Journal of Dental Sciences 20 (2025) 1709–1715

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Original Article

Performance of ChatGPT in answering the oral pathology questions of various types or subjects from Taiwan National Dental Licensing Examinations

Yu-Hsueh Wu [a,b], Kai-Yun Tso [a,b,c], Chun-Pin Chiang [d,e,f,g*]

[a] Department of Stomatology, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[b] Institute of Oral Medicine, School of Dentistry, National Cheng Kung University, Tainan, Taiwan
[c] Division of Endodontics, Department of Stomatology, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[d] Department of Dentistry, Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Hualien, Taiwan
[e] Institute of Oral Medicine and Materials, College of Medicine, Tzu Chi University, Hualien, Taiwan
[f] Graduate Institute of Oral Biology, School of Dentistry, National Taiwan University, Taipei, Taiwan
[g] Department of Dentistry, National Taiwan University Hospital, College of Medicine, National Taiwan University, Taipei, Taiwan

Received 22 March 2025; Final revision received 24 March 2025
Available online 5 April 2025

KEYWORDS
ChatGPT;
Artificial intelligence;
Oral pathology;
Dental education

Abstract *Background/purpose:* ChatGPT, a large language model, can provide an instant and personalized solution in a conversational format. Our study aimed to assess the potential application of ChatGPT-4, ChatGPT-4o without a prompt (ChatGPT-4o-P$^-$), and ChatGPT-4o with a prompt (ChatGPT-4o-P$^+$) in helping dental students to study oral pathology (OP) by evaluating their performance in answering the OP multiple choice questions (MCQs) of various types or subjects.

Materials and methods: A total of 280 OP MCQs were collected from Taiwan National Dental Licensing Examinations. The chatbots of ChatGPT-4, ChatGPT-4o-P$^-$, and ChatGPT-4o-P$^+$ were instructed to answer the OP MCQs of various types and subjects.

Results: ChatGPT-4o-P$^+$ achieved the highest overall accuracy rate (AR) of 90.0 %, slightly outperforming ChatGPT-4o-P$^-$ (88.6 % AR) and significantly exceeding ChatGPT-4 (79.6 % AR, $P < 0.001$). There was a significant difference in the AR of odd-one-out questions between ChatGPT-4 (77.2 % AR) and ChatGPT-4o-P$^-$ (91.3 % AR, $P = 0.015$) or ChatGPT-4o-P$^+$ (92.4 %

* Corresponding author. Department of Dentistry, Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, and Institute of Oral Medicine and Materials, College of Medicine, Tzu Chi University, No. 707, Section 3, Chung-Yang Road, Hualien, 970, Taiwan.
E-mail address: cpchiang@ntu.edu.tw (C.-P. Chiang).

https://doi.org/10.1016/j.jds.2025.03.030

AR, $P = 0.008$). However, there was no significant difference in the AR among three different models when answering the image-based and case-based questions. Of the 11 different OP subjects of single-disease, all three different models achieved a 100 % AR in three subjects; ChatGPT-4o-P^+ outperformed ChatGPT-4 and ChatGPT-4o-P^- in other 3 subjects; ChatGPT-4o-P^- was superior to ChatGPT-4 and ChatGPT-4o-P^+ in another 3 subjects; and ChatGPT-4o-P^- and ChatGPT-4o-P^+ had equal performance and both were better than ChatGPT-4 in the rest of two subjects.

Conclusion: In overall evaluation, ChatGPT-4o-P^+ has better performance than ChatGPT-4o-P^- and ChatGPT-4 in answering the OP MCQs.

Introduction

Artificial intelligence (AI) has gained a great attention in the past decade. One of the prominent applications of AI is natural language processing (NLP) algorithm, enabling computers to understand and generate natural language.[1,2] GPT, abbreviation for generative pre-trained transformers, is an NLP. The development of GPT can be traced back to 2018 by OpenAI, using a pre-trained large language model (LLM), which is pre-trained with a large-scale of texts including books, articles, and website texts and fine-tuned for specifying their own conversational task.[3] ChatGPT-4 (OpenAI Global, San Francisco, CA, USA, released on March 14, 2023), based on ChatGPT-3.5 LLM, which was first released as a public version in late November 2022. It is famous for its ability to handle language comprehension and generation tasks in a conversational format and soon holds worldwide attention. In 2024, ChatGPT-4o ("o" as "omni"; OpenAI Global, San Francisco, CA, USA, released on May 13, 2024) was released as the latest version with a strong update to provide further functions in the wake of ChatGPT-4. The users can input texts, voice, and visual images, thus more challenging multimodal tasks can be completed.[1,3]

There are a variety of applications of ChatGPT, one of which is serving as an interactive tool in the medical education. ChatGPT has been adopted in answering questions within the scope of the United States Medical Licensing Examination (USMLE) to check its performance, and the results showed accuracies of 42 %—64.4 %, indicating the potential of ChatGPT in supporting learning in the medical education.[4] Similarly, another study is to investigate the ChatGPT's performance on multiple choice questions (MCQs) of basic and clinical medical sciences, and it scored 74 % and 70 %, respectively.[5] Nevertheless, the study to assess ChatGPT-4's performance on the Japanese National Dental Examination (JNDE) revealed the correct response rates of above 70 % in the dental specialties, such as dental anesthesiology and endodontics, but a low correct response rate for the image-based questions (35.0 %) and for clinical practical questions (28.6 %).[6] Combined with another associated study that evaluated the image recognition capabilities of ChatGPT-4 and Gemini Pro (released by Google on December 13, 2023 through Google Cloud and AI Studio), their results might disclose the possible weakness of ChatGPT in handling image-intensive and complex clinical practical questions.[6,7]

In addition to the novelty, ChatGPT can provide an instant and personalized solution, rather than a long list of websites via traditional search engines. Thus, our study aimed to assess the potential application of ChatGPT-4, ChatGPT-4o without the prompt (so-called ChatGPT-4o-P^- in this study), and ChatGPT-4o with the prompt (so-called ChatGPT-4o-P^+ in this study) in helping the dental students to study oral pathology (OP) by evaluating their performances in answering the OP MCQs of various types or subjects.

Materials and methods

Collection of questions for dataset

The students graduated from the dental schools in Taiwan must pass the Taiwan National Dental Licensing Examination held by the Ministry of Examination to obtain a dentist license before they can practice. The Taiwan National Dental Licensing Examination consisted of two parts and was held twice a year.

The part I examination (mainly basic dental sciences) comprised Dentistry I examination (including basic dental specialties of oral anatomy, dental morphology, oral histology and embryology, biochemistry, and their relevant clinical knowledge) and Dentistry II examination (including basic dental specialties of OP, dental materials, oral microbiology, dental pharmacology, and their relevant clinical knowledge).

The part II examination (mainly clinical dental sciences) consisted of Dentistry III examination (including endodontics, operative dentistry, periodontology, and their relevant clinical cases and medical ethics), Dentistry IV examination (including oral and maxillofacial surgery, dental radiology, and their relevant clinical cases and medical ethics), Dentistry V examination (complete denture prosthodontics, removable partial prosthodontics, crown and bridge, occlusion, and their relevant clinical cases and medical ethics), and Dentistry VI examination (including orthodontics, pediatric dentistry, dental public health, and their relevant clinical cases and medical ethics). Each of the 6 Dentistry examinations has 80 MCQs, and each examination is worth 100 points, with 60 points being the passing score. There were 28 OP MCQs in each Dentistry II examination.

Processing the test questions

The test questions and their official correct answers were downloaded from the website of the Ministry of Examination from 2014 to 2024 as a PDF file. The Taiwan National Dental Licensing Examinations of 2015, 2016, 2018, 2021 and 2024 were randomly selected. They were then converted into editable text document files (Microsoft Word) to extract the OP questions and to rearrange the order of the test OP questions. There were 56 OP MCQs per year, and thus a total of 280 OP MCQs were collected and used as the test OP questions. The test OP questions were written in Chinese with the key medical terms written in English additionally. Every OP MCQ had four answer options and only one was correct.

Study design

All the OP questions were input in one day. A new chat was started for submitting the OP questions from another year. To assess the chatbot's performance under different conditions, different models (ChatGPT-4 and ChatGPT-4o) and the same model (ChatGPT-4o) with or without a certain prompt were applied for testing. The chatbot was instructed to answer the OP MCQs input. To further evaluate whether there was any difference in answering the OP MCQs between ChatGPT-4o-P^- and ChatGPT-4o-P^+, a prompt was added for ChatGPT-4o-P^+ to command the chatbot to play the role of a dental student and to answer the OP MCQs referring to two OP textbooks (*Oral Pathology: Clinical Pathologic Correlations* and *Oral and Maxillofacial Pathology*).[8,9]

Evaluation of the chatbot's performance

The answers generated by the chatbots of ChatGPT-4, ChatGPT-4o-P^-, and ChatGPT-4o-P^+ were recorded and classified as correct or incorrect using the standard answers provided by the website of the Ministry of Examination. Accuracy rates (ARs) were calculated as the proportion of the number of questions with correct answers to the total number of questions. For differentiating the three different chatbots' performance on a certain type of question, 170 selected OP MCQs were further divided into image-based questions ($n = 39$), case-based questions ($n = 39$), and odd-one-out questions ($n = 92$), respectively. Image-based questions were those questions with images in addition to the pure text. Case-based questions were those questions with a particular clinical situation, which could effectively test the application of relevant knowledge to diagnose a dental or medical disease.[10] Odd-one-out questions were characterized with an opposite stem orientation; that is, the correct answer of the question was the one that was false or the one with the lowest possibility.[11]

Regarding the OP subjects covered by the questions, the questions were first sorted into two groups by the involvement of single or multiple diseases. Subsequently, those questions involved by single disease were further subdivided into 11 OP subjects with different particular diseases. The subject 1 included those miscellaneous diseases, such as allergic reactions, drug-induced reactions, non-pathological oral conditions, sarcoidosis, foreign body reactions, osteoradionecrosis, medication-related osteonecrosis of the jaw, metallic intoxication, hematopoietic diseases, soft tissue tumors, and metastatic or relatively rare cancers. The subject 2 included developmental defects and oral manifestations of certain systemic diseases. The subject 3 included salivary gland pathology, comprising both neoplastic and non-neoplastic diseases. The subject 4 included infectious diseases, consisting of the viral, fungal, and bacterial infections. The subject 5 included odontogenic cysts and tumors. The subject 6 included pigmented lesions, from local lesions to systemic diseases. The subject 7 included bone pathology, mainly the fibro-osseous lesions. The subject 8 included oral cancers and precancers. The subject 9 included various abnormalities of teeth. The subject 10 included recurrent aphthous stomatitis and Behçet's disease. The subject 11 included different vesiculobullous diseases and oral lichen planus.

Statistical analysis

The comparisons of the differences in ARs of the OP MCQs of three different types (39 image-based questions, 39 case-based questions, and 92 odd-one-out questions) or of different OP subjects in two groups of single and multiple diseases among three different models (ChatGPT-4, ChatGPT-4o-P^-, and ChatGPT-4o-P^+) or between any two of the three different models were performed by the chi-square test. The significance level was set at $P < 0.05$ in all tests.

Results

The number of the correct answers and the ARs generated by the ChatGPT-4, ChatGPT-4o-P^-, and ChatGPT-4o-P^+ for the 28 OP MCQs per Dentistry II examination from 2015 to 2024 are recorded, calculated, compared, and shown in Table 1. The AR displayed as a percentage was the proportion of the number of questions with correct answers to the total number of questions. There was a statistically significant difference in the AR between ChatGPT-4 and ChatGPT-4o-P^- ($P = 0.006$). Although no significant difference in the AR was found between ChatGPT-4o-P^+ and ChatGPT-4o-P^-, ChatGPT-4o-P^+ achieved the highest overall AR of 90.0 %, slightly outperforming ChatGPT-4o-P^- (88.6 % AR, $P = 0.682$) and significantly exceeding ChatGPT-4 (79.6 % AR, $P < 0.001$).

A total of 170 OP MCQs were selected and classified into three different types: the image-based questions ($n = 39$), case-based questions ($n = 39$), and odd-one-out questions ($n = 92$). The numbers of questions with correct and incorrect answers generated by the ChatGPT-4, ChatGPT-4o-P^-, and ChatGPT-4o-P^+ for the three different types of OP questions were counted and compared (Table 2). For the 92 odd-one-out questions, either ChatGPT-4o-P^- or ChatGPT-4o-P^+ achieved a higher AR (91.3 % or 92.4 %, respectively) than ChatGPT-4 (77.2 %) ($P = 0.015$ or $P = 0.008$, respectively). However, for the 39 image-based questions and 39 case-based questions, no significant difference in the AR was discovered between ChatGPT-4 and ChatGPT-4o-P^- or between ChatGPT-4 and ChatGPT-4o-P^+ (all four P-values >0.05). In addition, for all three different

Table 1 The number of correct answers and the accuracy rates (ARs) generated by ChatGPT-4, ChatGPT-4o without the prompt (ChatGPT-4o-P$^-$), and ChatGPT-4o with the prompt (ChatGPT-4o-P$^+$) for 28 oral pathology multiple choice questions per Dentistry II examination from 2015 to 2024.

Year (number of question)	Number of questions with correct answers (%)		
	ChatGPT-4	ChatGPT-4o-P$^-$	ChatGPT-4o-P$^+$
2015-1 (n = 28)	23 (82.1)	27 (96.4)	26 (92.9)
2015-2 (n = 28)	19 (67.9)	26 (92.9)	27 (96.4)
2016-1 (n = 28)	25 (89.3)	26 (92.9)	26 (92.9)
2016-2 (n = 28)	26 (92.9)	23 (82.1)	24 (85.7)
2018-1 (n = 28)	25 (89.3)	26 (92.9)	27 (96.4)
2018-2 (n = 28)	24 (85.7)	24 (85.7)	26 (92.9)
2021-1 (n = 28)	17 (60.7)	24 (85.7)	25 (89.3)
2021-2 (n = 28)	24 (85.7)	27 (96.4)	28 (100.0)
2024-1 (n = 28)	17 (60.7)	22 (78.6)	21 (75.0)
2024-2 (n = 28)	23 (82.1)	23 (82.1)	22 (78.6)
Total (n = 280)	223 (79.6)	248 (88.6)	252 (90.0)
		[a]P = 0.006	[a]P < 0.001
			[b]P = 0.682

[a] Comparison of overall AR between ChatGPT-4 and ChatGPT-4o-P$^-$ or between ChatGPT-4 and ChatGPT-4o-P$^+$ by chi-square test.
[b] Comparison of overall AR between ChatGPT-4o-P$^-$ and ChatGPT-4o-P$^+$ by chi-square test.

Table 2 The number of questions with correct answers and incorrect answers generated by ChatGPT-4, ChatGPT-4o without the prompt (ChatGPT-4o-P$^-$), and ChatGPT-4o with the prompt (ChatGPT-4o-P$^+$) for the three different types of oral pathology questions.

	Image-based questions (n = 39)		Case-based questions (n = 39)		Odd-one-out questions (n = 92)	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
ChatGPT-4	28 (71.8)	11 (28.2)	30 (76.9)	9 (23.1)	71 (77.2)	21 (22.8)
ChatGPT-4o-P$^-$	31 (79.5)	8 (20.5)	33 (84.6)	6 (15.4)	84 (91.3)	8 (8.7)
[a]P		0.598		0.566		0.015
ChatGPT-4o-P$^+$	33 (84.6)	6 (15.4)	33 (84.6)	6 (15.4)	85 (92.4)	7 (7.6)
[a]P		0.273		0.566		0.008
[b]P		0.768		>0.999		>0.999

[a] Comparison of the AR of image-based, case-based, or odd-one-out questions between ChatGPT-4 and ChatGPT-4o-P$^-$ or between ChatGPT-4 and ChatGPT-4o-P$^+$ by chi-square test.
[b] Comparison of the AR of image-based, case-based, or odd-one-out questions between ChatGPT-4o-P$^-$ and ChatGPT-4o-P$^+$ by chi-square test.

types of questions, there was also no significant difference in the AR between ChatGPT-4o-P$^-$ and ChatGPT-4o-P$^+$ (all three P-values >0.05).

The 280 OP MCQs were first sorted into two groups by the involvement of multiple diseases (n = 37) or single disease (n = 243). The 243 OP MCQs of single disease were further subdivided into 11 different OP subjects. The numbers of questions with correct and incorrect answers generated by the ChatGPT-4, ChatGPT-4o-P$^-$, and ChatGPT-4o-P$^+$ for 243 OP MCQs of 11 OP subjects of single disease were recorded and compared (Table 3). For the multiple diseases or the single disease group, there was no significant difference in the AR between any two of the three different models (ChatGPT-4, ChatGPT-4o-P$^-$, and ChatGPT-4o-P$^+$) (all six P-values >0.05). For the 11 OP subjects of single-disease, statistical analysis was limited due to some small samples; thus, the results were interpreted by descriptive statistics and illustrated in Table 3 and Fig. 1. For subject 4 (infectious diseases), subject 10 (recurrent aphthous stomatitis and Behçet's disease), and subject 11 (vesiculobullous diseases and oral lichen planus), all three different models achieved an AR of 100 %. For the subjects 1, 2, and 7, ChatGPT-4o-P$^+$ outperformed ChatGPT-4 and ChatGPT-4o-P$^-$. For obtaining a correct response on the OP MCQs of subjects 3, 5, and 9, ChatGPT-4o-P$^-$ was superior to ChatGPT-4 and ChatGPT-4o-P$^+$. Moreover, for the rest of the two subjects 6 and 8, ChatGPT-4o-P$^-$ and ChatGPT-4o-P$^+$ had equal performance, but both had better performance in answering the OP MCQs correctly than ChatGPT-4 (Table 3 and Fig. 1).

Discussion

This study assessed the academic performance of three different models (ChatGPT-4, ChatGPT-4o-P$^-$, and ChatGPT-4o-P$^+$) in correctly answering the OP MCQs of three

Table 3 The number of questions with correct answers and incorrect answers generated by ChatGPT-4, ChatGPT-4o without the prompt (ChatGPT-4o-P$^-$), and ChatGPT-4o with the prompt (ChatGPT-4o-P$^+$) for 280 oral pathology questions of two different disease groups (multiple diseases and single disease) and for 243 questions of 11 oral pathology subjects of single disease.

Question subject	Number (%)					
	ChatGPT-4		ChatGPT-4o-P$^-$		ChatGPT-4o-P$^+$	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Multiple diseases (n = 37)	29 (78.4)	8 (21.6)	33 (89.2)	4 (10.8)	34 (91.9)	3 (8.1)
Single disease (n = 243)	194 (79.8)	49 (20.2)	215 (88.5)	28 (11.5)	218 (89.7)	25 (10.3)
1. Miscellaneous diseases (n = 80)	61 (76.3)	19 (23.7)	64 (80.0)	16 (20.0)	67 (83.8)	13 (16.2)
2. Developmental defects and oral manifestations of certain systemic diseases (n = 31)	27 (87.1)	4 (12.9)	27 (87.1)	4 (12.9)	29 (93.6)	2 (6.4)
3. Salivary gland pathology (n = 23)	18 (78.3)	5 (21.7)	23 (100.0)	0 (0.0)	22 (95.7)	1 (4.3)
4. Infectious diseases (n = 23)	23 (100.0)	0 (0.0)	23 (100.0)	0 (0.0)	23 (100.0)	0 (0.0)
5. Odontogenic cysts and tumors (n = 20)	15 (75.0)	5 (25.0)	18 (90.0)	2 (10.0)	17 (85.0)	3 (15.0)
6. Pigmented lesions (n = 18)	13 (72.2)	5 (17.8)	16 (88.9)	2 (11.1)	16 (88.9)	2 (11.1)
7. Bone pathology (n = 18)	15 (83.3)	3 (16.7)	16 (88.9)	2 (11.1)	17 (94.4)	1 (5.6)
8. Oral cancers and precancers (n = 16)	11 (68.8)	5 (31.2)	15 (93.8)	1 (6.2)	15 (93.8)	1 (6.2)
9. Abnormalities of teeth (n = 6)	3 (50.0)	3 (50.0)	5 (83.3)	1 (16.7)	4 (66.7)	2 (33.3)
10. Recurrent aphthous stomatitis and Behçet's disease (n = 6)	6 (100.0)	0 (0.0)	6 (100.0)	0 (0.0)	6 (100.0)	0 (0.0)
11. Vesiculobullous diseases and oral lichen planus (n = 2)	2 (100.0)	0 (0.0)	2 (100.0)	0 (0.0)	2 (100.0)	0 (0.0)

different question types (image-based questions, case-based questions, and odd-one-out questions), of two groups (single-disease and multiple diseases), and of 11 OP subjects of single disease. In general, ChatGPT-4o-P$^+$ achieved a higher overall AR of 90.0 % than ChatGPT-4 (79.6 % AR, $P < 0.001$, Table 1). Moreover, ChatGPT-4o-P$^-$ (88.6 % AR) also significantly outperformed ChatGPT-4 (79.6 % AR, $P = 0.006$, Table 1). These findings suggest that ChatGPT-4o is significantly better in answering the OP MCQs correctly than ChatGPT-4. In addition, when a prompt (such as please answer the OP questions based on two designated OP textbooks) was added to ChatGPT-4o (like ChatGPT-4o-P$^+$), the performance improved slightly but was not significantly better than the original ChatGPT-4o (ChatGPT-4o-P$^-$).

Three different models (ChatGPT-4, ChatGPT-4o-P$^-$, and ChatGPT-4o-P$^+$) were used to assess their performances in correctly answering the three types of OP questions. For the 92 odd-one-out questions, ChatGPT-4o-P$^+$ (92.4 % AR) slightly outperformed ChatGPT-4o-P$^-$ (91.3 % AR), but both ChatGPT-4o-P$^+$ and ChatGPT-4o-P$^-$ significantly exceeded ChatGPT-4 (77.2 % AR, $P = 0.008$ and $P = 0.015$, respectively, Table 2). For the 39 case-based questions, both ChatGPT-4o-P$^+$ (84.6 % AR) and ChatGPT-4o-P$^-$ (84.6 % AR) obtained equal performance, although both their ARs were higher than that (76.9 % AR) of ChatGPT-4, the differences were not significant (both P-values >0.05). Furthermore, for the 39 image-based questions, ChatGPT-4o-P$^+$ (84.6 % AR) slightly outperformed ChatGPT-4o-P$^-$ (79.5 % AR), although both their ARs were higher than that (71.8 % AR) of

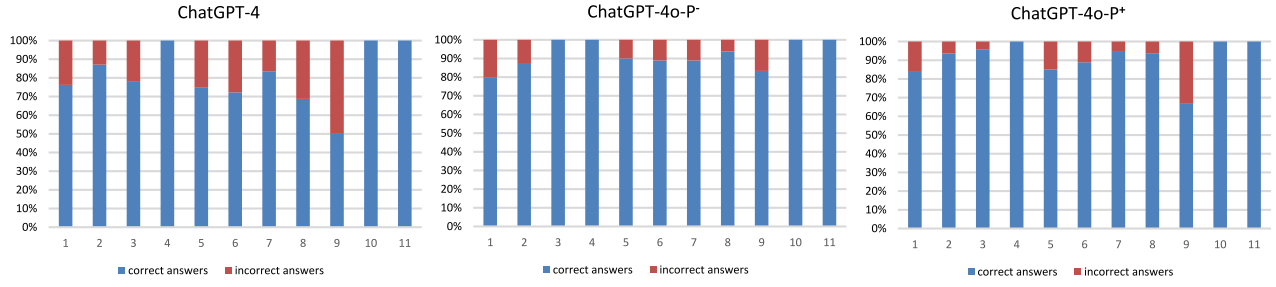

Figure 1 The percentage of questions with correct answers and incorrect answers generated by ChatGPT-4, ChatGPT-4o without the prompt (ChatGPT-4o-P$^-$), and ChatGPT-4o with the prompt (ChatGPT-4o-P$^+$) for 234 questions of 11 oral pathology subjects of single disease. The x-axis identifies 11 oral pathology subjects of single disease. Subject 1: Miscellaneous diseases; Subject 2: Developmental defects and oral manifestations of certain systemic diseases; Subject 3: Salivary gland pathology; Subject 4: Infectious diseases; Subject 5: Odontogenic cysts and tumors; Subject 6: Pigmented lesions; Subject 7: Bone pathology; Subject 8: Oral cancers and precancers; Subject 9: Abnormalities of teeth; Subject 10: Recurrent aphthous stomatitis and Behçet's disease; and Subject 11: Vesiculobullous diseases and oral lichen planus.

ChatGPT-4, the differences were also not significant (both *P*-values >0.05). These results indicate that different ChatGPT models may achieve various ARs for the three different types of OP questions, but in general the performance of ChatGPT-4o-P^+ is usually superior to those of the other two ChatGPT models, and the AR for odd-one-out questions is relatively higher than the ARs for either case-based questions or image-based questions, when the three different ChatGPT models were utilized to answer the three different types of OP questions. In addition, the specific and proper prompt can help the model generate more accurate and appropriate responses.[12]

Regarding the question types, most previous associated studies extracted text-based questions only for testing, but in addition to text-based questions we also included image-based and case-based questions for testing and compared the differences in ARs among three different ChatGPT models and between any two of the three different ChatGPT models.[13,14] Although there was no significant difference in ARs among them, our findings showed fair results in the overall ARs, even with those image-based questions being included. Jung et al. have proved no significant difference in the chatbot's performance of ChatGTP between case-based questions and non-case-based questions extracted from the German state examination in medicine, whereas the opposite result was reported by Tosun and Yilmaz.[14,15] In Tosun and Yilmaz's study, they evaluated AI-based chatbots of ChatGPT-3.5 and ChatGPT-4 in correctly answering prosthodontics questions from the Dental Specialty Exam in Turkey and found that ChatGPT-4 and ChatGPT-3.5 achieve significantly higher ARs in knowledge-based questions than in case-based questions. Considering that there might be discrepancy in the difficulty level between the case-based questions and the non-case-based questions, we did not focus on exploring the discrimination between them. However, case-based questions can test the capability of application of knowledge to diagnose the dental and medical diseases, it might be worthy to discover how to standardize these two kinds of questions for comparison in the future.[10]

Furthermore, we found no significant difference in the performances between OP questions in the single-disease group and those in the multiple disease group when using the three different ChatGTP models to answer the OP MCQs. Of the 11 different OP subjects of single-disease, all three different models (ChatGPT-4, ChatGPT-4o-P^-, and ChatGPT-4o-P^+) achieved a 100 % AR in three subjects (subjects 4, 10, and 11); ChatGPT-4o-P^+ outperformed ChatGPT-4 and ChatGPT-4o-P^- in other 3 subjects (subjects 1, 2, and 7); ChatGPT-4o-P^- was superior to ChatGPT-4 and ChatGPT-4o-P^+ in another 3 subjects (subjects 3, 5, and 9); and ChatGPT-4o-P^- and ChatGPT-4o-P^+ had equal performance and both were better than ChatGPT-4 in the rest of two subjects (subjects 6 and 8). These findings indicate that for correctly answering the questions of particular OP subjects of single disease, all the three ChatGPT models may have equal performance; however, for correctly replying the questions of most OP subjects of single disease, the performances of both ChatGPT-4o-P^- and ChatGPT-4o-P^+ are better than that of ChatGPT-4. In addition, for responding to the questions of most OP subjects of single disease, the performance of ChatGPT-4o-P^+ may be superior, equal, or inferior to that of ChatGPT-4o-P^-. This finding suggests that the performance of ChatGPT-4o-P^+ varies by the OP subject.

The variation in the performance on different subjects was also observed in other studies. Knoedler et al. proved that ChatGPT performs better in serology-related questions than in electrocardiography-related questions.[13] Bolgova et al. assessed ChatGPT's performance in answering questions on different topics in gross anatomy and found that ChatGPT answers better in questions of back than in those of other organs.[16] ChatGPT can be trained by the texts from a variety of sources, including books, scientific articles, or other website texts; however, it is unsettled whether the amount of literature of the particular disease influences the chatbot's performance or there might be other pivotal factors influencing the chatbot's performance.[3]

Our findings revealed the growing potential of the chatbots of different LLMs that can be applied in the medical and dental education. In general, an optimized LLM (such as ChatGPT-4o in this study) is usually superior to the original LLM (such as ChatGPT-4 in this study). When a prompt is given to an optimized LLM (such as ChatGPT-4o-P^+ in this study), compared to the original optimized LLM without a prompt (such as ChatGPT-4o-P^- in this study), the performance of ChatGPT-4o-P^+ may be superior, equal, or inferior to that of the ChatGPT-4o-P^-, depending on the questions of different subjects. This finding suggests that the selection of a precise and appropriate prompt is very important, because a precise and appropriate prompt may optimize the dataset to increase the AR of a LLM. On the contrary, an inexact and unappropriate prompt may narrow down the dataset to reduce the AR of a LLM. The prompt engineering is only one of the strategies that can improve the performance and responsiveness of a LLM. Other strategies such as continuously fine-tuning of the model on specific datasets, using a user feedback loop to let the model learn from its mistakes, integrating the model with external databases or tools for real-time information retrieval, etc. can also be used to optimize a LLM for better performance. Further studies are needed to evaluate how to include these strategies to a LLM for enhancing its ability to generate correct and relevant responses.

Declaration of competing interest

The authors have no conflicts of interest relevant to this article.

Acknowledgments

This study was partially supported by the grant [NSTC 111-2314-B-006-037-MY2] from the National Science and Technology Council, Taiwan.

References

1. Zhang N, Sun Z, Xie Y, Wu H, Li C. The latest version ChatGPT powered by GPT-4o: what will it bring to the medical field? *Int J Surg* 2024;110:6018—9.

2. Hirschberg J, Manning CD. Advances in natural language processing. *Science* 2015;349:261—6.
3. Wu T, He S, Liu J, et al. A brief overview of ChatGPT: the history, status quo and potential future development. *IEEE/-CAA J Autom Sin* 2023;10:1122—36.
4. 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.
5. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. *Healthcare (Basel)* 2023;11:2046.
6. Morishita M, Fukuda H, Muraoka K, et al. Evaluating GPT-4V's performance in the Japanese national dental examination: a challenge explored. *J Dent Sci* 2024;19:1595—600.
7. Fukuda H, Morishita M, Muraoka K, et al. Evaluating the image recognition capabilities of GPT-4V and Gemini Pro in the Japanese national dental examination. *J Dent Sci* 2025;20: 368—72.
8. Regezi JA, Sciubba JJ, Jordan RCK. *Oral pathology: clinical pathologic Correlations*, 7th ed. St Louis: WB Saunders, 2016.
9. Neville BW, Damm DD, Allen CM, Chi AC. *Oral and maxillofacial pathology*, 5th ed. St Louis: Elsevier, 2024.
10. Salam A, Yousuf R, Bakar SM. Multiple choice questions in medical education: how to construct high quality questions. *J Hum Health Sci* 2020;4:79—88.
11. Adeosun SO. Differences in multiple-choice questions of opposite stem orientations based on a novel item quality measure. *Am J Pharmaceut Educ* 2023;87:e8934.
12. Heston TF, Khun C. Prompt engineering in medical education. *IME* 2023;2:198—205.
13. Knoedler L, Knoedler S, Hoch CC, et al. In-depth analysis of ChatGPT's performance based on specific signaling words and phrases in the question stem of 2377 USMLE step 1 style questions. *Sci Rep* 2024;14:13553.
14. Jung LB, Gudera JA, Wiegand TLT, et al. ChatGPT passes German state examination in medicine with picture questions omitted. *Dtsch Arztebl Int* 2023;120:373—4.
15. Tosun B, Yilmaz ZS. Comparison of artificial intelligence systems in answering prosthodontics questions from the dental specialty exam in Turkey. *J Dent Sci* 2025;20:1454—9.
16. Bolgova O, Shypilova I, Sankova L, Mavrych V. How well did ChatGPT perform in answering questions on different topics in gross anatomy? *EJMHS* 2023;5:94—100.