Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Original Article

Impact of language and question types on ChatGPT-4o's performance in answering oral pathology questions 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 6 July 2025
Available online 19 July 2025

KEYWORDS
Artificial intelligence;
ChatGPT;
English-translated questions;
Oral pathology;
Dental education

Abstract *Background/purpose:* ChatGPT has been utilized in medical and dental education, but its performance is potentially influenced by factors like language, question types, and content complexity. This study aimed to assess how English translation and question types affect ChatGPT-4o's accuracy in answering English-translated oral pathology (OP) multiple choice questions (MCQs).

Materials and methods: A total of 280 OP MCQs were collected from Taiwan National Dental Licensing Examinations and English-translated as a testing set for ChatGPT-4o. The mean overall accuracy rates (ARs) for English-translated and non-translated MCQs were compared by the dependent *t*-test. The difference in ARs between English-translated and non-translated OP MCQs within each of three question types (image-based, case-based, and odd-one-out questions) was assessed by chi-square test. The binary logistic regression was used to determine which type of question was more likely to be answered incorrectly.

Results: ChatGPT-4o showed significantly higher mean overall AR (93.2 ± 5.7 %) for English-

* 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.2025.07.010

translated MCQs than for non-translated MCQs (88.6 ± 6.5 %, $P < 0.001$). There were no significant differences in the ARs between English-translated and non-translated MCQs within each question type. The binary logistic regression revealed that, within the English-translated condition, image-based questions were significantly more likely to be answered incorrectly (odds ratio = 9.085, $P = 0.001$).

Conclusion: Translation of exam questions into English significantly improved ChatGPT-4o's overall performance. Error pattern analysis confirmed that image-based questions were more likely to result in incorrect answers, reflecting the model's current limitations in visual reasoning. Nevertheless, ChatGPT-4o still demonstrated its strong potential as an educational support tool.

Introduction

Artificial intelligence (AI) has received increasing attention over the past decade, with natural language processing (NLP) emerging as one of its most prominent applications. NLP enables computers to understand, interpret, and generate human language.[1,2] Among these advancements, generative pre-trained transformers (GPT), developed by OpenAI in 2018, is a large language model (LLM), which is pre-trained with a vast collection of texts including books, articles, and website texts and subsequently fine-tuned to specialize in specific conversational tasks.[3]

ChatGPT-4 (OpenAI Global, San Francisco, CA, USA), released in 2023, and based on the earlier ChatGPT-3.5 model, rapidly gained global recognition for its impressive capabilities of language comprehension and generation within a conversational format. In May 2024, OpenAI released ChatGPT-4o ("o" standing for "omni"; OpenAI Global, San Francisco, CA, USA), introducing an improvement by enabling processing text, voice, and image inputs in real time.[1,3] This multimodal capability permits more natural and dynamic interactions between human and AI, expanding its potential applications in fields of education, healthcare, and research.[4]

ChatGPT has a wide range of applications in medicine, ranging from clinical support to medical education. It can assist not only in diagnostic procedures and patient communication, but also as an interactive tool to support learning and knowledge retrieval.[5–7] ChatGPT has been evaluated in answering multiple choice questions (MCQs) in licensing examinations for medical, dental, and pharmacy professionals across European, American and Asian nations. Reported accuracy rates (ARs) have differed among studies, generally varying from moderate to high performance depending on the testing models, subjects, and question types.[8–14] These variations may reflect not only the differences in exam structure and content, but also the model's varying ability to handle contextual interpretation, language complexity, and visual inputs, which might also affect the model's performance.

Given the increasing applications of AI chatbots in the medical education, it is important to understand the factors that may influence their performance. This study aimed to examine the impact of English translation of exam questions on ChatGPT-4o's performance by analyzing their capabilities in answering oral pathology (OP) MCQs to help clarify the model's strengths and limitations in serving as a support tool in dental education.

Materials and methods

Dataset construction and question processing

The Taiwan National Dental Licensing Examination (TNDLE) consisted of two parts and was held twice a year for dental students. If the dental students pass the TNDLE, they can acquire a dentist's license. The part I examination focused on basic dental sciences, comprising 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 (Dentistry III, IV, V, and VI) are mainly subjects of clinical dental sciences, such as endodontics, periodontology, operative dentistry, prosthodontics, pediatric dentistry, orthodontics, dental radiology, and oral and maxillofacial surgery. Each of the 6 Dentistry examinations has 80 MCQs, and there were 28 OP MCQs in every Dentistry II examination.

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 TNDLE 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, which were already utilized in our previous study.[15] A total of 280 OP MCQs (28 extracted OP questions twice per year in the selected five years) were collected as the dataset, and each MCQ had four answer options and only one was correct. These exam questions were originally written in Chinese with additional medical terms written in English. To analyze the impact of language, all questions and answer choices were translated into English and carefully reviewed and revised to ensure properness to the original text, accurate expression, and formatting consistency.

Study procedures and evaluation strategy

The English-translated OP MCQs were input into ChatGPT-4o to generate answers under standardized conditions without additional prompts. To minimize memory retention effects and ensure independence across every session, the previous records were ensured to be eliminated, and a new chat was initiated for each set of OP MCQs from a different examination year. The non-translated OP MCQs (original version with Chinese text and embedded English medical terms) were adopted and they had been previously tested and evaluated in our earlier study using the same model configuration, allowing for comparison of the accuracy rates (ARs, which were calculated as the proportion of the number of exam questions with correct answers to the total number of exam questions) between English-translated and non-translated OP MCQs.[15]

The answers generated by ChatGPT-4o were recorded and classified as correct or incorrect using the standard answers provided by the website of the Ministry of Examination. Three specific question types were selected and further classified as image-based MCQs (n = 39), case-based MCQs (n = 39), and odd-one-out MCQs (n = 92), according to their content features. Image-based MCQs referred to those questions that included visual content such as figures. Case-based MCQs were those questions with a particular clinical scenario for assessment of the application of relevant knowledge clinically.[16] Odd-one-out questions were defined by a reversed question orientation, where the correct choice was either the false statement or the option with the lowest likelihood.[17]

Statistical analysis

The performance of ChatGPT-4o on English-translated and non-translated OP MCQs was compared using a dependent *t*-test for paired samples, based on the ARs for each of the 10 test sets. To further assess whether English translation influenced the ChatGPT-4o's performance within specific question types (image-based, case-based, and odd-one-out questions), the chi-square test was conducted to evaluate the ARs between English-translated MCQs and non-translated MCQs within each specific question type. A binary logistic regression analysis was further performed on the English-translated MCQs to identify which question-type features were associated with incorrect answers, with graphical content, case-based context, and odd-one-out structure as independent variables. The significance level was set at $P < 0.05$ in all tests.

Results

The total number of correct answers in the English-translated OP MCQs was 261 (93.2 %) of 280 questions, while that in the non-translated MCQs was 248 (88.6 %) of 280 questions. Thus, when ChatGPT-4o was used to answer the 10 sets of exam questions (n = 28 for each set), the mean overall AR was significantly higher for the English-translated MCQs (93.2 % ± 5.7 %) than for the non-translated MCQs (88.6 % ± 6.5 %) ($P < 0.001$, Table 1). This pattern was consistent across every annual examination, with the English-translated OP MCQs showing higher ARs than the non-translated OP MCQs. These findings indicate that English translation can significantly improve the ChatGPT-4o performance.

Table 1 The number of correct answers and accuracy rates (ARs) generated by ChatGPT-4o for English-translated and non-translated oral pathology multiple choice questions (MCQs) in Dentistry II examinations from selected years between 2015 and 2024.

Year (number of question)	Number of questions with correct answers (%, AR)	
	Non-translated MCQs	English-translated MCQs
2015-1 (n = 28)	27 (96.4)	27 (96.4)
2015-2 (n = 28)	26 (92.9)	27 (96.4)
2016-1 (n = 28)	26 (92.9)	27 (96.4)
2016-2 (n = 28)	23 (82.1)	25 (89.3)
2018-1 (n = 28)	26 (92.9)	27 (96.4)
2018-2 (n = 28)	24 (85.7)	27 (96.4)
2021-1 (n = 28)	24 (85.7)	26 (92.9)
2021-2 (n = 28)	27 (96.4)	28 (100.0)
2024-1 (n = 28)	22 (78.6)	23 (82.1)
2024-2 (n = 28)	23 (82.1)	24 (85.7)
Total (n = 280)	248 (88.6)	261 (93.2)
Mean AR ± SD	88.6 ± 6.5	93.2 ± 5.7 [a]$P < 0.001$

[a] Comparison of mean overall AR between English-translated and non-translated oral pathology MCQs by dependent *t*-test for the paired samples.

Regarding the ARs generated by ChatGPT-4o between English-translated and non-translated OP MCQs within each question type, there were no statistically significant differences observed in any of the three question types (image-based, case-based, or odd-one-out questions) (Table 2). For image-based questions, the AR was 74.4 % for the English-translated MCQs and 79.5 % for the non-translated MCQs ($P = 0.788$, Table 2). For case-based questions, the AR was 82.1 % and 84.6 % for the English-translated MCQs and the non-translated MCQs, respectively ($P = 0.761$, Table 2). For odd-one-out questions, the AR (93.5 %) for the English-translated MCQs was slightly higher than the AR (91.3 %) for the non-translated MCQs, but the difference was also not significant ($P = 0.781$, Table 2). These findings suggest that within specific question types, English translation did not significantly affect the ARs generated by ChatGPT-4o.

To further evaluate the factors associated with errors generated by ChatGPT-4o on the English-translated OP MCQs, a binary logistic regression analysis was conducted and the results showed that image-based MCQs were significantly more likely to result in incorrect answers compared to the MCQs without images (odds ratio (OR) = 9.085, 95 % confidence interval (CI): 2.568–32.147, $P = 0.001$, Table 3). On the other hand, case-based contexts or odd-one-out structures did not significantly affect the likelihood of error ($P = 0.846$ and $P = 0.455$, respectively, Table 3). Taken together, within the English-translated condition, the presence of graphical content

Table 2 The accuracy rates (ARs) generated by ChatGPT-4o for English-translated and non-translated oral pathology multiple choice questions (MCQs) were compared within each question type by chi-square test.

	Image-based MCQs (n = 39)		Case-based MCQs (n = 39)		Odd-one-out MCQs (n = 92)	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
English-translated MCQs	29 (74.4)	10 (25.6)	32 (82.1)	7 (17.9)	86 (93.5)	6 (6.5)
Non-translated MCQs	31 (79.5)	8 (20.5)	33 (84.6)	6 (15.4)	84 (91.3)	8 (8.7)
[a]*P*-value		0.788		0.761		0.781

[a] The difference in ARs between English-translated and non-translated oral pathology MCQs within each question type was assessed by chi-square test.

Table 3 Binary logistic regression analysis of errors generated by ChatGPT-4o on English-translated oral pathology MCQs based on features of different question types.

Variable	Reference	Odds ratio [95 % confidence interval]	*P*-value
Image-based MCQs	MCQs without images	9.085 [2.568, 32.147]	0.001
Case-based MCQs	MCQs without case context	1.145 [0.292, 4.484]	0.846
Odd-one-out MCQs	MCQs without odd-one-out structure	1.532 [0.500, 4.691]	0.455

appeared to be a key factor contributing to errors, whereas the English translation alone may not affect ChatGPT-4o's performance on the image-based MCQs, as evidenced by the lack of significant difference between the English-translated and non-translated OP questions as shown in Table 2.

Discussion

This study demonstrated that translation of OP MCQs into English had a generally positive effect on ChatGPT-4o's performance, as reflected by the higher mean overall AR for the English-translated exam questions than the non-translated exam questions. Regarding different question types (image-based, case-based, or odd-one-out), no significant differences in the ARs were observed between the English-translated and non-translated OP exam questions. The above findings suggest that while English translation may have a positive impact on the overall ChatGPT-4o′s performance, its effect appears to be limited when question type is considered individually. Further analysis to examine error types on English-translated OP MCQs revealed that image-based questions were significantly more likely to generate incorrect answers, indicating that the graphical content poses a particular challenge for the ChatGPT-4o's performance, independent of language.

In our study, translation of the OP exam questions from Chinese (with the key medical terms written in English additionally) into English appeared to enhance ChatGPT-4o's overall performance on answering the OP MCQs correctly (mean overall AR increased from 88.6 % to 93.2 %). Similar findings with minor divergence were noted in other two associated studies.[10,14] Wang et al. reported the comprehensively-improved ChatGPT-3.5's performance on the English-translated version of the Taiwan national pharmacist licensing examination.[10] Fang et al. also found that translation of Chinese medical licensing examination questions into English increases the number of correct answers from 256 to 260 when using ChatGPT-4, although the difference does not reach the significant level ($P = 0.728$).[14] One possible explanation could be that English translations may be closer to the linguistic patterns and conceptual representations present in training dataset of GPT, which are predominantly English-based.[18] Even though key medical terms were provided in English within the Chinese exam questions, sentence structure, context framing, and grammatical cues in Chinese might still produce ambiguities that affect the model comprehension. This could be partially supported by our previous findings that ChatGPT-4 achieved significantly lower AR on the Chinese odd-one-out questions (77.2 %) compared to ChatGPT-4o (91.3 %, $P = 0.015$), suggesting that the earlier versions of the model might be struggled with subtle semantic distinctions in Chinese exam question formats.[15]

Further analysis of the English-translated OP MCQs showed that image-based questions were associated with more incorrect answers, highlighting that visual information might introduce unique difficulties for the model unrelated to language. Morishita et al. assessed the capabilities of ChatGPT-4V with image recognition in answering 160 image-based questions from the Japanese national dental licensing examination to explore its potential as an educational support tool for dental students. They demonstrated that the overall AR of ChatGPT-4V for 160 image-based questions is 35.0 %. Moreover, a higher number of images in questions is correlated with lower accuracy, suggesting an impact of the number of images on the ARs.[12] In addition, Fukuda et al. also compared the performance of two LLMs with the image recognition capabilities, ChatGPT-4V and Gemini Pro, in answering the 160 image-based questions from the Japanese national dental licensing examination. They found that the overall AR of ChatGPT-4V (35.0 %) is higher than that of Gemini Pro

(28.1 %), but the difference is not statistically significant. Furthermore, they also observed that the ARs tend to decrease with an increased number of images within a question, suggesting that the number of images influences the correctness of the responses.[11] This difficulty in visual processing persists even when questions are presented in English originally, as demonstrated by Hayden et al., who reported that ChatGPT-4 performs dramatically worse on the image-based radiology questions than on the text-only radiology questions (47.8 % vs. 81.5 %, $P < 0.001$).[19] Moreover, Kaneyasu et al. used ChatGPT-4.5 to answer the 213 MCQs (including 74 text-only and 139 visually-based question) from the 34th Japanese national dental hygienist examination. They found the ChatGPT-4.5's ARs are 76.1 % for all questions, 87.8 % for text-only questions, and 69.8 % for visually-based questions, indicating a relatively higher AR for text-only questions than for visually-based questions.[20] All the above-mentioned findings collectively point to a consistent limitation in current LLMs' capabilities of processing visual information and spatial reasoning, underlining the need for continued model refinement in multimodal understanding.

Based on our findings, although the restriction in handling visual information of the model remained apparent, translation of Chinese OP MCQs into English significantly improved the ChatGPT-4o's overall performance. Therefore, given its strong performance on text-based question and convenience, ChatGPT-4o is indeed a promising tool to support learning in dental education, although it still needs the continued development in enhancing the integration of visual and linguistic information in the current LLMs.

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. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. *Clin Med* 2023;23:278—9.
5. Lin YC, Cheng FC, Lin WC, Chiang CP. Artificial intelligence measurement of multi-layer tooth structures using semantic segmentation and computer vision. *J Dent Sci* 2025;20:723—5.
6. Skryd A, Lawrence K. ChatGPT as a tool for medical education and clinical decision-making on the wards: case study. *JMIR Form Res* 2024;8:e51346.
7. Su AY, Wu ML, Wu YH. Deep learning system for the differential diagnosis of oral mucosal lesions through clinical photographic imaging. *J Dent Sci* 2025;20:54—60.
8. 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.
9. Jin HK, Kim E. Performance of GPT-3.5 and GPT-4 on the Korean pharmacist licensing examination: comparison study. *JMIR Med Educ* 2024;10:e57451.
10. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. *J Chin Med Assoc* 2023;86:653—8.
11. 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.
12. 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.
13. 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.
14. Fang C, Wu Y, Fu W, et al. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. *PLoS Digit Health* 2023;2:e0000397.
15. Wu YH, Tso KY, Chiang CP. Performance of ChatGPT in answering the oral pathology questions of various types or subjects from Taiwan National Dental Licensing Examinations. *J Dent Sci* 2025;20:1709—15.
16. 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.
17. 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.
18. Johnson R, Pistilli G, Menédez-González N, et al. The Ghost in the Machine has an American accent: value conflict in GPT-3. *arXiv* 2022;2005:14165.
19. Hayden N, Gilbert S, Poisson LM, Griffith B, Klochko C. Performance of GPT-4 with vision on text- and image-based ACR diagnostic radiology in-training examination questions. *Radiology* 2024;312:e240153.
20. Kaneyasu Y, Mine Y, Niitani Y, et al. Analysis of multimodal large language models on visually-based questions in the Japanese National Examination for Dental Hygienists: a preliminary comparative study. *J Dent Sci* 2025 (in press).