score of 295 (49.2%) and a maximum score of 474 (79.0%).

The result shows ChatGPT’s responses according to examination year. The chi-square test yielded results of $\chi^2=10.79$, degrees of freedom (df) = 1. This result indicates that ChatGPT performed in the 2021 exams significantly better than in the 2022 exams (P = 0.001).

Table1 shows ChatGPT's responses according to different units. Based on a two-year assessment, ChatGPT demonstrated higher performance in Unit 4 (47.7%), whereas its results in the other three units were relatively lower (Unit 1 43.4%, Unit 2 36.7%, Unit 3 36.7%), indicating a statistical difference (P=0.012).

Table 1
The scores of ChatGPT in 4 units.

	Unit1	Unit2	Unit3	Unit4
2021	66	54	70	85
2022	65	56	40	58
Total Score	132	110	110	143
Accuracy	43.4%,	36.7%	36.7%	47.7%

Discussion

The development of ChatGPT has been a major milestone in the field of NLP (Natural Language Processing) and AI. [2] Its continuous improvement and evolution is likely to have a major impact on the future of conversational AI. The consequences are sure to affect all sectors of society, from business development to medicine, from education to research, from coding to entertainment and art. [13]

ChatGPT, as a cutting-edge and massive language model, is capable of learning from vast amounts of text data to generate human-like language. [2] In the education sector, the utilization of ChatGPT offers personalized learning materials and the ability to answer questions related to medical student exams. [10] Through utilizing its advanced natural language processing capabilities, ChatGPT effectively enhances the overall learning experience for students, making it more efficient and engaging. Its ability to process and comprehend natural language inputs combined with its vast knowledge base makes it a valuable tool for medical educators to enhance the learning process and support student success.

With its unique advantages, ChatGPT can be used in medical education for various purposes to enhance the quality of education. ChatGPT can effectively be used for evaluating students' essays and papers, analyzing sentence structure, vocabulary, grammar, and clarity of the paper. [14] Another use of ChatGPT is its ability to generate exercises, quizzes, and scenarios which can be used in the classroom to aid in practice and assessment. Additionally, ChatGPT is capable of writing basic medical reports, which assists students in identifying areas for improvement and deepening their understanding of complex medical concepts. [15] Its ability to generate translations, explanations, and summaries can also be used to help students understand complex learning material more easily. [16] ChatGPT can be used to provide accurate and up-to-date information on medical topics upon immediate notification. This could include a range of topics from diseases and their treatments to medical procedures. [17] For these applications to

be effective, ChatGPT must perform similarly to human experts in medical knowledge and reasoning tests so that users have confidence in its responses.

In both 2021 and 2022, ChatGPT's scores in the Chinese National Medical Licensing Examination did not meet the passing requirements. According to statistics, the national pass rate of the exam in 2021 was 50%, and in 2022 was 55%. Compared to medical students who have undergone traditional 5-year medical education in a medical school, ChatGPT's performance is currently not sufficient. This may be due to several reasons. Firstly, all questions in the Chinese NMLE are multiple-choice questions that require selection of the best answer, while some questions have multiple-choice answers provided by ChatGPT, which are considered as suboptimal rather than incorrect in clinical practice. Secondly, there are differences in medical policies and laws between China and the United States, such as issues related to abortion, which is not allowed in the United States by law, while it is allowed in China under certain medical conditions. Thirdly, some unique epidemiological data in China are beyond the knowledge scope of ChatGPT, and some data are only available in Chinese. This phenomenon is similar to the situation mentioned in a paper published by Korean colleagues. In addition, some question types in the Chinese NMLE are based on a patient-centered clinical scenario, followed by 2 to 3 related questions, each of which is related to the initial clinical scenario but tests different points, and the questions are independent of each other. ChatGPT performed poorly on these questions.

However, we have also observed several interesting phenomena. ChatGPT performed relatively well in Unit 4, which covers subjects such as pediatrics, gynecology, and surgery and is less affected by national conditions. ChatGPT's performance in the 2021 exam was higher than that in the 2020 exam, which may be due to more people seeking relevant information online to prepare for the exam, allowing ChatGPT to learn more knowledge through big data. We tried 10 questions that ChatGPT answered incorrectly and after being told the correct answers, ChatGPT was able to provide a correct answer in response (data not shown).

The above results should not be extrapolated to other subjects or medical schools, as chatbots are likely to continue to rapidly evolve through user feedback. Future trials with the same items may yield different results. The present results reflect the abilities of ChatGPT on February 1, 2023. The input for the question items for ChatGPT was not exactly the same as for medical students. The chatbot cannot receive information about differences in medical policies between the US and other regions, and this information needs to be learned by the software. Additionally, the interpretation of explanations and correct answers may vary depending on the perspectives of clinical experts, although the author has been working in the field of medicine in China for 15 years (2009–2023). Patient care best practices may also vary depending on the region and medical environment.

At present, ChatGPT's level of understanding and ability to interpret information is not adequate to be utilized by medical students, particularly in medical school exams and high-stakes exams such as health licensing exams. However, it is anticipated that with deep learning, ChatGPT's knowledge and interpretation abilities will improve at a rapid pace, similar to AlphaGo's performance. Thus, medical and

health professors and students should be mindful of incorporating this AI platform into medical and health education in the near future. In addition, the integration of AI into the medical school curriculum is already underway in some institutions.

In conclusion, ChatGPT's proficiency and interpretation abilities for questions pertaining to the Chinese NMLE are not yet at par with Chinese medical students. Nevertheless, it is probable that these abilities will improve through deep learning. Medical education authorities and students ought to be cognizant of the developments in this AI chatbot and ponder its potential utilization in learning and education. In conclusion, this research indicates that ChatGPT holds the possibility of serving as a virtual medical mentor, but further examination is required to fully evaluate its efficiency and applicability in this context.

Declarations

Ethical Approval

This was not a study of human subjects, but an analysis of the results of an educational examination routinely conducted. Therefore, neither receiving approval from the institutional review board nor obtaining informed consent was required.

Competing interests

The authors declare that there is no conflflict of interest regarding the publication of this paper.

Authors' contributions

WXY and LXY conceived and designed the study, developed the study protocol, statistical analyses and wrote the manuscript. WGX, JJD, XY, ZJL, FQY, and SW encoded and input the data into ChatGPT. GZY and HWG performed quality control, and statistical analyses.

Funding

The work received no external funding.

References

1. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, Moy L. ChatGPT and Other Large Language Models Are Double-edged Swords.Radiology. 2023 Jan 26:230163. https://doi.org/10.1148/radiol.230163.
2. Som Biswas. ChatGPT and the Future of Medical Writing .Radiology.Feb 2 2023 :223312 https://doi.org/10.1148/radiol.223312
3. Shuai Wang, Harrisen Scells, Bevan Koopman, Guido Zuccon. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? arXiv. Preprint posted online on 3 Feb 2023 https://doi.org/10.48550/arXiv.2302.03495

4. Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, Yupeng Wu. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv. Preprint posted online on 18 Jan 2023 https://doi.org/10.48550/arXiv.2301.07597
5. King, M.R. The Future of AI in Medicine: A Perspective from a Chatbot. Ann Biomed Eng 51, 291–295 (2023). https://doi.org/10.1007/s10439-022-03121-w
6. Avisha Das, Salih Selek, Alia R. Warner, Xu Zuo, Yan Hu, Vipina Kuttichi Keloth, Jianfu Li, W. Jim Zheng, and Hua Xu. 2022. Conversational Bots for Psychotherapy: A Study of Generative Transformer Models Using Domain-specific Dialogues. In Proceedings of the 21st Workshop on Biomedical Language Processing, pages 285–297, Dublin, Ireland. Association for Computational Linguistics.https://doi.org/10.18653/v1/2022.bionlp-1.27
7. Mijwil, M., Mohammad Aljanabi, & Ahmed Hussein Ali. (2023). ChatGPT: Exploring the Role of Cybersecurity in the Protection of Medical Information . Mesopotamian Journal of CyberSecurity, 2023, 18–21. https://doi.org/10.58496/MJCS/2023/004
8. Bommarito, J., Bommarito, M., Katz, D. M. & Katz, J. GPT as Knowledge Worker: A Zero-Shot Evaluation of (AI)CPA Capabilities.arXiv preprint posted online on 11 Jan 2023 https://doi.org/10.48550/arXiv.2301.04408
9. Bommarito II, M. & Katz, D. M. GPT Takes the Bar Exam.arXiv preprint posted online on 29 Dec 2022 https://doi.org/10.48550/arXiv.2212.14402
10. Aidan Gilson, Conrad W Safranek, Thomas Huang, 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. JMIR Med Educ. 2023 Feb 8;9:e45312. https://doi.org/10.2196/45312
11. Sun Huh. Are ChatGPT’s knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof. 2023;20:1. https://doi.org/10.3352/jeehp.2023.20.1
12. Xiancheng Wang. Experiences, challenges, and prospects of National Medical Licensing Examination in China. BMC Med Educ. 2022 May 8;22(1):349. https://doi.org/10.1186/s12909-022-03385-9
13. Philipp Hacker, Andreas Engel, Marco Mauer.Regulating ChatGPT and other Large Generative AI Models. arXiv. Preprint posted online on 10 Feb 2023 https://doi.org/10.48550/arXiv.2302.02337
14. Kung TH, Cheatham M, Medinilla A, Sillos C, De Leon L, Elepano C, et al. Performance of ChatGPT on USMLE: Potential for AIAssisted Medical Education Using Large Language Models. medRxiv 2022.12.19.22283643 https://doi.org/10.1101/2022.12.19.22283643
15. Katharina Jeblick BS, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, Michael Ingrisch. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. arXiv preprint posted online on 30 Dec 2022 https://doi.org/10.48550/arXiv.2212.14882

16. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers.bioRxiv 2022.12.23.521610 https://doi.org/10.1101/2022.12.23.521610
17. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. arXiv preprint posted online on30 Dec 2022 https://doi.org/10.48550/arXiv.2212.14882

Figures

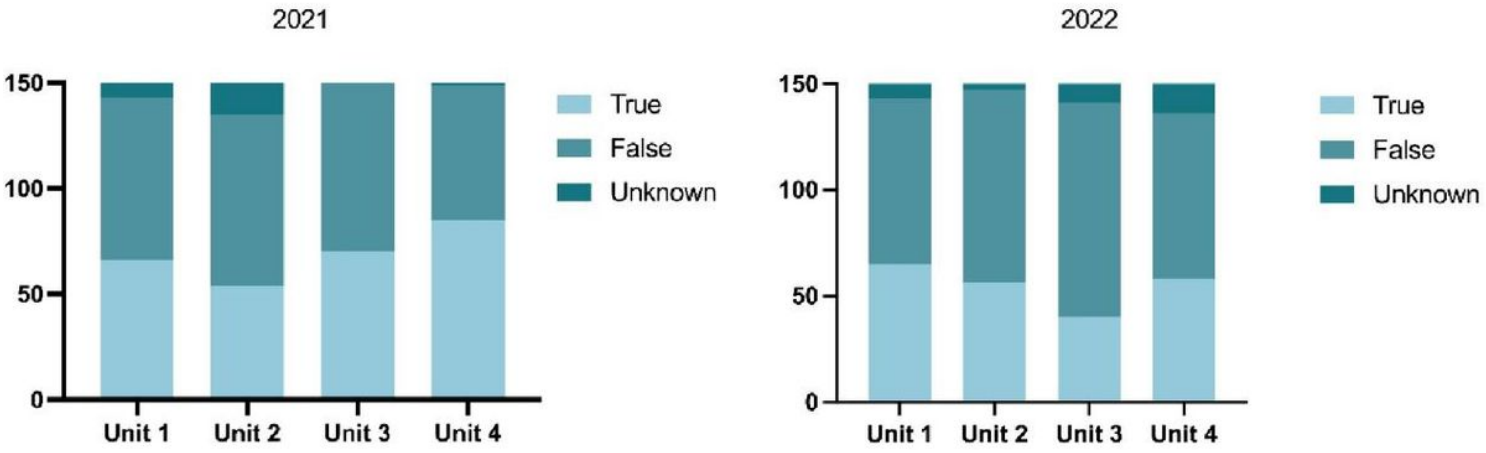

Figure 1

The performance of ChatGPT in 4 units (2021&2022)