PLOS DIGITAL HEALTH

RESEARCH ARTICLE

How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language

Changchang Fang[1,2‡], Yuting Wu[1‡], Wanying Fu[1,2], Jitao Ling[1], Yue Wang[3,4], Xiaolin Liu[3,4], Yuan Jiang[3,4], Yifan Wu[1], Yixuan Chen[1], Jing Zhou[1], Zhichen Zhu[1], Zhiwei Yan[5], Peng Yu[1,6]*, Xiao Liu[3,4,6]*

1 Department of Endocrine, the Second Affiliated Hospital of Nanchang University, Jiangxi, China, **2** Queen Mary College, Nanchang University, Jiangxi, China, **3** Department of Cardiology, Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Guangzhou, China, **4** Guangdong Province Key Laboratory of Arrhythmia and Electrophysiology, Guangzhou, China, **5** Provincial University Key Laboratory of Sport and Health Science, School of Physical Education and Sport Sciences, Fujian Normal University, Fuzhou, China, **6** Institute for the Study of Endocrinology and Metabolism in Jiangxi, the Second Affiliated Hospital of Nanchang University, Jiangxi, China

‡ These authors share first authorship on this work.
* yu8220182@163.com (PY); Liux587@mail.sysu.edu.cn (XL)

Citation: Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, et al. (2023) How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health 2(12): e0000397. https://doi.org/10.1371/journal.pdig.0000397

Editor: Nicole Yee-Key Li-Jessen, McGill University, CANADA

Received: May 5, 2023

Accepted: October 23, 2023

Published: December 1, 2023

Data Availability Statement: The data is available in the supplemental material.

Funding: This work was supported by the Natural Science Foundation of China (No. 82160371 to J. Z., No. 82100869 and No. 82360162 to P.Y., 202201011395 to X.L.); Natural Science Foundation of Guangdong (202201011395 to X. L.); Basic and Applied Basic Research Project of Guangzhou (202201011395 to X.L.); Natural Science Foundation in Jiangxi Province grant [20224ACB216009 to J.Z.]; the Jiangxi Province

Abstract

ChatGPT, an artificial intelligence (AI) system powered by large-scale language models, has garnered significant interest in healthcare. Its performance dependent on the quality and quantity of training data available for a specific language, with the majority of it being in English. Therefore, its effectiveness in processing the Chinese language, which has fewer data available, warrants further investigation. This study aims to assess the of ChatGPT's ability in medical education and clinical decision-making within the Chinese context. We utilized a dataset from the Chinese National Medical Licensing Examination (NMLE) to assess ChatGPT-4's proficiency in medical knowledge in Chinese. Performance indicators, including score, accuracy, and concordance (confirmation of answers through explanation), were employed to evaluate ChatGPT's effectiveness in both original and encoded medical questions. Additionally, we translated the original Chinese questions into English to explore potential avenues for improvement. ChatGPT scored 442/600 for original questions in Chinese, surpassing the passing threshold of 360/600. However, ChatGPT demonstrated reduced accuracy in addressing open-ended questions, with an overall accuracy rate of 47.7%. Despite this, ChatGPT displayed commendable consistency, achieving a 75% concordance rate across all case analysis questions. Moreover, translating Chinese case analysis questions into English yielded only marginal improvements in ChatGPT's performance ($p = 0.728$). ChatGPT exhibits remarkable precision and reliability when handling the NMLE in Chinese. Translation of NMLE questions from Chinese to English does not yield an improvement in ChatGPT's performance.

Thousands of Plans (No. jxsq2023201105 to P.Y.); and the Hengrui Diabetes Metabolism Research Fund (No. Z-2017-26-2202-4 to P.Y.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Artificial Intelligence (AI) has been making significant strides in various fields, including healthcare. This study examines the proficiency of an AI system known as ChatGPT in understanding and responding to medical exam questions posed in the Chinese language. The researchers used a dataset from the Chinese National Medical Licensing Examination (NMLE) and found that ChatGPT performed well, scoring above the passing threshold. However, when the questions were altered to be more open-ended, the AI's performance declined significantly. Interestingly, translating the questions into English did not improve its performance. This research highlights the potential of AI, like ChatGPT, in assisting with medical education and clinical decision-making, particularly within a Chinese context. However, the findings also emphasize the need for further improvements, particularly for handling more complex, open-ended queries. The study contributes to understanding how AI can be effectively employed in non-English medical settings, and can be a steppingstone for future research in this area.

Introduction

AI (Artificial Intelligence) has gained significant influence in recent years by simulating human intelligence and cognitive processes to tackle complex problems [1]. Trained on specific datasets, AI systems enhance prediction accuracy and address complex challenges [2–4]. They assist doctors in rapidly searching through medical data, augmenting creativity, and facilitating error-free decision-making [5,6]. ChatGPT is a Large Language Model that predicts word sequences based on context and generates novel sequences resembling natural human language. These novel sequences have not been previously observed by other AI systems [7].

ChatGPT shows promise in medical education, performing well in Certified Public Accountant (CPA) exams and generating accurate responses to complex inputs [8]. Applied in the United States Medical Licensing Examination and South Korean parasitology exams, ChatGPT demonstrates significant advancements, despite discrepancies with medical students' scores [9]. However, ChatGPT's proficiency relies on available training data quality and quantity in the languages, and most of it is in English. With over 1.3 billion speakers, the amount and quality of training data in Chinese language may not be comparable to that in English, necessitating further research into ChatGPT's performance with Chinese medical information. The Chinese National Medical Licensing Examination (NMLE) is a legally mandated qualification for doctors [10]. This comprehensive, standardized assessment poses conceptually and linguistically challenging questions across medical domains, which makes it an excellent input for ChatGPT in clinical decision-making.

Given this background, this study aims to evaluate ChatGPT's performance on the Chinese NMLE conducted within the Chinese context.

Methods

Artificial intelligence

ChatGPT is an advanced language model that leverages self-attention mechanisms and extensive training data to deliver natural language responses within conversational settings. Its primary strengths include managing long-range dependencies and producing coherent, contextually appropriate responses. Nevertheless, it is essential to recognize that GPT-4 is a server-based language model without internet browsing or search capabilities. Consequently,

all generated responses rely solely on the abstract associations between words, or "tokens," within its neural network [7].

Input source

The official website does not release the 2022 NMLE test questions. However, for this study, a complete set of 600 questions, with a total value of 600 points, is available online (**S1 Table**) and considered as original questions. These questions are divided into four units, with each question worth one point.

The four units encompass the following areas: Unit 1 assesses medical knowledge, policies, regulations, and preventive medicine; Unit 2 focuses on cardiovascular, urinary, muscular, and endocrine systems; Unit 3 addresses digestive, respiratory, and other related systems; while Unit 4 evaluates knowledge of female reproductive systems, pediatric diseases, and mental and nervous systems.

All inputs provided to the GPT-4 model are valid samples that do not belong to the training dataset, as the database has not been updated since September 2021, predating the release of these questions. This was further confirmed by randomly spot checking the inputs. To facilitate research efforts, the 600 questions have been organized into distinct categories based on their question type and units.

1. Common Questions (n = 340): These questions are distributed across all units, including Unit 1 (n = 108), Unit 2 (n = 82), Unit 3 (n = 79), and Unit 4 (n = 71). They aim to evaluate basic science knowledge in physiology, biochemistry, pathology, and medical humanities. Each question has four choices, and the AI must select the single correct answer. An example from Unit 1 is: "What type of hypoxia is likely to be caused by long-term consumption of pickled foods? A. Hypoxia of blood type B. Hypoxia of tissue type C. Circulatory hypoxia D. Anoxic hypoxia E. Hypoxia of hypotonic type."

2. Case Analysis Questions (n = 260): These questions are also distributed across all units, including Unit 1 (n = 42), Unit 2 (n = 68), Unit 3 (n = 71), and Unit 4 (n = 79). These questions, employed in clinical medicine, examine and evaluate patient cases through a thorough review of medical history, symptoms, and diagnostic findings to determine a diagnosis and treatment plan. Each question has four choices, and the AI must select the single correct answer. An example from Unit 1 is: "A 28-year-old male complains of muscle and joint pain in his limbs three days after diving. He experienced respiratory equipment failure during diving three days ago and immediately ascended rapidly to the surface. Subsequently, he experienced symptoms such as dizziness, orientation disorder, nausea, and vomiting. After rest and oxygen inhalation, the symptoms improved, but he continued to experience persistent muscle spasms, convulsions, and joint pain in his limbs. Therefore, what is the most likely cause of the patient's pain? A. Chronic inflammation and cell infiltration B. Stress ulcers C. Local tissue coagulative necrosis D. Increased carbon dioxide concentration in the blood E. Gas embolism in the blood vessel lumen."

Scoring

We assembled a dataset of NMLE questions and their corresponding answers, maintaining validity by cross-verification with senior medical professionals. This dataset was used to evaluate ChatGPT's performance on the exam by comparing its responses to the standard answers and calculating the scores it achieved. A high score would indicate that ChatGPT effectively tackled this task.

Moreover, the passing score for the NMLE, established by the Department of Health's Board of Medical Examiners, is intended to determine whether an individual possesses the necessary skills to practice medicine independently and safely. In the 2022 Chinese NMLE, the passing score was set as 360. As such, ChatGPT's performance will be evaluated based on this benchmark.

Encoding

To better reflect the actual clinical situation, we modified the case analysis questions to be open-ended. Questions were formatted by deleting all the choices and adding a variable lead-in imperative or interrogative phrase, requiring ChatGPT to provide a rationale for the answer choice. Examples include: "What could be the most plausible explanation for the patient's nocturnal symptoms? Justify your answer for each option," and "Which mechanism is most likely responsible for the most fitting pharmacotherapy for this patient? Provide an explanation for its correctness."

However, a unique subset of questions, which required selecting from provided choices, could not be encoded in the same open-ended manner. These questions required selecting one provided choice, so we transformed them into a special form (n = 3). For example, the original question, "Which can inhibit insulin secretion? A. Increased free fatty acids in blood B. Increased gastric inhibitory peptide secretion C. Sympathetic nerve excitation D. Growth hormone secretion increases" was encoded as "Can an increase in free fatty acids in the blood, an increase in gastric inhibitory peptide secretion, an increase in sympathetic nerve excitation, or an increase in growth hormone secretion inhibit insulin secretion?" This encoding was present only in Unit 1.

To minimize the potential for ChatGPT to 'remember' previous answers and bias its responses, a new chat session was initiated for each question.

Adjudication

In our study, two physicians, blinded to each other's assessments, independently scored the AI outputs from the two types of encoders for accuracy and concordance. Scoring was based on predefined criteria (**S2 Table**). To train the physician adjudicators, a subset of 20 questions was used. ChatGPT's responses were categorized into three categories: accurate, inaccurate, and indeterminate. Accurate responses indicated that ChatGPT provided the correct answer, while inaccurate responses encompassed no answer, incorrect answers, or multiple answers with incorrect options. Indeterminate responses implied that the AI output did not provide a definitive answer selection or believed there was insufficient information to do so. Concordance was defined as when ChatGPT's explanation confirmed its provided answer, while discordant explanations contradicted the answer.

To minimize within-item anchoring bias, adjudicators first evaluated accuracy for all items, followed by concordance. Two physicians were blinded to each other's evaluations. In cases where the two initial physicians disagreed on a score, a third physician adjudicator was consulted. Ultimately, 17 items (2.7% of the dataset) required the intervention of a third physician adjudicator. The interrater agreement between the physicians was assessed using the Cohen kappa (κ) statistic for the questions (**S3 Table**).

A schematic overview of the study protocol is provided in **Fig 1.**

Translation

To evaluate if translating questions from Chinese to English could enhance ChatGPT's performance, We utilized ChatGPT to translate the case analysis questions, which had not yet been

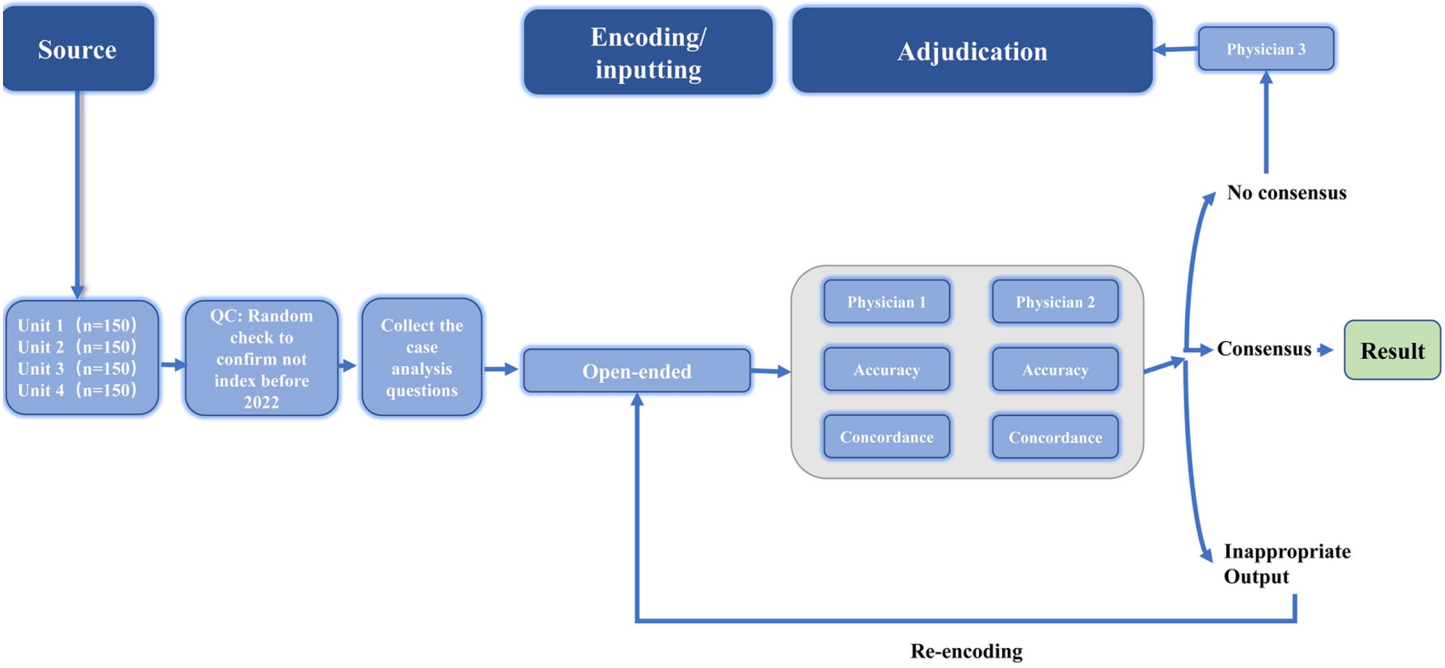

Fig 1. Schematic of workflow for sourcing, encoding, and adjudicating results. The 600 questions were categorized into 4 units. The accuracy of the open-ended encoded questions was evaluated, while the answer with forced justification encoded questions were also assessed for the accuracy, concordance. The adjudication process was carried out by two physicians, and in case of any discrepancies in the domains, a third physician was consulted for adjudication. Additionally, any inappropriate output was identified and required re-encoding.

https://doi.org/10.1371/journal.pdig.0000397.g001

encoded for the AI's testing. We then assessed ChatGPT's performance on the translated exam by comparing its responses to standard answers and calculating its scores (S4 Table). We compared the scores obtained from the original questions to those from the translated questions and employed the chi-square test to determine performance improvement.

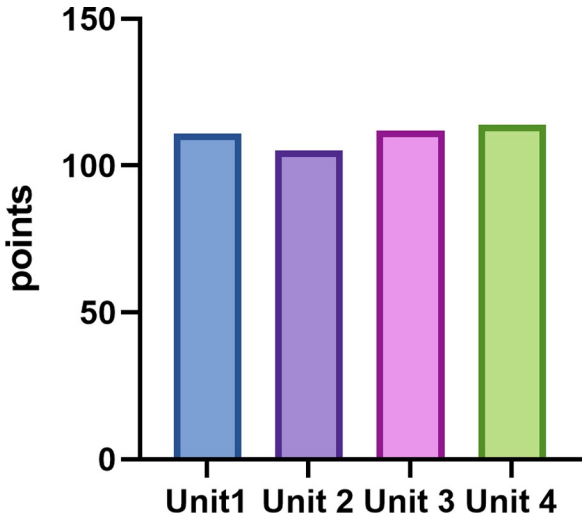

Fig 2. Score of ChatGPT4 on Chinese National Medical Licensing Examination before encoding. ChatGPT's outputs from Units 1, 2, 3, and 4 were scored for each unit.

https://doi.org/10.1371/journal.pdig.0000397.g002

Statistical analysis

Initially, we used the Cohen kappa (κ) statistic to assess the degree of consensus among physicians. This was achieved by comparing the responses of ChatGPT to the standard answers. In cases where the initial two physician adjudicators disagreed on a score, a third physician was consulted. Following this, we examined if translating questions from Chinese to English could potentially enhance ChatGPT's performance. We juxtaposed ChatGPT's performance on the original and translated questions, and utilized the chi-square test to determine if there were any statistically significant performance improvements.

All statistical calculations were performed using the SPSS software package.

We believe these revisions address your comments adequately and we eagerly await any other feedback you may have.

Result

ChatGPT passed Chinese NMLE with a high score

In the Chinese NMLE, ChatGPT correctly answered 442 (73.67%) out of 600 items, a score significantly higher than the passing threshold of 360 as defined by official agencies.

The score of each unit is shown in Fig 2. The performance of ChatGPT varied across the four units of questions, with the highest accuracy being in Unit 4 (76.0%), followed by Unit 3 (74.7%), Unit 1 (74.0%) and Unit 2 (70.0%), while there was no statistically difference among four units ($\chi^2 = 0.66$, $p = 0.883$).

ChatGPT's performance declines when handling encoded questions

Test questions were encoded as open-ended for case analysis questions, simulating scenarios where a student poses a common medical question without answer choices or a doctor diagnoses a patient based on multimodal clinical data (e.g., symptoms, history, physical examination, laboratory values). The accuracy was 40.5%, 60.3%, 42.3%, and 34.2% for Units 1, 2, 3, and 4, respectively (**Fig 3A**). Compared to the original questions, the accuracy of the encoded questions decreased by 40.5%, 9.7%, 32.4%, and 41.8% for Units 1, 2, 3, and 4, respectively (**Fig 3B**).

These findings demonstrate that while ChatGPT's ability to answer common medical questions in Chinese is commendable, there is still room for improvement. During the adjudication stage, physician agreement was good for open-ended questions (with a κ range from 0.83 to 1.00).

ChatGPT demonstrates high internal concordance

Concordance is a measure of the agreement or similarity between the option selected by AI and its subsequent explanation. The results indicated that ChatGPT maintained a >75% concordance across all questions, and this high level of concordance was consistent across all four units (**Fig 4**). Furthermore, we examined the concordance difference between correct and incorrect answers, discovering that concordance was perfect and significantly higher among accurate responses compared with inaccurate ones (85% vs. 59.5%, $p<0.005$) (**Fig 4**).

These findings suggest that ChatGPT exhibits a high level of answer-explanation concordance in Chinese, which can be attributed to the strong internal consistency of its probabilistic language model.

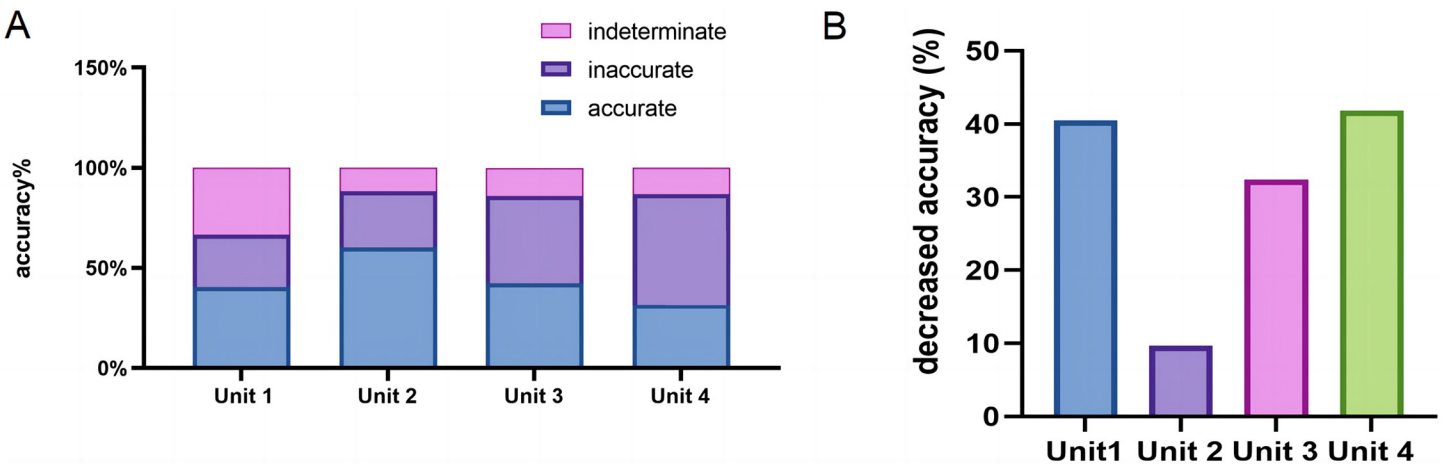

Fig 3. Accurancy of ChatGPT4 on Chinese National Medical Licensing Examination before encoding. ChatGPT's outputs for Units 1, 2, 3, and 4 were evaluated as accurate, inaccurate, or indeterminate using the scoring system outlined in S2 Table after encoding. (A) Assessment of accuracy for open-ended question encodings. (B) Reduced accuracy of encoded questions across Units 1, 2, 3, and 4.

https://doi.org/10.1371/journal.pdig.0000397.g003

Translating the input into English may not improve ChatGPT's performance

After translating the original case analysis questions in Chinese into English to explore a potential way to improve ChatGPT's performance, the improvement for Units 1, 2, 3, and 4 was minimal, with only one point gained in each unit. The total number of correct answers increased from 256 to 260. The accuracy improvement for translated case analysis questions was subtle ($\chi^2 = 0.1206$, P = 0.728). This suggests that ChatGPT's performance when facing

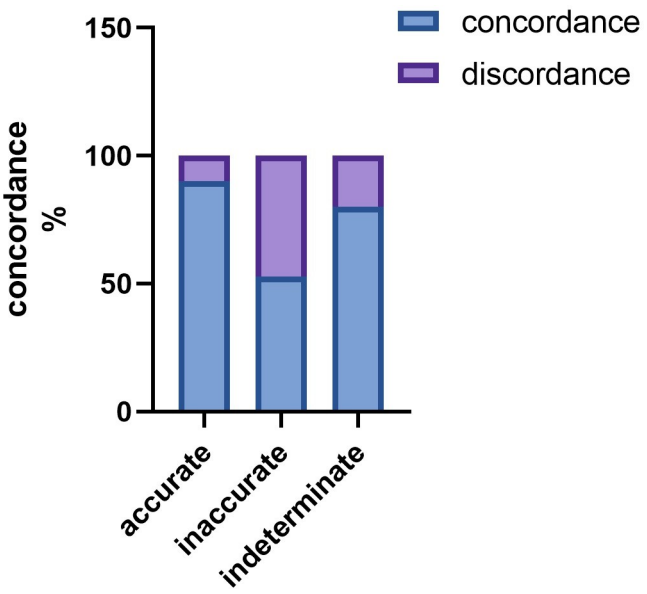

Fig 4. Concordance of ChatGPT4 on Chinese National Medical Licensing Examination after encoding. For Units 1, 2, 3, and 4 after encoding, ChatGPT's outputs were evaluated as concordant or discordant, based on the scoring system detailed in S2 Table. This figure demonstrates the concordance rates stratified between accurate, inaccurate, and indeterminate outputs, across all case analysis questions.

https://doi.org/10.1371/journal.pdig.0000397.g004

questions in Chinese may not be improved by translating them into English, and solely relying on translating other languages into English and building a database exclusively in English may not be the most effective approach.

Discussion

In this study, we firstly investigated ChatGPT's performance on the Chinese NMLE. Our findings can be summarized under two major themes: (1) ChatGPT's score is satisfactory but requires improvement when addressing questions posed in the Chinese language; and (2) Translation into English may not improve ChatGPT's performance. This study provides new evidence for the ability of ChatGPT in medical education and clinical decision-making within the Chinese context, offering valuable insights into the applicability of AI language models for non-English medical education settings and laying the groundwork for future research in this area.

ChatGPT's performance in the Chinese NMLE is acceptable, yet further improvement

In the Chinese NMLE, ChatGPT achieved a score of 442 (73.67%), exceeding the passing requirement of 360 points for the exam in the Chinese language. In the 2022 NMLE, 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%). According to the statistics, the national pass rate for the exam in 2022 was 55%. When compared to medical students who have undergone a traditional 5-year medical education and a one-year internship, ChatGPT's performance is currently satisfactory, however, there is still potential for improvement. Several underlying reasons may be responsible for this, such as: 1) Limitations in training data: If ChatGPT's training data contains less information about the Chinese medical field, its performance when handling Chinese medical questions could be impacted, resulting in a lower accuracy rate for such queries. 2) Knowledge updates: With a knowledge cutoff date in September 2021, the most recent developments in the Chinese medical field may not have been adequately learned by the model, affecting its accuracy when answering Chinese medical questions.

ChatGPT's accuracy can be improved by addressing data limitations, refining its architecture, and using domain-specific knowledge

Moreover, we observed that outputs with high accuracy showed high concordance, while lower accuracy was associated with reduced concordance. Consequently, we speculate that ChatGPT's inaccurate responses primarily arise from missing information, leading to indetermination or inaccurate in the AI. Language models like ChatGPT are built on vast amounts of text, and their accuracy depends on the quality and diversity of their training data [11]. When the model encounters scenarios with limited or underrepresented data, its performance may suffer, leading to indecision or inaccurate responses. To address this issue, one could consider expanding the training data to cover a broader range of contexts or refining the model's architecture to handle uncertainty more effectively. Additionally, incorporating domain-specific knowledge and data sources can help improve the model's performance in specialized areas.

ChatGPT performs best in English, with accuracy affected by translation issues and data limitations in other languages

ChatGPT's performance varies across languages due to different factors, primarily the quality and quantity of the training data available in each language [12]. English, as the language with the most abundant training data, typically yields the highest accuracy. However, our findings

indicate that translating questions into English did not significantly improve ChatGPT's performance. This suggests that a strategy relying solely on English databases or translations might not be the most effective approach for enhancing ChatGPT's proficiency in handling medical tasks. There are two potential reasons for this observation: 1) Translation limitations: The process of translation can often lead to the loss of certain nuances or the inaccurate translation of specific terms, which could impact the AI's understanding of the question and subsequently its performance [13]. Additionally, some languages may have unique expressions or cultural contexts that are difficult to convey accurately in English. This can lead to potential misunderstandings or misinterpretations that could significantly influence the performance of language models like ChatGPT. 2) Different languages may have semantic and cultural differences, and certain unique expressions or cultural contexts in one language may be challenging to accurately convey in English. This can result in potential misunderstandings or misinterpretations, significantly influencing the performance of language models like ChatGPT.

Therefore, in order to accurately assess the performance of an AI model like ChatGPT for a particular language or task, a more targeted evaluation may be necessary. This would involve conducting a thorough analysis of the AI's performance using a diverse range of tasks and data sources that are representative of the language in question.

GPT-4 shows progress, but addressing healthcare standards, ethics, and culture is crucial for AI integration in medicine

ChatGPT-3.5 achieved near-passing threshold accuracy of 60% on the United States Medical Licensing Exam [7]. Furthermore, our previous study also showed a similar performance by ChatGPT-3.5 on the Clinical Medicine Entrance Examination for Chinese Postgraduates (scored 153.5/300, 51%) in Chinese language [12]. In the present study, a significant higher score was found by GPT-4 (scored 442/600,73.6%). This improvement may be attributed to differences in model sizes and training data. GPT-4's larger model size enables it to handle more complex tasks and generate more accurate responses due to its extensive training dataset, broader knowledge base, and improved contextual understanding [13]. On the other hand, Chinese medical licensing exams have many common-sense questions and fewer case analysis questions than United States Medical Licensing Exam, which may be other reasons for the relatively high pass rates.

Despite the promising potential of AI in medicine, it also faces several challenges. The development of standards for AI use in healthcare is still necessary [14,15], encompassing clinical care, quality, safety, malpractice, and communication guidelines. Moreover, the implementation of AI in healthcare necessitates a shift in medical culture, posing challenges for both medical education and practice. Ethical considerations, such as data privacy, informed consent, and bias prevention, must also be addressed to ensure that AI is employed ethically and for the benefit of patients.

Limitations

Several limitations should be noted. Firstly, clinical tasks are highly complicated, the exams cannot fully stimulate the problems in clinical practices. Secondly, the limited input sample size may preclude us from performing analyses in depth and range, potentially limiting the generalizability of our findings. Thus, before large-scale application of AI based on Large Language Models in medical education or clinical practice, their utility should be further studied under real-world conditions.

Moreover, we acknowledge the potential benefits of diversifying datasets in our analysis and evaluating examples. However, due to time limitations during data collection and analysis,

we exclusively used the one-year NMLE dataset in our study. Our research was conducted after September 2022 when the ChatGPT dataset was updated. To avoid encountering the same questions used for training purposes, we relied solely on data from 2022 onwards. Additionally, since questions for the year 2023 were unavailable at the time of our research, we were constrained to using only one year's worth of medical licensing examination questions.

Conclusion

ChatGPT demonstrated impressive performance on the Chinese NMLE, exceeding the passing threshold and exhibiting high internal consistency. Nevertheless, its performance waned when faced with open-ended encoded questions. Translation into English did not substantially boost its performance. The findings emphasize ChatGPT's ability for comprehensible reasoning in medical education and clinical decision-making in Chinese.

Supporting information

S1 Table. The original question.
(DOCX)

S2 Table. Adjudication criteria for accuracy, concordance.
(DOCX)

S3 Table. Kappa statistic for interrater agreement between adjudicating physicians.
(DOCX)

S4 Table. Crosstab for evaluating translation effectiveness.
(DOCX)

Author Contributions

Conceptualization: Peng Yu, Xiao Liu.

Data curation: Changchang Fang, Yuting Wu, Wanying Fu, Jitao Ling, Yue Wang, Yuan Jiang, Yixuan Chen, Jing Zhou, Zhichen Zhu.

Formal analysis: Yuting Wu, Wanying Fu, Jitao Ling, Yuan Jiang, Yixuan Chen, Jing Zhou.

Funding acquisition: Peng Yu.

Investigation: Changchang Fang, Yixuan Chen, Jing Zhou.

Methodology: Yuting Wu, Jitao Ling, Yixuan Chen, Jing Zhou.

Project administration: Xiao Liu.

References

1. Haleem A, Javaid M, Khan IH. Current status and applications of Artificial Intelligence (AI) in medical field: An overview. Current Medicine Research and Practice. 2019; 9(6):231–7. https://doi.org/10.1016/j.cmrp.2019.11.005
2. Haleem A, Vaishya R, Javaid M, Khan IH. Artificial Intelligence (AI) applications in orthopaedics: An innovative technology to embrace. Journal of Clinical Orthopaedics and Trauma. 2019;(0976–5662 (Print)). https://doi.org/10.1016/j.jcot.2019.06.012 PMID: 31992923
3. Jha S, Topol EJ. Information and artificial intelligence. Journal of the American College of Radiology. 2018; 15(3):509–11. https://doi.org/10.1016/j.jacr.2017.12.025 PMID: 29398501
4. Lupton ML, editor Some ethical and legal consequences of the application of artificial intelligence in the field of medicine2018.

5. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;(1538–3598 (Electronic)). https://doi.org/10.1001/jama.2013.393 PMID: 23549579

6. Misawa M, Kudo S-e, Mori Y, Cho T, Kataoka S, Yamauchi A, et al. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. Gastroenterology. 2018; 154(8):2027–9. https://doi.org/10.1053/j.gastro.2018.04.003 PMID: 29653147

7. 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 digital health. 2023; 2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198 PMID: 36812645

8. Bommarito J, Bommarito M, Katz DM, Katz JJapa. GPT as Knowledge Worker: A Zero-Shot Evaluation of (AI) CPA Capabilities. 2023. https://doi.org/10.48550/arXiv.2301.04408

9. Sun H. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. Journal of Educational Evaluation for Health Professions. 2023; 20:1. https://doi.org/10.3352/jeehp.2023.20.01

10. Xiancheng W. Experiences, challenges, and prospects of National Medical Licensing Examination in China. BMC Medical Education. 2022; 22(1):349. https://doi.org/10.1186/s12909-022-03385-9 PMID: 35527240

11. Almazyad M, Aljofan F, Abouammoh NA, Muaygil R, Malki KH, Aljamaan F, et al. Enhancing Expert Panel Discussions in Pediatric Palliative Care: Innovative Scenario Development and Summarization With ChatGPT-4. Cureus. 2023; 15(4).

12. Liu X, Fang C, Wang J. Performance of ChatGPT on Clinical Medicine Entrance Examination for Chinese Postgraduate in Chinese. medRxiv. 2023:2023.04. 12.

13. Butler S. GPT 3.5 vs GPT 4: What's Difference 2023 [cited 2023 MAR 31]. Available from: https://www.howtogeek.com/882274/gpt-3-5-vs-gpt-4/.

14. F. D-V. Considerations for the Practical Impact of AI in Healthcare Food and Drug Administration. 2023.

15. Zweig M EBRH. How should the FDA approach the regulation of AI and machine learning in healthcare? 2018. Available from: https://rockhealth.com/how-should-the-fda-approach-the-regulation-of-ai-and-machine-learning-in-healthcare/.