Chen *et al. BMC Medical Education* (2024) 24:1372
https://doi.org/10.1186/s12909-024-06309-x

BMC Medical Education

RESEARCH **Open Access**

Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study

Yikai Chen[1†], Xiujie Huang[1†], Fangjie Yang[1], Haiming Lin[2,4], Haoyu Lin[3], Zhuoqun Zheng[1], Qifeng Liang[1], Jinhai Zhang[1*] and Xinxin Li[1*]

Abstract

Background This study aimed to evaluate the performance of GPT-3.5, GPT-4, GPT-4o and Google Bard on the United States Medical Licensing Examination (USMLE), the Professional and Linguistic Assessments Board (PLAB), the Hong Kong Medical Licensing Examination (HKMLE) and the National Medical Licensing Examination (NMLE).

Methods This study was conducted in June 2023. Four LLMs (Large Language Models) (GPT-3.5, GPT-4, GPT-4o and Google Bard) were applied to four medical standardized tests (USMLE, PLAB, HKMLE and NMLE). All questions are multiple-choice questions and were sourced from the question banks of these examinations.

Results In USMLE step 1, step 2CK and Step 3, there are accuracy rates of 91.5%, 94.2% and 92.7% provided from GPT-4o, 93.2%, 95.0% and 92.0% provided from GPT-4, 65.6%, 71.6% and 68.5% provided from GPT-3.5, and 64.3%, 55.6%, 58.1% from Google Bard, respectively. In PLAB, HKMLE and NMLE, GPT-4o scored 93.3%, 91.7% and 84.9%, GPT-4 scored 86.7%, 89.6% and 69.8%, GPT-3.5 scored 80.0%, 68.1% and 60.4%, and Google Bard scored 54.2%, 71.7% and 61.3%. There was significant difference in the accuracy rates of four LLMs in the four medical licensing examinations.

Conclusion GPT-4o performed better in the medical licensing examinations than other three LLMs. The performance of the four models in the NMLE examination needs further improvement.

Clinical trial number Not applicable.

Keywords ChatGPT, Google bard, Large language models, Artificial intelligence, Medical licensing examination, Medical education

---

[†] Yikai Chen and Xiujie Huang contributed equally to this work. Xinxin Li and Jinhai Zhang are both corresponding authors.

*Correspondence:
Jinhai Zhang
stuzjh@163.com
Xinxin Li
xxli2023@163.com

[1] Department of Gastrointestinal Surgery, The First Affiliated Hospital of Shantou University Medical College, No. 57 Changping Road, Jinping District, Shantou, Guangdong 515000, China
[2] Department of Orthopaedics, The First Affiliated Hospital of Shantou University Medical College, Shantou, Guangdong 515000, China
[3] Department of Thyroid Breast Surgery, The First Affiliated Hospital of Shantou University Medical College, Shantou, Guangdong 515000, China
[4] School of Dentistry, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

The large language models (LLMs) developed by artificial intelligence (AI) has experienced explosive growth. LLMs will revolutionize medical education [1–3]. LLMs can help medical students memorize and understand a large amount of information, enabling them to get personalized and interactive learning experiences [2, 4]. LLMs can encourage self-directed learning and provide the opportunity to practice clinical decision-making for medical students [4]. LLMs could assist students in preparing for multiple-choice examinations and objective structured clinical examinations [5]. In China, a prominent advantage of LLMs may be that it can enable medical students in remote or resource-poor areas to receive better assistance.

The Chat Generative Pre-Trained Transformer (ChatGPT) was developed by OpenAI (San Francisco, California) and was released in November 2022. ChatGPT is the most extensively studied LLMs, including GPT-3.5, GPT-4 and GPT-4o versions [6]. In March 2023, Google Bard, which is the LLM developed by Google (Mountain View, California), was released [7]. Both ChatGPT and Google Bard have the ability to generate text, translate languages, write different kinds of creative content, and answer questions in an informative way. They have both been used in studies for medical examinations in various countries [8–12]. The performance of ChatGPT and Google Bard varies across different medical examinations. In English-speaking countries, ChatGPT has achieved exciting results on the United States Medical Licensing Examination (USMLE) and the UK Standardized Admission Tests [13, 14]. While in the National Medical Licensing Examination (NMLE) in China, chatGPT did not meet the passing criteria and performed worse than Chinese students [15, 16]. While GPT-4 has demonstrated excellent performance in most examinations [8–10], Bard has also shown certain advantages in some examinations [9, 12]. Although the latest version of GPT, GPT-4o, has only been released for a short period, research has already shown that it performs well in the European Interventional Radiology Examination and the Japanese Medical Licensing Examination(JMLE) [17–19]. Notably, GPT-4o demonstrates higher accuracy than GPT-4 specifically in JMLE [19].

Current investigations remain devoid of explorations regarding the performance of ChatGPT and Bard, particularly GPT-4o, in medical standardized tests of different cultures. There is also a lack of comparative studies on the application of different models in the Chinese NMLE. Consequently, in the pursuit of this research, we propose an evaluation of GPT-4's capabilities in clinical reasoning and medical education by scrutinizing its performance in standardized examinations such as the USMLE, the Professional and Linguistic Assessments Board (PLAB) test, the Hong Kong Medical Licensing Examination (HKMLE) and the NMLE. Among these exams, USMLE, PLAB, and HKMLE serve as benchmarks for global medical standards in English-speaking contexts, while the NMLE represents a standardized medical examination in Chinese-speaking regions. Collectively, these four standardized exams are highly recognized on an international level.

This study aims to evaluate the performance of four LLMs (GPT-3.5, GPT-4, GPT-4o and Bard) on four licensing examinations (USMLE, PLAB, HKMLE and NMLE) and assess their knowledge and potential applications for medical education.

Materials and methods

Study design

The cross-sectional study was conducted in June 2023. Four LLMs (GPT-3.5, GPT-4, GPT-4o and Google Bard) were applied to four medical standardized tests (USMLE, PLAB, HKMLE and NMLE) to compare the performance of the LLMs. As GPT-3.5 lacks image input capabilities, we excluded all questions containing images from its test set.

Data set

All questions were multiple-choice questions. The questions of USMLE, PLAB and HKMLE were written in English, while the questions of NMLE were written in Chinese. The sample questions were obtained from USMLE (www.usmle.org/prepare-your-exam), PLAB (https://www.gmc-uk.org/registration-and-licensing/join-the-register/plab), and HKMLE (https://leip.mchk.org.hk/EN/quiz.html). In detail, we acquired 592 USMLE sample questions, 30 PLAB part I sample questions, and 48 HKMLE part I sample questions. For USMLE, the update times for the Step 1, Step 2CK, and Step 3 question banks were June 2022, March 2023, and August 2022, respectively. The HKMLE question bank was last updated in March 2021. We also referred to the NMLE official sample question book published in 2023, extracting 10 questions from each topic (or all if less than 10), ultimately integrating into a collection of 139 questions to evaluate ChatGPT's performance in NMLE. We also categorized the test questions based on each official exam guideline to evaluate the performance of the four LLMs across different medical topics. Although the passing thresholds for the USMLE and PLAB vary slightly each year, they generally remain around the 60% level [20, 21]. Similarly, the NMLE's passing threshold is set at 60%. As for the HKMLE, its pass rate varies by the year of the paper, with an average pass rate of 32.8% over the past three years [22].

Prompt engineering

The input format of the dataset was standardized first, as prompt engineering has been proven to significantly affect the output of generative language models [23]. The input directives were established as follows: "Please answer the following medical question, choose the correct answer, and provide an explanation.". Notably, since HKMLE Part II requires answers to be written on a specific answer book and mainly evaluates the English proficiency of participants. Each candidate must pass Part I and Part II Examinations before applying for Part III. Moreover, Part III is an OSCE examination, so we chose only the first part's questions for our test. This was done to ensure the accuracy of ChatGPT's responses relied entirely on its ability to synthesize medical knowledge from descriptive texts and make accurate judgments, rather than parsing complex textual inputs [14]. Example of question prompt and answers is presented in Fig. 1.

No additional pre-training was conducted in the study. The researchers manually input all questions into GPT-4o, GPT-4, GPT-3.5, and Google Bard, collected their responses. Subsequently, the answers generated by the models were compared with the standard answers, and the corresponding scores were calculated.

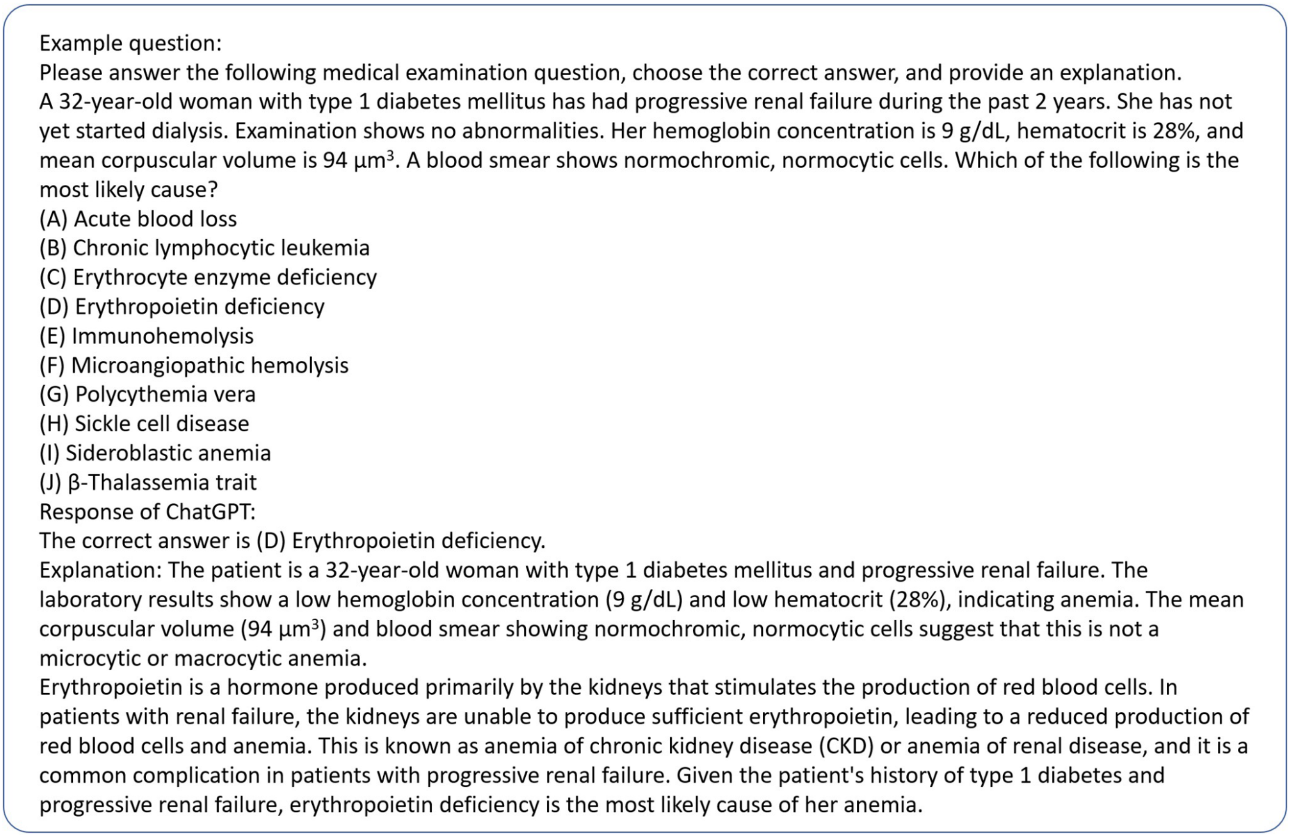
Statistical methods

Statistical analyses were conducted using the SPSS 21.0 (IBM Corp., Amronk, NewYork, United States). The responses generated by GPT-4o, GPT-4, GPT-3.5 and Bard were documented and corroborated, followed by categorization of the questions by subject. Categorical variables are presented as frequencies and percentages. Chi-square tests and Fisher exact tests were used for comparisons between groups. $P<0.05$ was considered statistically significant. For post-hoc tests following a Chi-square, the Bonferroni correction was applied for adjusting P-value.

Result

Performance of GPT-4o, GPT-4, GPT-3.5 and Google Bard on USMLE, PLAB, HKMLE and NMLE

Due to the limitations of GPT-3.5 in processing image inputs and the ethical review of questions by the four LLMs, some questions cannot be answered by these LLMs. GPT-4o, GPT-4, GPT-3.5, and Google Bard answered 592, 591, 542, and 516 questions from four examinations, respectively, with corresponding accuracy rates of 87.1%, 67.2%, and 60.8%. In USMLE step 1, step 2CK and Step 3, GPT-4o provided correct answers with accuracy rates of 91.5%, 94.2%, and 92.7%, which GPT-4 provided correct answers with accuracy rates of 93.2%, 95.0%, and 92.0%, while GPT-3.5 provided answers with accuracy rates of 65.6% 71.6% and 68.5%, and Google Bard provided answers with accuracy rates of 64.3%, 55.6%, and 58.1%,



Example question:
Please answer the following medical examination question, choose the correct answer, and provide an explanation.
A 32-year-old woman with type 1 diabetes mellitus has had progressive renal failure during the past 2 years. She has not yet started dialysis. Examination shows no abnormalities. Her hemoglobin concentration is 9 g/dL, hematocrit is 28%, and mean corpuscular volume is 94 μm^3. A blood smear shows normochromic, normocytic cells. Which of the following is the most likely cause?
(A) Acute blood loss
(B) Chronic lymphocytic leukemia
(C) Erythrocyte enzyme deficiency
(D) Erythropoietin deficiency
(E) Immunohemolysis
(F) Microangiopathic hemolysis
(G) Polycythemia vera
(H) Sickle cell disease
(I) Sideroblastic anemia
(J) β-Thalassemia trait
Response of ChatGPT:
The correct answer is (D) Erythropoietin deficiency.
Explanation: The patient is a 32-year-old woman with type 1 diabetes mellitus and progressive renal failure. The laboratory results show a low hemoglobin concentration (9 g/dL) and low hematocrit (28%), indicating anemia. The mean corpuscular volume (94 μm^3) and blood smear showing normochromic, normocytic cells suggest that this is not a microcytic or macrocytic anemia.
Erythropoietin is a hormone produced primarily by the kidneys that stimulates the production of red blood cells. In patients with renal failure, the kidneys are unable to produce sufficient erythropoietin, leading to a reduced production of red blood cells and anemia. This is known as anemia of chronic kidney disease (CKD) or anemia of renal disease, and it is a common complication in patients with progressive renal failure. Given the patient's history of type 1 diabetes and progressive renal failure, erythropoietin deficiency is the most likely cause of her anemia.

Fig. 1 Template of question posed to ChatGPT and the responses from ChatGPT

Table 1 Performance of GPT-4o, GPT-4, GPT-3.5 and Google Bard in USMLE, PLAB, HKMLE and NMLE.

		GPT-4o (n/N, %)	GPT-4 (n/N, %)	GPT-3.5 (n/N, %)	Google Bard (n/N, %)
Overall		538/592(90.9%)	515/591(87.1%)	364/542(67.2%)	314/516(60.9%)
USMLE	Step 1 (119)	108/118(91.5%)	109/117(93.2%)	61/93(65.6%)	73/114(64.3%)
	Step 2CK (120)	113/120(94.2%)	114/120(95.0%)	78/109(71.6%)	50/90(55.6%)
	Step 3 (137)	127/137(92.7%)	126/137(92.0%)	85/124(68.5%)	61/105(58.1%)
PLAB (30)		28/30(93.3%)	26/30(86.7%)	24/30(80.0%)	13/24(54.2%)
HKMLE (48)		44/48(91.7%)	43/48(89.6%)	32/47(68.1%)	33/46(71.7%)
NMLE (139)		118/139(84.9%)	97/139(69.8%)	84/139(60.4%)	84/137(61.3%)

USMLE=United State Medical Licensing Examination; PLAB=Professional and Linguistic Assessments Board; HKMLE=Hong Kong Medical Licensing Examination; NMLE=National Medical Licensing Examination

Table 2 Comparison of performance among GPT-4o, GPT-4, GPT-3.5 and Google Bard and Google Bard in USMLE, PLAB, HKMLE and NMLE

		P						
		4o,4.0, 3.5 & Bard	**4o vs. 4.0****	**4o vs. 3.5****	**4o vs. bard****	**4.0 vs. 3.5****	**4.0vsBard****	**3.5vsBard****
Overall		<0.001	0.040	<0.001	<0.001	<0.001	<0.001	0.027
USMLE	Step 1 (119)	<0.001	0.637	<0.001	<0.001	<0.001	<0.001	0.814
	Step 2CK (120)	<0.001	0.776	<0.001	<0.001	<0.001	<0.001	0.019
	Step 3 (137)	<0.001	0.820	<0.001	<0.001	<0.001	<0.001	0.101
PLAB (30)		<0.005	0.671*	0.254*	<0.001	0.488	0.008	0.042
HKMLE (48)		<0.005	1.000*	0.004	0.012	0.010	0.028	0.701
NMLE (139)		<0.001	0.003	<0.001	<0.001	0.102	0.102	1.000

Note * Fisher exact tests

** The significance of the P value was assessed at 0.0125 (0.05/4) after Bonferroni correction

4o=GPT-4o; 4.0=GPT-4; 3.5=GPT-3.5;bard=Google Bard

respectively. In the other three examinations, PLAB, HKMLE and NMLE, GPT-4 scored 86.7%, 89.6% and 69.8%, GPT-3.5 scored 80.0%, 68.1% and 60.4%, and Google Bard scored 54.2%, 71.7% and 61.3% (Table 1).

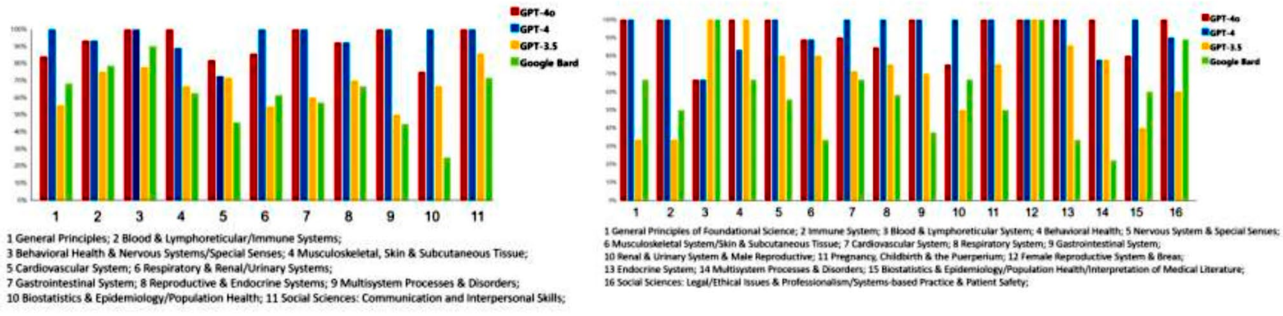
On the overall questions (all the questions in total) from USMLE, PLAB, HKMLE and NMLE, the accuracy rate for answering questions is ranked in descending order as follows: GPT-4o (90.9%), GPT-4 (87.1%), GPT-3.5 (67.2%), and Google Bard (60.9%). There was not significantly difference on the performance on the overall questions between GPT-4o and GPT-4 ($p=0.040$). The accuracy rate of GPT-4o and GPT-4 is significantly higher than that of GPT-3.5 and Google Bard (both $p<0.001$). In the meanwhile, there was not significantly difference on the performance on the overall questions between GPT-3.5 and Google Bard ($p=0.027$). The performance on USMLE, PLAB, HKMLE and NMLE significantly vary among four LLMs. Except NMLE ($p=0.003$), there was not significantly difference on the performance on the overall questions between GPT-4o and GPT-4 on USMLE step 1 ($p=0.637$), step 2CK ($p=0.776$), Step 3 ($p=0.820$), PLAB ($p=0.671$) and HKMLE ($p=1.000$). GPT-4o's score surpassed GPT-3.5 on USMLE step 1 ($p<0.001$), step 2CK ($p<0.001$), Step 3 ($p<0.001$) and HKMLE ($p=0.010$), and surpassed Google Bard on USMLE step 1($p<0.001$), step 2CK($p<0.001$), Step 3 ($p<0.001$), HKMLE (0.012) and PLAB($p<0.001$). GPT-4's score surpassed GPT-3.5 on USMLE step 1 ($p<0.001$), step 2CK ($p<0.001$), Step 3 ($p<0.001$) and HKMLE ($p=0.010$), and surpassed Google Bard on USMLE step 1($p<0.001$), step 2CK($p<0.001$), Step 3 ($p<0.001$) and PLAB($p=0.008$). On the contrary, there was not significantly difference between GPT-3.5 and Google Bard on USMLE, PLAB, HKMLE and NMLE. There was no significant difference in the accuracy rates between any two of GPT-4, GPT-3, and Google Bard in the NMLE ($p=0.102$, $p=0.102$, $p=1.000$) (Table 2).

Performance of GPT-4o, GPT-4 GPT-3.5 and Google Bard on USMLE, PLAB, HKMLE and NMLE, stratified by subject category

In the medical licensing examination, the performance of the four LLMs varied across different subjects. Specifically, in the USMLE, both GPT-4o and GPT-4 excelled by scoring over 80% in most subjects. However, their performance dipped in the categories of the hematolymphoid systems, and behavioral health, which was lower than 70%. As for PLAB and HKMLE, GPT-4o scored over 80% in all subjects. However, for PLAB and HKMLE, GPT-4's accuracy was notably lower in medical ethics and orthopedics questions, with accuracies of only 50% and 66.6% respectively. On the other hand, GPT-3.5 and Google Bard scored below 60% in several subjects for both the USMLE and PLAB. In USMLE Step 2 CK, GPT-3.5's accuracy was only 33.3% for questions related to the immune system and basic sciences. For questions involving multiple disciplines, Google Bard's accuracy stood at just 22.2%. Their performance in questions on doctor-patient communication and medical ethics also lagged significantly behind that of GPT-4o and GPT-4. As for NMLE, GPT-4o performed well in NMLE by scoring over 70% in all subjects except neurology scored 60.0%. Overall, these four LLMs faced challenges in the NMLE, with the accuracy of GPT-4, GPT-3.5, and Bard all failing to reach 70% accuracy. GPT-4o's accuracy also declined compared to its performance on the other three English-based exams. Notably, in neurology, the accuracy rates of GPT-4, GPT-3.5 and Google Bard was 30%, 30% and 40%, respectively. Additionally, in pediatrics and surgery, the rates did not surpass 60%. (Supplementary Table 1, Fig. 2).

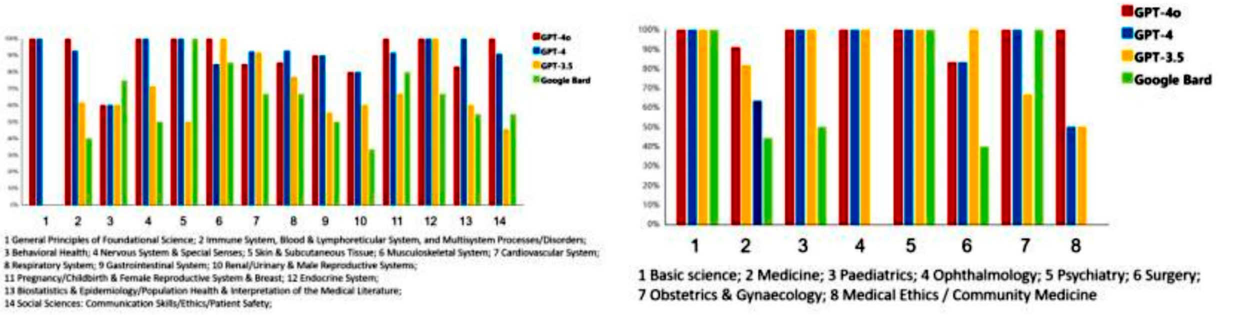
Performance of GPT-4o, GPT-4, and Google Bard on imaging-based questions

In the overall questions based on images from USMLE, the accuracy rates of GPT-4o, GPT-4, GPT-3.5 and Google Bard significantly varied from 44/49(89.8%), 87.5% (42/48), 12/24(50%), respectively. However, there was no significant difference found between GPT-4o and GPT-4, while both them outperformed Google Bard by 39.8% ($p < 0.001$) and 37.5% ($p < 0.001$). GPT-4o scored 80.0%



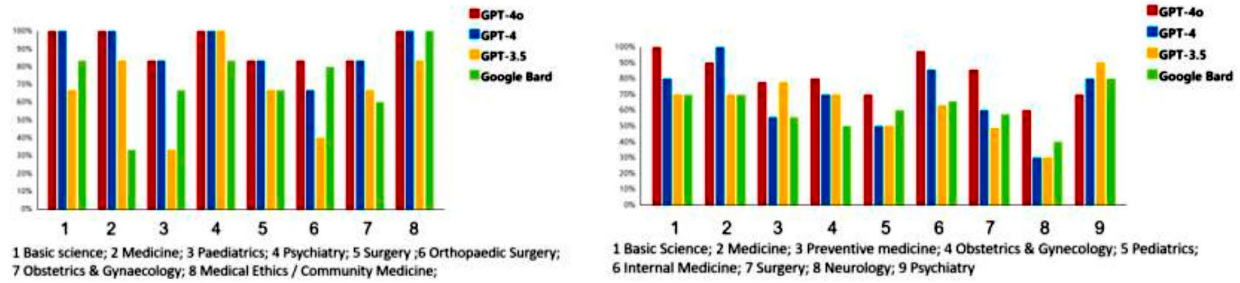
(a) USMLE Step 1

(b) USMLE Step 2CK



(c) USMLE Step 3

(d) PLAB



(e) HKMLE

(f) NMLE

Fig. 2 The performance of GPT-4o, GPT-4, GPT-3.5 and Google Bard in USMLE (**a**) Step 1, (**b**) Step 2, (**c**) Step 3, (**d**)PLAB, (**e**)HKMLE and (**f**) NMLE, stratified by subject category

(20/25), 100% (11/11) and 100% (13/13) in USLME Step 1, Step 2CK and Step 3. GPT-4 scored 83.3% (20/24), 100% (11/11) and 84.6% (11/13) in USLME Step 1, Step 2CK and Step 3. Google Bard scored 50% (11/22) in USLME Step 1 which was significantly lower than GPT-4, and 50% (1/2) in Step 3 (Table 3).

Discussion

The present study assessed the performance of GPT-3.5, GPT-4, GPT-4o and Google Bard in medical licensing examinations. Compared to GPT-3.5 and Google Bard, GPT-4o and GPT-4 exhibits superior capabilities in handling complex medical and clinical tasks. To the best of our knowledge, our study is among the few that have evaluated this aspect to date. Our findings can be divided into two main themes: [1] GPT-4o and GPT-4, in comparison to GPT-3.5 and Google Bard, has significantly improved in accuracy, exceeding the passing threshold of four standardized medical examinations; [2] The varied performance of the four LLMs across different cultures, disciplines, and subjects.

Following the release of GPT-3.5 by OpenAI, researchers quickly tested its performance in the USMLE, finding that GPT-3.5's accuracy exceeded 50% in most assessments, and in some analyses, even surpassed 60% [14, 20]. Upon GPT-4's introduction, subsequent studies indicated a notable increase in accuracy for USMLE [14]. At the same time, Google Bard also performed well in medical examinations [8, 24]. However, the situation differs significantly in the NMLE. Our research indicates that all four LLMs showed a decline in accuracy on the NMLE, with GPT-4, GPT-3.5, and Bard all failing to reach the 70% accuracy threshold. Even GPT-4o demonstrated lower accuracy compared to its performance on the other three English-based exams. This outcome could be attributed to multiple factors. Firstly, the training datasets of ChatGPT and Google Bard were mainly derived from English data, potentially restricting their effectiveness in non-English contexts. Previous studies highlighted similar situations [25], for example, ChatGPT's accuracy in the NMLE examination was 56%, but in translated versions, the accuracy was 76% [26]. In addition, previous studies have also discovered that ChatGPT scored 79.9% in the JMLE, another non-English medical examination, which is lower than the average candidate score of 84.9% [27]. Recent research also demonstrates that GPT-4o's accuracy on the JMLE is significantly higher than that of GPT-4, although it remains lower than its performance on the USMLE [19]. This finding indirectly supports the reliability of our study. Secondly, NMLE is structured as a multiple-choice examination, requiring the model to select the best answer from given options. This indicates LLMs still face challenges in identifying the most suitable answer. Furthermore, the NMLE's examination content is focuses more on China's healthcare situation, touching upon medical policies and legal contexts unique to China. For instance, there are significant legal differences between China and Western countries on issues such as abortion and euthanasia [28]. Such differences might result in biases in LLMs when tackling these issues, given their training predominantly on Western healthcare data and laws. Simultaneously, epidemiological data unique to China is a significant factor affecting LLMs' performance. Some diseases have markedly different incidences and prevalence between China and Western countries, with some data only available in Chinese. The potential inadequacy of exposure to such data during training might lead to subpar performance by LLMs on relevant matters. Lastly, there are differences in medical communication styles between China and Western countries. In the patient communication questions of NMLE, the three LLMs performed worse than in the physician examinations of other countries. This may be due to difficulties LLMs have in simulating communication between Chinese doctors and patients, failing to accurately understand and respond to patients' needs and emotions. In conclusion, factors such as language, culture, law, and data acquisition may all impact LLMs' performance in medical examinations within a Chinese context.

The four LLMs demonstrated varying accuracies in different medical licensing examinations. Overall, GPT-4o performed the best in these examinations, with its accuracy in all four medical licensing examinations surpassing that of the other three LLMs. It also outperformed GPT-4

Table 3 Performance of GPT-4o, GPT-4 and Google Bard in imaging-based questions in USMLE

		GPT-4o (*n*/*N*, %)	**GPT-4 (*n*/*N*, %)**	**Google Bard (*n*/*N*, %)**	***P***			
					4o, 4.0 & bard	**4o vs. 4.0****	**4o vs. bard****	**4.0 vs. bard****
Overall		44/49(89.8%)	42/48(87.5%)	12/24(50%)	<0.001	0.721432	<0.001	<0.001
USMLE	Step 1 (119)	20/25(80.0%)	20/24(83.3%)	11/22(50%)	0.022645	1.000*	0.030	0.016
	Step 2CK (120)	11/11(100%)	11/11(100%)	-	-	-	-	-
	Step 3 (137)	13/13(100%)	11/13(84.6%)	1/2(50%)	0.103175*	0.480*	0.133*	0.255

Note * Fisher exact tests

** The significance of the P value was assessed at 0.0167 (0.05/3) after Bonferroni correction

4o=GPT-4o; 4.0=GPT-4; 3.5=GPT-3.5;bard=Google Bard and Bard in accuracy on imaging-based questions. In the three steps of the USMLE, GPT-4o's and GPT-4's accuracy improved compared to before, and its accuracy in HKMLE was significantly higher than the examination's average passing rate of 32.8% over the past three years [29]. Furthermore, in the Peruvian medical licensing examination, GPT-4 achieved an accuracy of 86%, surpassing the average candidate score of 54% [10]. In JMLE, GPT-4o reached an accuracy rate of 89.2%, exceeding the average candidate score [19]. These achievements should be attributed to OpenAI's continuous optimization of the model, also demonstrating that GPT-4o and GPT-4 performs consistently in English-language medical licensing examinations. In contrast, Google Bard and GPT-3.5 exhibited similar performance levels across the four medical licensing examinations, consistent with previous research findings. Some studies have found that Bard even outperforms GPT-4 in certain specific examinations [24]. In contrast to ChatGPT, Bard exhibits considerable fluctuations in accuracy across standardized medical exams. Its accuracy in USMLE Step 1, USMLE Step 2, and PLAB exam remains below 60%. In the three English-language standardized medical exams, GPT-4o consistently achieved an accuracy rate above 90%, significantly surpassing the passing scores of all three exams. In the NMLE, GPT-4o also attained a score of 84.9%, notably outperforming GPT-4. These results can be attributed to OpenAI's ongoing model optimizations, which also highlight the stability of GPT-4o's performance in English-language medical licensing exams. Naturally, GPT-4o, similar to other LLMs, also faces the issue of "hallucination", meaning the generated text might include errors that appear semantically or grammatically logical but are in fact incorrect or nonsensical [30]. GPT-4o may occasionally exhibit flawed reasoning while still arriving at the correct final answer [31]. However, despite these imperfections, OpenAI's ongoing updates and refinements have enabled GPT-4o to be available for limited free use. Currently, GPT-4o demonstrates the most impressive performance in standardized medical exams, potentially making it the most suitable LLM for medical students and educators.

Our research additionally revealed notable variations in the accuracy rates of the four LLMs across various disciplines. In USMLE, GPT-4o and GPT-4 demonstrated higher accuracy in questions involving medical knowledge of a single system compared to the other two LLMs. However, when the questions integrated knowledge from multiple systems, the performance of the three LLMs other than GPT-4o was less than satisfactory, failing to provide satisfactory answers. This finding indicates that although LLMs excel in certain specific medical fields, they still face challenges in dealing with more complex, interdisciplinary medical scenarios, further confirming previous research results about limitations of LLMs in complex medical tasks [32–35]. Simultaneously, in tests of certain subjects, we found that the LLMs performed relatively poorly, with a higher error rate. GPT-4's accuracy in orthopedics in the HKMLE and neurology in the NMLE was only 66.7% and 30%, respectively. Previous studies have also demonstrated the limitations of LLMs when applied in specialized medical fields [36, 37]. This may be due to subspecialty examinations typically presenting more complex clinical scenarios, requiring higher levels of reasoning skills and the ability to synthesize information from multiple sources. Since LLMs are not specifically trained for the medical field, although they can understand context, they may struggle with these complex specific scenarios, possibly leading to insufficient accuracy due to a lack of data representativeness. With the continuous updates of these models, such as GPT-4o and future iterations, this challenge can gradually be overcome.

We also conducted an interesting experiment aimed at testing the self-correction capabilities of three LLMs. When faced with their incorrect responses, we attempted to correct them and provided the right answers, which they would change after our correction. However, when we initiated a new session and re-entered the same questions, these models still reverted to the original incorrect responses. This finding indicates that although the models are capable of self-correction within a single session, they are unable to retain this corrected knowledge across sessions. This limitation may stem from the fact that current LLMs largely depend on extensive data supplied by developers during training, without effectively integrating feedback from regular users into the training process. This means that although users can correct errors in a single session, these corrections fail to change the model's behavior in a broader context. It has also been noted that continuous user feedback and model improvements by developers are expected to improve the self-correcting ability of LLM chatbots over time [38].

Although challenges such as ethical and legal issues, bias risks, content accuracy, and public acceptance must still be addressed, numerous practical and observational studies support the utilization of LLMs in the medical field [39]. Moreover, it's undeniable that the emergence of GPT-4o and Google Bard has showcased different directions in AI development. GPT-4o represents continuous iterations on its original training to achieve higher performance levels. The internet connectivity feature of Google Bard enables AI to persistently learn current information, thereby furnishing users with superior responses. With the ongoing iteration and refinement of LLMs, their latest versions have progressively enhanced capabilities for processing complex text, images, video, and audio. Medical students can leverage LLMs to synthesize knowledge across diverse medical disciplines, enabling them to manage complex, interdisciplinary cases more effectively. Additionally, LLMs' interactive features allow for clinical scenario simulations, providing performance-based feedback that supports medical students in steadily developing their clinical competence. Besides medical education, LLMs like ChatGPT have shown extensive application potential in clinical practice and scientific research. Research shows that ChatGPT can bolster clinical decision support [40] and act as an auxiliary resource for patients [41, 42], while also serving as a powerful tool for scientific research [43]. We believe that with the continuous development of LLMs, they are poised to become indispensable support tools for medical scholars, practitioners, patients, and researchers, significantly contributing to the advancement of medical science and patient care.

Limitation

There are a few limitations to this study. Firstly, ChatGPT was initially trained on a corpus comprised of data available up to or before 2021. Although some of the sample questions we used are from 2022 and beyond, the examination syllabus has not undergone significant changes, thus not being affected by the timeliness of the ChatGPT training dataset. Secondly, the ethical constraints imposed on LLMs mean that certain questions may not be processed through their review process, thus hindering the acquisition of valid responses. Finally, we repeatedly attempted to correct the LLMs, but when starting a new session, these models still returned the original erroneous answers. This limitation may stem from the fact that current LLMs still largely depend on the extensive data provided by developers during training. However, these are the common problems faced by researchers. With the ongoing learning and iteration of LLMs, we are optimistic about the progressive enhancement of their AI capabilities, which is expected to enable researchers to conduct more comprehensive tests in the medical field.

Conclusion

In conclusion, our research results show that GPT-4o and GPT-4 demonstrated consistently strong performance across evaluations, while the results of GPT-3.5 and Google Bard were more variable and inconsistent, with Google Bard even failing to meet the passing threshold in certain examinations. However, the performance of the four models in the NMLE examination needs further improvement. GPT-4o and GPT-4 have a significant performance advantage over GPT 3.5 and Google Bard in the English-language medical licensing examinations. Taking into account the internet connectivity and better accessibility of GPT-4o, it may be more suitable for medical students after continuous improvement and development. Moreover, we acknowledge that while LLMs exhibit significant potential in medical examination preparation, they also encounter several challenges and limitations. Therefore, ongoing strategy exploration and evaluations are essential to enhance and confirm the effectiveness of these tools further. As technology continuously advances and models are further optimized, we anticipate more advanced LLM chatbots playing an increased role in medical education and examination preparation, offering more accurate and efficient learning support and decision-making assistance to medical students, residents, and experienced physicians.

Abbreviations

USMLE	United States Medical Licensing Examination
PLAB	Professional and Linguistic Assessments Board
HKMLE	Hong Kong Medical Licensing Examination
NMLE	National Medical Licensing Examination
JMLE	Japanese Medical Licensing Examination
LLMs	Large Language Models
ChatGPT	Chat Generative Pre-Trained Transformer

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-024-06309-x.

Supplementary Material 1

Author contributions

Conceptualization: X. L., J. Z., and Y. C.; methodology: F. Y., Y. C. and X. H.; Software, Y. C. and X. H.; validation, H. L., H. L., Z. Z and Q. L., formal analysis, F. Y., Y. C. and X. H.; writing—original draft preparation, F. Y. and Y. C., writing—review and editing, F. Y., Y. C., X. H., H. L., H. L., Z. Z., Q. L., J. Z. and X. L.; supervision, J. Z. and X. L. All authors have read and agreed to the published version of the manuscript.

Funding

Funding was provided by Guangdong Province "New medical" teaching committee teaching reform project (2023 NO.122); Shantou University Medical School undergraduate teaching quality and teaching reform project (2023 No.25); Guangdong Province clinical teaching base teaching reform project (2023 NO.32);Guangdong Province undergraduate higher education teaching reform project No. 583; Guangdong Province clinical teaching base teaching reform project No. 128 and Ministry of Education industry-university cooperative education project No.230805236050915.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Informed consent

Informed consent were not required.

Institutional review board

This study does not involve human or animal subjects and therefore did not require approval from the Institutional Ethical Board.

Competing interests

The authors declare no competing interests.

Received: 14 June 2024 / Accepted: 5 November 2024

Published online: 26 November 2024

References

1. Shaheen A, Azam F, Amir M. ChatGPT and the future of Medical Education: opportunities and challenges. J Coll Physicians Surg Pak. 2023;33(10):1207.
2. Feng S, Shen Y. ChatGPT and the future of Medical Education. Acad Med. 2023;98(8):867–8.
3. Extance A. ChatGPT has entered the classroom: how LLMs could transform education. Nature. 2023;623(7987):474–7.
4. Daungsupawong H, Wiwanitkit V. ChatGPT and the future of Medical Education: correspondence. J Coll Physicians Surg Pak. 2024;34(2):244–5.
5. Seetharaman R. Revolutionizing Medical Education: can ChatGPT boost subjective learning and expression? J Med Syst. 2023;47(1):61.
6. ChatGPT [Available from. https://openai.com/blog/chatgpt/
7. Bard. [https://bard.google.com
8. Cheong RCT, Pang KP, Unadkat S, McNeillis V, Williamson A, Joseph J, et al. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. Eur Arch Otorhinolaryngol. 2024;281(4):2137–43.
9. Ohta K, Ohta S. The performance of GPT-3.5, GPT-4, and Bard on the Japanese national dentist examination: a comparison study. Cureus. 2023;15(12):e50369.
10. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, Arteaga-Cisneros KF, Chalco XCB, Ordoñez MAB, et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical examination: a cross-sectional study. J Educ Eval Health Prof. 2023;20:30.
11. Thibaut G, Dabbagh A, Liverneaux P. Does Google's Bard Chatbot perform better than ChatGPT on the European hand surgery exam? Int Orthop. 2024;48(1):151–8.
12. Meo SA, Al-Khlaiwi T, AbuKhalaf AA, Meo AS, Klonoff DC. The Scientific Knowledge of Bard and ChatGPT in Endocrinology, Diabetes, and Diabetes Technology: Multiple-Choice Questions Examination-Based Performance. Journal of Diabetes Science and Technology. 2023. https://doi.org/10.1177/19322968231203987
13. Giannos P, Delardas O. Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. JMIR Med Educ. 2023;9:e47737.
14. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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.
15. Shang L, Xue M, Hou Y, Tang B. Can ChatGPT pass China's national medical licensing examination? Asian J Surg. 2023;46(12):6112–3.
16. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. J Med Syst. 2023;47(1):86.
17. Besler MS, Oleaga L, Junquero V, Merino C. Evaluating GPT-4o's performance in the official European board of radiology exam: a comprehensive assessment. Acad Radiol. 2024;31(11):4365–71.
18. Ebel S, Ehrengut C, Denecke T, Gossmann H, Beeskow AB. GPT-4o's competency in answering the simulated written European Board of Interventional Radiology exam compared to a medical student and experts in Germany and its ability to generate exam items on interventional radiology: a descriptive study. J Educ Eval Health Prof. 2024;21:21.
19. Liu M, Okuhara T, Dai Z, Huang W, Okada H, Furukawa E et al. Performance of Advanced Large Language Models (GPT-4o, GPT-4, Gemini 1.5 Pro, Claude 3 Opus) on Japanese Medical Licensing Examination: A Comparative Study. medRxiv. 2024:2024.07.09.24310129.
20. 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 Digit Health. 2023;2(2):e0000198.
21. Recent pass rates for PLAB 1 and PLAB 2. Accessed February 11. 2024. [https://www.gmc-uk.org/registration-and-licensing/join-the-register/plab/recent-pass-rates-for-plab-1-and-plab-2
22. LEIP - MCHK, Accessed. February 11, 2024. [https://leip.mchk.org.hk/EN/aexam_IIa.html
23. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023;35(7):1098–102.
24. Gan RK, Ogbodo JC, Wee YZ, Gan AZ, Gonzalez PA. Performance of Google bard and ChatGPT in mass casualty incidents triage. Am J Emerg Med. 2024;75:72–8.
25. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT Across different versions in Medical Licensing examinations Worldwide: systematic review and Meta-analysis. J Med Internet Res. 2024;26:e60807.
26. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inf. 2023;177:105173.
27. 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;9:e48002.
28. Lee YT, Kleinbach R, Peng HPC, Chen ZZ. Cross-cultural research on euthanasia and abortion. J Soc Issues. 1996;52(2):131–48.
29. HKMLE pass rate. [https://leip.mchk.org.hk/EN/aexam_IIa.html
30. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: implications in Scientific writing. Cureus. 2023;15(2):e35179.
31. Jin Q, Chen F, Zhou Y, Xu Z, Cheung JM, Chen R et al. Hidden Flaws Behind Expert-Level Accuracy of Multimodal GPT-4 Vision in Medicine. ArXiv. 2024.
32. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, et al. ChatGPT-Generated Differential diagnosis lists for Complex Case-Derived Clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inf. 2023;11:e48808.
33. Currie G, Barry K. ChatGPT in Nuclear Medicine Education. J Nucl Med Technol. 2023;51(3):247–54.
34. 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;47(1):33.
35. Kuang YR, Zou MX, Niu HQ, Zheng BY, Zhang TL, Zheng BW. ChatGPT encounters multiple opportunities and challenges in neurosurgery. Int J Surg. 2023;109(10):2886–91.
36. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and Orthopaedic Resident performance on Orthopaedic Assessment examinations. J Am Acad Orthop Surg. 2023;31(23):1173–9.
37. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. BMJ Neurol Open. 2023;5(1):e000451.
38. Pushpanathan K, Lim ZW, Er Yew SM, Chen DZ, Hui'En Lin HA, Lin Goh JH, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. iScience. 2023;26(11):108163.
39. Sallam M. ChatGPT Utility in Healthcare Education, Research, and practice: systematic review on the promising perspectives and valid concerns. Healthc (Basel). 2023;11(6).
40. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inf Assoc. 2023;30(7):1237–45.
41. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic surgery advice and counseling from Artificial Intelligence: a Rhinoplasty Consultation with ChatGPT. Aesthetic Plast Surg. 2023;47(5):1985–93.
42. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023;29(3):721–32.
43. Lahat A, Shachar E, Avidan B, Shatz Z, Glicksberg BS, Klang E. Evaluating the use of large language model in identifying top research questions in gastroenterology. Sci Rep. 2023;13(1):4164.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.