ORIGINAL ARTICLE

GPT-3.5-turbo versus GPT-4.0 for medical education

Hong Yu Ryan Fong[*], *Sin Kiu Hui*[*], *Sum Yuet Ng*[*], *Chun Hei Chow*, *Yuk Ming Lai*, *Lok Hang To*, *Ruyue Shen*, *Xiaoyan Hu*, *Carol Y Cheung*[†], *Dawei Gabriel Yang*[†]

Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China

[*]*Contributed equally*

[†]*Contributed equally*

Correspondence and reprint requests:

Dr Dawei Gabriel Yang, Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China. Email: gabrielyang@link.cuhk.edu.hk

Abstract

Objectives: To compare the performances of GPT-3.5-turbo and GPT-4.0 for medical education.

Methods: The performances of GPT-3.5-turbo (via Poe) and GPT-4.0 (via Microsoft 365 Copilot) were compared in terms of medical terminology translation (283 medical terms from six specialties), situational judgment (30 situations in seven contexts), medical knowledge (80 multiple choice questions and 80 clinical scenario questions), medical studies (10 case scenarios), and clinical communication (10 case scenarios).

Results: GPT-4.0 outperformed GPT-3.5-turbo in accuracy in terms of medical terminology translation (98.6% vs 91.5%), situational judgment (83.3% vs 63.3%), medical knowledge (93.8% vs 85.0% on multiple choice questions and 82.5% vs 72.5% on clinical scenario questions), medical studies (in 3 of 10 case scenarios), and clinical communication (in 4 of 10 case scenarios).

Conclusions: GPT-3.5-turbo and GPT-4.0 performed reasonably well on medical knowledge and medical studies, with GPT-4.0 performing slightly better in medical terminology translation, situational judgment, and clinical communication. These results are promising for the incorporation of large language models in medical education. Nonetheless, overreliance should be avoided as responses can be inaccurate or irrelevant.

Key words: Education, medical; Generative artificial intelligence; Large language models

Introduction

Generative artificial intelligence (AI) excels in producing diverse content such as text, images, and videos.[1,2] Large language models, a form of generative AI, are customized for text-based tasks such as natural language comprehension, text generation, and language translation. These models can potentially reshape healthcare and education by creating high-quality educational materials and enabling interactive learning with data-driven insights.[3-5]

ChatGPT by OpenAI has demonstrated its performance on the United States Medical Licensing Examination[6] and graduate-level business and law school examinations.[7,8] Compared with GPT3.5-turbo, GPT-4.0 has an enhanced contextual understanding and expanded knowledge base.[9] GPT's utilities for answering physicians' questions,[10] making diagnoses,[11,12] assisting triage,[13] and translating language[14,15] have been evaluated.

In Hong Kong, English is the medium of medical education, whereas Cantonese is the medium of communication between clinicians and patients. It is important to assess the accuracy of GPT in multilingual settings before its potential integration into medical education and clinical practice. This study aimed to compare the performances of GPT-3.5-turbo and GPT-4.0 for medical education.

Methods

The performances of GPT-3.5-turbo (via Poe) and GPT-4.0 (via Microsoft 365 Copilot) were compared in terms of medical terminology translation, situational judgment, medical knowledge, medical studies, and clinical communication.

For medical terminology translation, we used 283 terminologies from six specialties: ophthalmology (n=183), general medicine (n=20), general surgery (n=20), obstetrics and gynecology (n=20), pediatrics (n=20), and psychiatry (n=20) from the Medical Audio Glossary[16] of the Faculty of Medicine at the Chinese University of Hong Kong, using the prompt "Help me translate the following medical terms into Hong Kong Chinese."

For situational judgment, 30 questions in seven contexts from the University Clinical Aptitude Test[17] were used for ethical decision making. Each question had four options: A (very important/appropriate), B (important/appropriate), C (of minor importance/inappropriate but not awful), and D (not important/very inappropriate).

For medical knowledge, 80 multiple choice questions (MCQs) from *Robbins and Cotran Review of Pathology* (fourth edition)[18] and 80 clinical scenario questions from *Surgical Recall* (eighth edition)[19] were used.

For medical studies, five cases related to the role of AI in medicine and five cases related to textbook recommendations, note-taking tools, preparation for clinical rotations, and clinical skills were used.

For clinical communication, 10 cases from the Practical Assessment of Clinical Examination Skills from the Membership of the Royal Colleges of Physicians[20-22] of the United Kingdom were used. Information and scenarios involving a doctor and a patient were input. The information included the patients' chief complaints, investigation results, characters, moods, and concerns.

Six medical students independently performed the medical studies and clinical communication inquiries. Responses were evaluated based on the modified AI-DISCERN criteria[23,24] in terms of technical, clinical, and societal perspectives on a 5-point Likert scale ranging from 1 to 5. Total scores ranged from 10 to 50; higher scores indicated better performance.

The performances of GPT-3.5-turbo and GPT-4.0 were compared using the Mann-Whitney *U* test. Statistical analyses were performed using SPSS (Windows version 19.0; IBM Corp, Armonk [NY], USA). A p value of <0.05 was considered statistically significant.

Results

For medical terminology translation, the accuracy was 91.5% (259/283) [95% confidence interval (CI)=88.4%-94.9%] for GPT-3.5-turbo and 98.6% (279/283) [95% CI=97.0%-99.9%] for GPT-4.0. Most inaccuracies were due to a lack of context; for example, 'floaters' was translated as '漂浮物', 'tunnel surgery' as '隧道手術', and 'synechia' as '粘連' or '瞳孔粘連'. GPT-3.5-turbo falsely translated 'slit lamp' as '縫合鏡' and 'nephroblastoma (Wilm's tumor)' as '腎芽腫', whereas GPT-4.0 falsely translated 'scotomas' as '視野缺損' instead of '暗點／盲點' (**Table 1**).

For situational judgment, the accuracy was 63.3% (19/30) [95% CI=46.1%-80.5%] for GPT-3.5-turbo and 83.3% (25/30) [95% CI=70.0%-96.6%] for GPT-4.0. Among the inaccurate answers, GPT-3.5-turbo's responses had a one-level difference in nine questions and a two-level difference in two questions, whereas GPT-4.0's responses had a one-level difference in in five questions only.

For medical knowledge, the accuracy of the MCQs was 85.0% (68/80) [95% CI=77.2%-92.8%] for GPT-3.5-turbo and 93.8% (75/80) [95% CI=88.4%-99.0%] for GPT-4.0. GPT-4.0 outperformed GPT-3.5-turbo on eight questions, whereas GPT-3.5-turbo outperformed GPT-4.0 on one question. Additionally, the accuracy of the clinical scenario questions was 72.5% (58/80) [95% CI=62.7%-82.3%] for GPT-3.5-turbo and 82.5% (66/80) [95% CI=74.2%-90.8%] for GPT-4.0. GPT-4.0 outperformed GPT-3.5-turbo on 13 questions, whereas GPT-3.5-turbo outperformed GPT-4.0 on five questions. Notably, both chatbots made the same mistake by wrongly diagnosing inner thigh pain as meralgia paresthetica (which causes outer thigh pain).

For medical studies inquiry, in the five cases related to the role of AI in medicine, the mean AI-DISCERN scores were 47.3 for GPT-3.5-turbo and 48.2 for GPT-4.0 (**Table 2**). Both chatbots provided guidelines on identifying false information and advice on the proper use of chatbots. However, in cases related to clinical practice, specifically, "What are some items that a medical student must bring to the ward during their clinical rotations?", GPT-3.5-turbo accurately answered a stethoscope and white coat, whereas GPT-4.0 failed to do so. When asked to assist medical students in performing a physical examination to diagnose aortic regurgitation, GPT-4.0 was superior by providing more suggestive findings and additional indicative signs.

For clinical communication, both chatbots performed comparably in six of 10 cases (**Table 3**); for example, regarding initiating communication between doctors and patients on smoking cessation (case 1), and hydration and feeding to patients with severe dementia (case 2), both chatbots could explain patients' conditions and showed

Table 1. Translation errors made by GPT-3.5-turbo and/or GPT-4.0, together with their correct answers.

Medical terminology	GPT-3.5-turbo	GPT-4	Correct answer
Ophthalmology			
Canaliculitis	淚囊炎*	淚小管炎	淚小管炎
Chalazion (internal hordeolum)	麥粒腫†	霰粒腫	霰粒腫（瞼板囊腫）（眼瘡）
Cone	圓錐†	視錐細胞	視錐細胞
Cycloplegia	睫狀肌麻痹	瞳孔麻痹*	睫狀肌麻痹
Dacryostenosis	淚囊閉塞*	淚管狹窄	淚管狹窄
Electroretinography	電視網膜圖†	視網膜電圖	視網膜電圖
Epicanthus	內眥†	內眥贅皮	內眥贅皮
Epiphora	流淚†	溢淚	淚溢（淚水過多）
Epiretinal membrane	玻璃膜†	視網膜前膜	視網膜前膜
Episcleritis	結膜炎*	鞏膜炎	表層鞏膜炎
Floaters	漂浮物†	飛蚊症	飛蚊症
Hemorrhage, petechial	瘀斑†	瘀點出血	瘀點
Lid lag	眼瞼下垂†	眼瞼遲滯	眼蓋活動遲鈍
Limbus	眼瞳*	角膜緣	眼緣（角鞏膜交界）
Pilocarpine hydrochloride	益眼水†	鹽酸匹羅卡品†	縮瞳孔藥物（毛果芸香鹼）
Rod	桿細胞†	視桿細胞	視桿細胞
Slit lamp	縫合鏡†	裂隙燈	裂隙燈
Scotoma	盲點	視野缺損†	暗點／盲點
Supraorbital	額上*	眶上	眼眶上
Sympathetic ophthalmia	同情性眼炎†	交感性眼炎	交感性眼炎
Synechia	粘連†	瞳孔粘連*	虹膜粘連
General Medicine			
Dystrophy (dystrophia)	萎縮症†	肌肉營養不良†	增長不良
General Surgery			
Continent cutaneous urinary diversion	可控性皮膚尿道導管†	可控性皮膚尿流改道術	可控性皮膚尿流改道手術
Laparotomy	腹腔鏡手術†	剖腹手術	剖腹術
Tunnel surgery	隧道手術†	隧道手術†	腕管綜合症手術
Pediatrics			
Immotile cilia syndrome	無鞭毛症候群†	纖毛不動症候群	纖毛不動症候群
Nephroblastoma (Wilm's tumor)	腎芽腫†	腎母細胞瘤	腎胚胎癌／腎母細胞瘤

* *Anatomical translation error*

† *Literal translation error*

empathy and understanding. In four cases, GPT-4.0 outperformed GPT-3.5-turbo; for example, when asked to explain a patient's condition to his mother without exposing the patient's recreational drug use (case 5), GPT-3.5-turbo violated the patient's autonomy by disclosing this confidential information. Both chatbots fell short in understanding human emotions. When discussing needle phobia (case 4), neither chatbot displayed embarrassment. Moreover, vague terms instead of specific details were documented during doctor-patient interactions. For instance, both chatbots merely mentioned 'various treatment options' to patients with HIV (case 8) without further elaboration.

Discussion

GPT-4.0 outperformed GPT-3.5-turbo in medical terminology translation, situational judgment, and clinical communication, whereas the two chatbots were comparable in medical knowledge and medical studies. Both chatbots have shown high accuracy in translating medical English to

Table 2. GPT-3.5-turbo and GPT-4.0 responses to questions related to medical studies measured using the AI-DISCERN criteria.

Question	AI-DISCERN score*		p Value
	GPT-3.5-turbo	GPT-4	
Question related to the role of AI in medicine			
As a medical student, how should I use ChatGPT?	49.67±0.82	42.67±6.53	0.03
As a medical student, how can I identify false information generated by ChatGPT? Can you provide a stepwise approach for me please?	49.67±0.82	50.00±0.00	0.69
What is the correct attitude for medical students who use ChatGPT in their medical studies?	50.00±0.00	49.67±0.82	0.69
How should medical doctors integrate AI in their practice?	38.00±3.35	49.00±2.45	0.01
I am a medical student. The medical recommendations provided by ChatGPT and the doctor that I am attached to are different. What should I do?	49.33±1.03	49.67±0.82	0.69
Question related to medical practices			
Which book should first-year medical students read to prepare for their surgical ward attachment?	41.00±6.54	35.67±6.12	0.23
Which note-taking tool should a medical student use? Can you recommend one for me?	40.67±2.07	40.33±3.44	0.87
What are some items that a medical student must bring to the ward during their clinical rotations?	46.00±2.83	17.00±4.69	0.01
I am a medical student. How can I use Copilot to initiate Objective Structured Clinical Examination history-taking practice?	26.34±5.28	38.00±4.20	0.01
I am a medical student. How do I know that a patient has aortic regurgitation from a physical examination?	35.67±1.51	41.67±4.27	0.02

* *Data are presented as mean±standard deviation*

Table 3. GPT-3.5-turbo and GPT-4.0 responses to scenarios related to clinical communication measured using the AI-DISCERN criteria.

Case scenario	AI-DISCERN score*		p Value
	GPT-3.5-turbo	GPT-4	
Case 1: Cessation of smoking in a patient with type 2 diabetes mellitus and transient ischemic attack	33.33±4.32	34.00±1.26	0.13
Case 2: Discussing issues relating to hydration and feeding in a patient with severe dementia	36.67±4.32	40.33±2.94	0.13
Case 3: Explaining an error to a patient	38.30±2.34	40.00±2.53	0.30
Case 4: Explaining the importance of treatment to a patient with diabetes mellitus	28.67±1.63	31.33±1.63	0.04
Case 5: A relative requesting information that the patient has instructed must not be divulged	16.83±2.04	36.33±2.34	0.01
Case 6: Explaining the need for lifestyle changes to a patient with rheumatoid arthritis	40.67±5.16	40.67±3.72	0.81
Case 7: Explain the delay in diagnosis of multiple myeloma	37.30±3.50	42.33±3.44	0.04
Case 8: Explaining newly diagnosed HIV in a sexual partner	36.00±3.35	42.67±2.73	0.02
Case 9: Discussing valid complaints from a patient with unresolved acute pyelonephritis who wishes to discharge herself	36.00±2.53	39.33±2.42	0.07
Case 10: Discussing further management of a seriously ill patient	35.33±6.89	31.00±1.67	0.34

* *Data are presented as mean±standard deviation*

Spanish,[25] Japanese,[26] and Chinese.[27,28] For translation into Hong Kong Chinese, both chatbots had an overall accuracy of ≥90%, indicating their potential role in facilitating doctor-patient communication. Nevertheless, both chatbots failed to provide context-aware translation of some terminologies (eg, 'tunnel surgery' was translated as '隧道手術' rather than '腕管綜合症手術'), even though the prompt already clarified the request for medical terms. Incorrect translations might result from the chatbots' inability or imprecision to identify the exact anatomical locations of a health problem. For example, GPT-3.5-turbo falsely translated 'canaliculitis' as '淚囊炎' rather than '淚小管炎', whereas GPT-4.0 falsely translated 'synechia' as '瞳孔粘連' rather than '虹膜粘連'. Caution should be exercised when translating specialist terminology; reliable resources should be used for verification.

GPT-4.0 outperformed GPT-3.5-turbo in situational judgment, probably because of GPT-4.0's expanded knowledge base with enhanced contextual understanding.[10,29] Although GPT-3.5-turbo correctly answered only 19 of 30 questions, its responses provided justifications (eg, relevant

details and references) and were reasonably close to the correct answers. Both chatbots have sufficient knowledge to answer questions related to skills and techniques. However, GPT-3.5-turbo performed worse in scenarios requiring human rationale and feeling, whereas GPT-4.0's responses more closely resembled real-life healthcare providers by providing conflict resolution and addressing emotional concerns empathetically. Nonetheless, both chatbots generally performed well, answering most questions related to medical knowledge correctly, especially the MCQs. Critical thinking and profound reflection remain fundamental to mastering specialized knowledge and distinguishing fact from fiction.

Regarding medical knowledge and medical studies, both chatbots failed to provide the most relevant recommendations and suggestions in some cases. For instance, when asked which book first-year medical students should read to prepare for their surgical ward attachment, the chatbots recommended the *Oxford Handbook of Surgical Nursing*, which is better suited to nurses. Additionally, neither chatbot could recommend the popular note-taking tools such as Goodnotes and Notability, indicating the chatbots' lack of timeliness and relevancy.

Regarding clinical communication, GPT-4.0 was superior to GPT-3.5-turbo; however, both chatbots failed to ask for important information such as smoking status, which is crucial in Objective Structured Clinical Examinations for history taking. The current AI settings prevent the chatbots from appreciating and expressing non-verbal communication in the forms of body language, speech patterns, and facial expressions. Interaction with chatbots in doctor-patient communication can sometimes be ineffective, as the responses might sound indifferent and unempathetic due to their lack of cognition and intuition to provide human contact and emotional support.[30]

Medical students increasingly use GPT to acquire specialized knowledge[31,32] and prepare for examinations[33,34] and Objective Structured Clinical Examinations practices.[35,36] Use of GPTs in medical education needs further evaluation for their complexity, ethical and legal responsibilities, and impact on human lives. Previous studies have demonstrated the potential of GPTs to reshape medical education,[1-3] but none was conducted in multilingual settings. Our study found that both chatbots are reliable and can provide well-organized output with relevant details and references for verification. However, for more advanced interactions such as initiating conversation in role play, both chatbots fell short in addressing real-life non-verbal communication. In the real world, patients might exhibit atypical symptoms that are not included in textbook descriptions. Therefore, virtual exposure to diverse patient personalities should be supplemented with real clinical experience. It is imperative that medical students understand what scenarios chatbots can be a valid auxiliary tool and when overreliance should be avoided, as it might lead to misuse or adverse consequences.

There were several limitations to our study. Only six medical students were involved, which is not representative of all medical students. Nevertheless, all questions were selected based on necessity and were agreed upon unanimously. A sample size calculation was not performed due to a lack of consensus or formula to define the validity of chatbot responses. However, the sample size is comparable to that in other studies.[37-39] Evaluation guidelines were strictly followed, ensuring fair and justified interpretation and comparison.

Conclusion

GPT-3.5-turbo and GPT-4.0 performed reasonably well on medical knowledge and medical studies, with GPT-4.0 performing slightly better in medical terminology translation, situational judgment, and clinical communication. These results are promising for the incorporation of large language models in medical education. Nonetheless, vigilance should be maintained regarding the authenticity and accuracy of chatbot outputs. Future chatbots might be trained on broader sources for more context-relevant outputs and more advanced logic and empathy for real-world situations.

Contributors

All authors designed the study, acquired the data, analyzed the data, drafted the manuscript, and critically revised the manuscript for important intellectual content. All authors had full access to the data, contributed to the study, approved the final version for publication, and take responsibility for its accuracy and integrity.

Conflicts of interest

All authors have disclosed no conflicts of interest.

Funding/support

This study received financial support from CUHK Direct Grant (reference: 4054419 & 4054487).

Data availability

All data analyzed in this study are available from the corresponding author upon reasonable request.

References

1. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023; 9:e46885. Crossref
2. Toma A, Senkaiahliyan S, Lawler PR, Rubin B, Wang B. Generative AI could revolutionize health care – but not if control is ceded to big tech. Nature 2023; 624:36-8. Crossref
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023; 29:1930-40. Crossref
4. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023; 3:141. Crossref
5. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. Ann Intern Med 2024; 177:210-20. Crossref
6. 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. Crossref
7. Choi JH, Hickman KE, Monahan AB, Schwarcz D. ChatGPT goes to law school. J Legal Educ 2022; 71:387. Crossref
8. Terwiesch C. Would ChatGPT3 get a Wharton MBA. A prediction based on its performance in the operations management course. Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania; 2023.
9. Kozachek D. Proceedings of the 15th Conference on Creativity and Cognition. Association for Computing Machinery; 2023.
10. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023; 6:e2336483. Crossref
11. Delsoz M, Madadi Y, Munir WM, et al. Performance of ChatGPT in diagnosis of corneal eye diseases. medRxiv 2023; 2023.08.25.23294635. Crossref
12. Hu X, Ran AR, Nguyen TX, et al. What can GPT-4 do for diagnosing rare eye diseases? A pilot study. Ophthalmol Ther 2023; 12:3395-402. Crossref
13. Waisberg E, Ong J, Zaman N, et al. GPT-4 for triaging ophthalmic symptoms. Eye (Lond) 2023; 37:3874-5. Crossref
14. Grimm DR, Lee YJ, Hu K, et al. The utility of ChatGPT as a generative medical translator. Eur Arch Otorhinolaryngol 2024; 281:6161-5. Crossref
15. Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023; 6:9. Crossref
16. Medical Audio Glossary. The Chinese University of Hong Kong. Accessed 12 July 2024. Available from: https://intranet.mfpo.cuhk.edu.hk/med/eweb/Terminology/med-term.html
17. Kulkarni S, Parry J, Sitch A. An assessment of the impact of formal preparation activities on performance in the University Clinical Aptitude Test (UCAT): a national study. BMC Med Educ 2022; 22:747. Crossref
18. Klatt EC, Kumar V. Robbins and Cotran Review of Pathology. Elsevier Health Sciences; 2014.
19. Blackbourne L. Surgical Recall. Wolters Kluwer Health; 2020.
20. Henry ES, Quinton N. How to prepare for the Membership of the Royal College of Physicians (UK) Part 2 Clinical Examination (Practical Assessment of Clinical Examination Skills): an exploratory study. Educ Med J 2021; 13:77-91. Crossref
21. Membership of the Royal Colleges of Physicians of the United Kingdom. PACES Station 4: Communication Skills & Ethics. Accessed 12 July 2024. Available from: https://www.thefederation.uk/sites/default/files/documents/Station%204%20Scenario%20Pack%20%2824%29.pdf
22. Membership of the Royal Colleges of Physicians of the United Kingdom. PACES Marksheets Sample 2023. Accessed 12 July 2024. Available from: https://www.thefederation.uk/sites/default/files/PACES%20New%20Marksheets%20-%20for%20website.pdf
23. Hua HU, Kaakour AH, Rachitskaya A, Srivastava S, Sharma S, Mammo DA. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. JAMA Ophthalmol 2023; 141:819-24. Crossref
24. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999; 53:105-11. Crossref
25. Garcia Valencia OA, Thongprayoon C, Jadlowiec CC, et al. AI-driven translations for kidney transplant equity in Hispanic populations. Sci Rep 2024; 14:8511. Crossref
26. Ando K, Sato M, Wakatsuki S, et al. A comparative study of English and Japanese ChatGPT responses to anaesthesia-related medical questions. BJA Open 2024; 10:100296. Crossref
27. Fang C, Wu Y, Fu W, et al. How does ChatGPT-4 perform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health 2023; 2:e0000397. Crossref
28. 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 Inform 2023; 177:105173. Crossref
29. Ballard DH. Inconsistently accurate: repeatability of GPT-3.5 and GPT-4 in answering radiology board-style multiple choice questions. Radiology 2024; 311:e241173. Crossref
30. Silverman J, Kinnersley P. Doctors' non-verbal behaviour in consultations: look at the patient before you look at the computer. Br J Gen Pract 2010; 60:76-8. Crossref
31. Ayoub NF, Lee YJ, Grimm D, Divi V. Head-to-head comparison of ChatGPT versus Google search for medical knowledge acquisition. Otolaryngol Head Neck Surg 2024; 170:1484-91. Crossref
32. Jo E, Song S, Kim JH, et al. Assessing GPT-4's performance in delivering medical advice: comparative analysis with human experts. JMIR Med Educ 2024; 10:e51282. Crossref
33. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023; 9:e45312. Crossref
34. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. Medicine (Baltimore) 2023; 102:e34673. Crossref
35. Ramchandani R, Biglou SG, Gupta M, Guo E. Using AI to revolutionize clinical training through OSCE-GPT: a focused exploration of user feedback on otolaryngology and neurology cases. Can J Neurol Sci 2024; 51:S35. Crossref
36. Misra SM, Suresh S. Artificial intelligence and Objective Structured Clinical Examinations: using ChatGPT to revolutionize clinical skills assessment in medical education. J Med Educ Curric Dev 2024; 11: 23821205241263475. Crossref
37. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023; 93:1090-8. Crossref
38. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep 2023; 13:20512. Crossref
39. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023; 3:100324. Crossref