BMJ Health & Care Informatics

ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model

Manoochehr Ebrahimian ,[1] Behdad Behnam,[2] Negin Ghayebi,[3] Elham Sobhrakhshankhah[2]

To cite: Ebrahimian M, Behnam B, Ghayebi N, *et al*. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. *BMJ Health Care Inform* 2023;**30**:e100815. doi:10.1136/bmjhci-2023-100815

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2023-100815).

Received 30 May 2023
Accepted 28 November 2023

[1]Pediatric Surgery Research Center, Research Institute for Children's Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2]Gastrointestinal and Liver Disease Research Center, Iran University of Medical Sciences, Tehran, Iran
[3]School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Correspondence to
Dr Manoochehr Ebrahimian; manoochehrebrahimian@gmail.com

ABSTRACT

Introduction Large language models such as ChatGPT have gained popularity for their ability to generate comprehensive responses to human queries. In the field of medicine, ChatGPT has shown promise in applications ranging from diagnostics to decision-making. However, its performance in medical examinations and its comparison to random guessing have not been extensively studied.

Methods This study aimed to evaluate the performance of ChatGPT in the preinternship examination, a comprehensive medical assessment for students in Iran. The examination consisted of 200 multiple-choice questions categorised into basic science evaluation, diagnosis and decision-making. GPT-4 was used, and the questions were translated to English. A statistical analysis was conducted to assess the performance of ChatGPT and also compare it with a random test group.

Results The results showed that ChatGPT performed exceptionally well, with 68.5% of the questions answered correctly, significantly surpassing the pass mark of 45%. It exhibited superior performance in decision-making and successfully passed all specialties. Comparing ChatGPT to the random test group, ChatGPT's performance was significantly higher, demonstrating its ability to provide more accurate responses and reasoning.

Conclusion This study highlights the potential of ChatGPT in medical licensing examinations and its advantage over random guessing. However, it is important to note that ChatGPT still falls short of human physicians in terms of diagnostic accuracy and decision-making capabilities. Caution should be exercised when using ChatGPT, and its results should be verified by human experts to ensure patient safety and avoid potential errors in the medical field.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Recent studies reported promising results in various medical licensing examinations within multiple countries.

WHAT THIS STUDY ADDS

⇒ This study specifically focused on the Iranian medical licensing examination, which serves as the final comprehensive test for medical students which examines the general information of a medical student before entering to clinical practice. Additionally, this study assessed various aspects of ChatGPT, including its proficiency in basic sciences, diagnosis and decision-making. Furthermore, the performance of ChatGPT was compared with that of randomly selected answers of computer examinees.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The future of medical education will be affected remarkably by artificial intelligence systems. ChatGPT, which is a publicly available response chatbot, can be replaced with more specific and customised chatbots that can respond to queries based on textbooks, articles and scientific papers. Also, policymakers may use more advanced versions of these chatbots as an initial evaluation of non-emergent diseases.

INTRODUCTION

With the emergence of large language models (LLMs) and advancements in natural language processing, numerous artificial intelligence (AI) models have been developed to generate comprehensive responses to human queries.[1] Recently, among these LLMs, ChatGPT, a generative pretrained transformer developed by OpenAI, has gained significant popularity.[2] This chatbot uses its embedded neural network to interact with users and provide precise responses. Therefore, ChatGPT has found utility in various fields, including the social sciences, language editing and translation, becoming a valuable tool for many individuals in their daily tasks. It has also made a remarkable impact on the field of medical sciences.[3] Numerous studies have explored different applications of ChatGPT in medical sciences, ranging from diagnostic evaluation and basic sciences to treatment, decision-making and even scholarly writing such as manuscripts, letters, presentations and also books.[4] However, it is important to note that despite its promising capabilities, ChatGPT still requires debugging and further development, as it has been found to make errors in medical decision-making.[5] But it still works better than the majority of real human physicians in medical examinations. Recently, the performance of ChatGPT in medical licensing examinations have been investigated in different countries.[1 6 7] With all this in mind, we intended to evaluate the performance of ChatGPT in medical licensing examination of Iran, called preinternship, which is the final comprehensive assessment of medical students before graduation. Furthermore, we sought to compare the results obtained from ChatGPT with those from a self-made random computer examinee in order to identify the chance of passing the exam by computer without using AI models.

METHODS

We selected the latest preinternship examination (March 2023), which is a standardised integrated medical exam that medical students in Iran are required to pass after completing their clerkship. This is the final comprehensive medical examination in which every medical doctor encounters before graduation. This comprehensive examination serves as the final assessment before graduation and covers various medical specialties. The exam primarily focuses on four major specialties: internal medicine, surgery, paediatrics and gynaecology, which collectively account for over 50% of the questions. The remaining questions cover other specialties including neurology, infectious diseases, radiology, pathology, dermatology, orthopaedics, urology, ophthalmology, otorhinolaryngology, pharmacology and biostatistics. Each examination consists of 200 multiple-choice questions (MCQs) with 4 options provided for each question and only 1 option being correct. The questions can be categorised into three types, namely basic science evaluation, diagnosis and decision-making. All the questions were translated to English before entering to ChatGPT chatbot. Regarding the complexity of Farsi language, online translators are unable to translate accurately from Farsi to English. For this reason, an expert medical specialist who was familiar with English medical exams changed the structure of questions to be readable for a native English speaker. Each examinee must pass the exam with at least 90 correct answers (45%). Unfortunately, due to confidentiality of information, we have not access to each individual's score. However, the average result of this particular exam was 114, with 6.8% of students failed the exam. To ensure accuracy, each question was carefully reviewed by two authors for any language or grammar errors. The questions were then entered into separate ChatGPT message boxes in consecutive order. We used GPT-4 (23 March 2023) for this study. After receiving responses from ChatGPT, the selected option for each question was entered into an electronic collaborative spreadsheet. The authors verified the results, and finally, the data were analysed and compared with the pass mark for each question type and specialty (figure 1).

Random test group

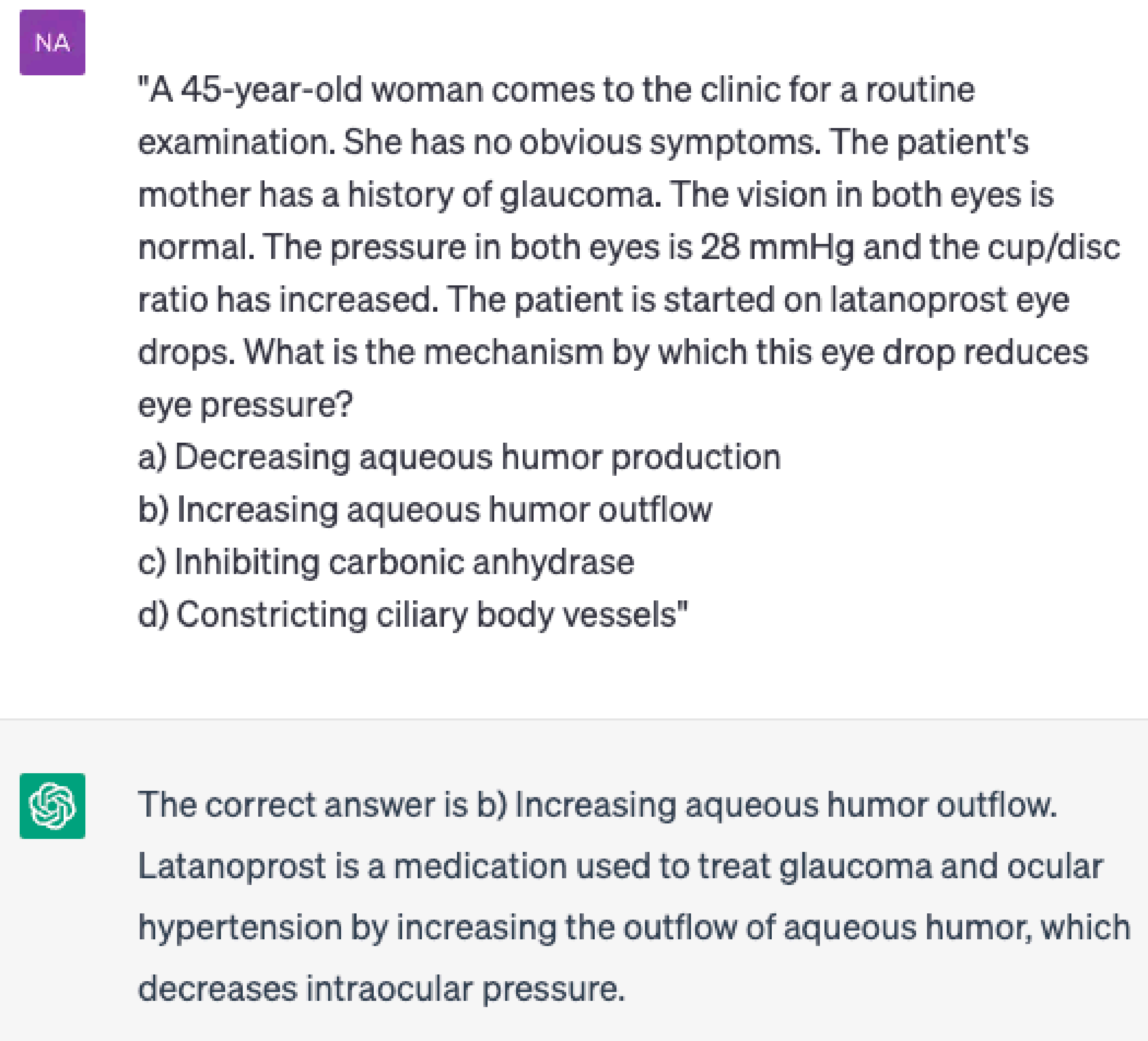
In addition to using ChatGPT as a test participant, we also employed computer random testers to generate results representing the performance of random examinees. To achieve this, we used a Java class called MassMCQExaminees, which is provided in the appendix file. This software incorporates three input variables: the total number of tests, the number of choices for each question and the number of examinees. The programme then randomly selects answers for the 200-question test for each computer participant (see online supplemental appendix). The average percentage of correct answers for each examinee will be calculated and presented in the console (eg, 24.6%). This virtual examination could show us what would be the exam result, if a participant chooses the answers randomly. The random average mark also could be compared with ChatGPT in order to show the rational thinking and reasoning abilities of the AI system.

Statistical analysis

To assess the performance of ChatGPT in the preinternship examination, we employed several statistical analyses, including frequency, χ^2 and simple t-tests. Initially, we calculated the frequency distribution of correct answers for each question type and specialty. This allowed us to understand the distribution of correct responses across different categories. Next, we conducted χ^2 tests to assess the association between correct answers and question types or specialties. This analysis helped determine if there were any significant relationships or dependencies between the correct answers and specific question types or specialties. The simple t-test was used to compare the mean correct answers of ChatGPT with the pass mark cutoff (0.45 for our study) and random test group results. Also, we used the cumulative distribution function (CDF) of the binomial distribution to find the chance of passing the exam with random answers.

RESULTS

Table 1 presents the results of the exams conducted for each question type. It is evident that ChatGPT performed exceptionally well and was able to pass the examination with a high score. The number of correct answers obtained by ChatGPT was 137 (68.5%), which is significantly higher than the pass mark of 45% ($p<0.001$). Although ChatGPT demonstrated a superior performance in the area of decision-making compared with other question types, there were no significant differences observed between each question type. Additionally, table 2 shows the performance of ChatGPT in each specialty. In the table, there are two columns indicating the rate of correct answers regarding specialty, and correct diagnosis rate (eg, in some questions, AI could not be able to find the correct answer but diagnosed the disease). ChatGPT successfully



"A 45-year-old woman comes to the clinic for a routine examination. She has no obvious symptoms. The patient's mother has a history of glaucoma. The vision in both eyes is normal. The pressure in both eyes is 28 mmHg and the cup/disc ratio has increased. The patient is started on latanoprost eye drops. What is the mechanism by which this eye drop reduces eye pressure?

a) Decreasing aqueous humor production

b) Increasing aqueous humor outflow

c) Inhibiting carbonic anhydrase

d) Constricting ciliary body vessels"

The correct answer is b) Increasing aqueous humor outflow. Latanoprost is a medication used to treat glaucoma and ocular hypertension by increasing the outflow of aqueous humor, which decreases intraocular pressure.

Figure 1 An example of MCQ test given to ChatGPT. MCQ, multiple-choice question.

Table 1 The results of exam sorted by the type of questions

Questions (%)	Type	Correct answers	Pass mark (%)	P value*
Diagnosis	43 (21.5)	29 (67.4)	45	0.003
Basic science	84 (42.0)	54 (64.3)	45	<0.001
Decision-making	73 (36.5)	54 (74)	45	<0.001
Total	200	137 (68.5)	45	<0.001

*P values were calculated with one-sample t-test with the test value of 0.45. Also, the p value for question type and correct answers was estimated by χ^2 test (0.422).

passed all of the specialties, with the highest performance observed in gynaecology and surgery (78.9% and 75%, respectively). However, the performance of ChatGPT in various specialties was not significantly different (p=0.702). It is noteworthy to mention that even in questions where ChatGPT failed to find the correct answer, the exact diagnosis in some cases was correct (175 correct diagnosis). However, diagnostic ability was much higher in surgery and internal medicine, in comparison to paediatrics, gynaecology and other specialties (p=0.032).

Comparison to random examinees

To provide a benchmark for comparison, we used the previously mentioned Java class to generate results from 5000 computer participants who answered each question randomly, without recalling any database. As anticipated,

Table 2 Comparison of each subspecialty

Field	Total questions	Correct answers (%)	Correct diagnosis (%)
Internal medicine	46	32 (69.6)	44 (95.7)
Surgery	24	18 (75)	24 (100)
Paediatrics	24	16 (66.7)	19 (79.2)
Gynaecology	19	15 (78.9)	17 (89.5)
Other specialties*	87	56 (64.4)	71 (81.6)
Total	200	137 (68.5)	175 (87.5)
P value		0.702	0.032

*Other specialties include neurology, infectious diseases, radiology, pathology, dermatology, orthopaedics, urology, ophthalmology, otorhinolaryngology, pharmacology and biostatistics.

the average percentage of correct answers among the random testers was approximately 25%. This aligns with expectations since each MCQ had one correct answer out of four choices. Conducting a simple t-test, we found that the performance of ChatGPT was significantly higher than that of the random test group ($p<0.001$). This indicates that ChatGPT's performance surpasses that of random guessing, demonstrating its ability to provide more accurate responses and reasoning in the context of the preinternship examination.

Probability of passing the examination with an excellent mark

As we stated earlier, for probability estimation of at least 90 correct answers (the cut-off used by Iranian National Organization of Medical Educational Testing), we employed CDF of the binomial distribution. Therefore, the probability of getting less than 90 correct answers is:

$$\mathrm{P}\left(\mathrm{X}<90\right)=\sum_{k=0}^{89}\binom{200}{k}\left(\frac{1}{4}\right)^{k}\left(\frac{3}{4}\right)^{200-k}$$

Consequently, the probability of having at least 90 correct answers (passing the exam) with random testing is the complementary probability as shown below:

$$P\left(X\geq 90\right)=1-P\left(X<90\right)\approx 0.041=4.1\%$$

As observed, while computer random-test models were able to answer an average of 50 questions correctly (25%), the probability of passing the exam with random answers was very low, at 4.1%. However, ChatGPT passed the exam with the notable mark at 68.5%. To investigate whether ChatGPT's performance could be attributed to chance, we conducted a simple t-test using test values of both 25% and 4.1%. The resulting p values were very small, which rejects the hypothesis that ChatGPT's exam results were achieved by random test answering systems. It is noteworthy that if we use the binomial distribution to estimate the probability of a person answering exactly 137 questions correctly, the probability is calculated to be 0.0038%. This highlights the exceptional performance of ChatGPT and demonstrates that its success in the exam is highly unlikely to be a result of random chance.

DISCUSSION

The results demonstrate that ChatGPT achieved an excellent mark in the examination, surpassing the pass mark by a significant margin. This indicates that ChatGPT has the potential to accurately answer medical-related MCQs and perform well in different question types, including basic science evaluation, diagnosis and decision-making. The superior performance of ChatGPT in decision-making aligns with its advanced natural language processing capabilities and its ability to generate precise responses. These findings are consistent with previous studies that have highlighted the effectiveness of ChatGPT in various domains, including medical sciences.

In a similar study evaluating the performance of ChatGPT in the US Medical Licensing Examination (USMLE), the AI-based model achieved a correct answer rate of over 60%, which is equivalent to the passing score of a well-trained third-year medical student.[8] Moreover, another study focusing on the three steps of the USMLE, involving 376 MCQs, also reported similar passing scores of 60%.[9] Additionally, multiple studies have assessed the performance of public AI language models, such as ChatGPT, in medical licensing examinations, particularly in various specialties and subspecialties. These studies have demonstrated the success of AI models in disciplines such as radiology, ophthalmology, medical physiology and plastic surgery.[10–13] It is important to note that these studies primarily evaluated the performance of ChatGPT with exams conducted in the English language. However, studies on Chinese medical exams have reported suboptimal results with ChatGPT.[14] This suggests that the accuracy of ChatGPT may vary across different languages which is consistent with existing data. This forced the researchers to translate their MCQs to English in order to mitigate this bias. In this study, we also ensured the translation of questions to English by an expert in medical field terminology. On the other hand, there are reports suggesting the incompetency of ChatGPT in specific medical examinations. For instance, ChatGPT did not reach the passing threshold for life support exams, in a scenario-based examination.[5] However, these controversial results may be due to technical problems like using non-standardised questions, unclear queries and also designing exams with not publicly available information.

As mentioned previously, several studies confirmed the accuracy of ChatGPT in various medical licensing examinations. However, there is a shortage of large-scale comprehensive studies in this context. It is important to note that despite its impressive performance, ChatGPT may not be able to answer all questions correctly, which is to be expected by researchers. One possible reason for this is that ChatGPT is a publicly accessible AI bot that relies on publicly available large databases to generate responses, rather than specific textbooks or specialised medical resources.

It is expected that using custom learning models based on ChatGPT's application programming interface (API), can improve the performance of ChatGPT in medical exams.[15 16] These personalised APIs may improve the results significantly in every medical field, especially on those where ChatGPT could not reach the passing mark.

An intriguing question that arises is whether AI-based models, such as ChatGPT, can replace human physicians. At present, the answer remains negative. While ChatGPT has shown remarkable performance in certain tasks, it still lags behind human physicians in terms of diagnostic accuracy and decision-making capabilities. The expertise, clinical judgement and nuanced understanding of complex medical cases that human physicians possess are not yet fully replicated by AI models.

On the other hand, the impact of using AI-based models in medical field led to some concerns for health policy-makers. Although using these models may come with potential advantages such as understanding complex language structures and categorising unstructured data, there are also some drawbacks. For example, the level of transparency in decision-making progress, ethical concerns and comprehensibility of AI-based answers may lead to potentially harmful consequences.[17] Fortunately, in this study, the performance of ChatGPT in decision-making was acceptable. But using these models in real-time clinical situations is still under question. Additionally, medical students will use AI-based models in their daily educational programmes, which is another potential risk for their traditional human learning process, and in a long-time period may decrease the capabilities of human doctors.[18] Therefore, regarding the usage of AI-based models in the modern era is inevitable, health and medical policy-makers should be aware of the advantages and disadvantages of these AI tools and adjust their national laws due to newly coming changes. In addition, there are limited data on usefulness as well as accuracy of AI in different aspects of medicine for patients and healthcare workers. Although AI may be a powerful tool in medicine, it should be used with caution by healthcare workers.[19] Overall, the ability of AI to answer MCQs will not reflect its usefulness in clinical practice as well as it could not been replaced by expert's decision yet.

Finally, like any other study, there are several limitations to consider in this work. First, the evaluation of ChatGPT's performance was solely based on its ability to answer MCQs in the preinternship examination. This narrow focus may not fully capture the complex and nuanced decision-making process required in real-world medical scenarios. Additionally, the study only assessed the performance of ChatGPT in a single country's medical licensing examination, which may limit the generalisability of the findings to other healthcare systems and contexts. Further research is needed to explore the potential biases, limitations and ethical considerations associated with the use of AI models like ChatGPT in medical settings.

CONCLUSION

The results indicated that ChatGPT showed promising capabilities in answering MCQs and achieved an excellent mark, surpassing the pass mark by a significant margin. This highlights the potential of ChatGPT in assisting medical professionals by providing accurate responses and supporting decision-making processes. However, it is important to acknowledge that ChatGPT still has limitations and should not be seen as a substitute for human physicians. Also, personalised custom API's with the ability to answer the prompts based on medical textbooks may improve the performance of AI-based models significantly.

Contributors ME and BB designed the study; NG wrote the manuscript; ES evaluated the validity of data; ME cowrote the manuscript. ME is guarantor.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Manoochehr Ebrahimian http://orcid.org/0000-0003-1286-3308

REFERENCES

1 Gilson A, Safranek C, Huang T, *et al*. How does Chatgpt perform on the medical licensing exams the implications of large language models for medical education and knowledge assessment. *medRxiv* 2022.
2 Vaishya R, Misra A, Vaish A. Chatgpt: is this version good for healthcare and research?. *Diabetes Metab Syndr* 2023;17:102744.
3 Homolak J. Opportunities and risks of Chatgpt in medicine, science, and academic publishing: a modern Promethean dilemma. *Croat Med J* 2023;64:1–3.
4 van Dis EAM, Bollen J, Zuidema W, *et al*. Chatgpt: five priorities for research. *Nature* 2023;614:224–6.
5 Fijačko N, Gosak L, Štiglic G, *et al*. Can Chatgpt pass the life support exams without entering the American heart association course? *Resuscitation* 2023;185.
6 Kaneda Y, Tanimoto T, Ozaki A, *et al*. Can Chatgpt pass the 2023 Japanese national medical licensing examination? *Med Pharmacol* [Preprint] 2023.
7 Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: is Chatgpt a blessing or blight in disguise? [Taylor & Francis]. *Med Educ Online* 2023;28:2181052.
8 Gilson A, Safranek CW, Huang T, *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;9:e45312.

9 Kung TH, Cheatham M, Medenilla A, *et al*. Performance of Chatgpt on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198.

10 Bhayana R, Krishna S, Bleakney RR. Performance of Chatgpt on a radiology board-style examination: insights into current strengths and limitations. *Radiology* 2023;307:e230582.

11 Humar P, Asaad M, Bengur FB, *et al*. Chatgpt is equivalent to first-year plastic surgery residents: evaluation of Chatgpt on the plastic surgery in-service examination. *Aesthet Surg J* 2023;43:1085–9.

12 Antaki F, Touma S, Milad D, *et al*. Evaluating the performance of Chatgpt in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci* 2023;3:100324.

13 Subramani M, Jaleel I, Krishna Mohan S. Evaluating the performance of Chatgpt in medical physiology University examination of phase I MBBS. *Adv Physiol Educ* 2023;47:270–1.

14 Wu J, Wu X, Qiu Z, *et al*. Qualifying Chinese medical licensing examination with knowledge enhanced generative pre-training model [arXiv:230510163]. *arXiv* 2023.

15 Tsai Y-C. Empowering learner-centered instruction: integrating Chatgpt python API and Tinker learning for enhanced creativity and problem-solving skills [arXiv:230500821]. *arXiv* 2023.

16 Chen E, Huang R, Chen H-S, *et al*. Gptutor: a Chatgpt-powered programming tool for code explanation [arXiv:230501863]. *arXiv* 2023.

17 Sifat RI. Chatgpt and the future of health policy analysis: potential and pitfalls of using Chatgpt in policymaking. *Ann Biomed Eng* 2023;51:1357–9.

18 Khan RA, Jawaid M, Khan AR, *et al*. Chatgpt - reshaping medical education and clinical management. *Pak J Med Sci* 2023;39:605–7.

19 Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med* 2023;388:2400.