Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Original Article

Performance of ChatGPT-4, Gemini, and DeepSeek-V3 on answering the multiple choice questions from Taiwan national dental technician licensing examinations and their self-learning abilities over a three-week period

Ching-Yi Huang [a,b†], Yi-Pang Lee [a,b†], Andy Sun [c**], Chun-Pin Chiang [a,b,c,d*]

[a] *Institute of Oral Medicine and Materials, College of Medicine, Tzu Chi University, Hualien, Taiwan*
[b] *Department of Dentistry, Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Hualien, Taiwan*
[c] *Department of Dentistry, National Taiwan University Hospital, College of Medicine, National Taiwan University, Taipei, Taiwan*
[d] *Graduate Institute of Oral Biology, School of Dentistry, National Taiwan University, Taipei, Taiwan*

Received 6 July 2025
Available online 21 July 2025

KEYWORDS
Artificial intelligence;
Dental technology;
Large language model;
ChatGPT-4;
Gemini;
DeepSeek-V3

Abstract *Background/purpose:* Large language models (LLMs) can help the students to learn specific dental subjects and thus can be used as educational support tools for dental students. This study evaluated whether LLMs could correctly answer the multiple-choice questions (MCQs) selected from the 2023 Taiwan national dental technician licensing examination (TNDTLE) and whether the LLMs had the self-learning ability to improve their performance on correctly answering the exam questions over a three-week period.

Materials and methods: Three different LLMs, ChatGPT-4, Gemini, and DeepSeek-V3, were used to answer the 194 text-based MCQs selected from the 2023 TNDTLE and the initial accuracy rates (ARs) were recorded. The same process was performed one, two, and three weeks

[*] 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.

[**] Corresponding author. Department of Dentistry, National Taiwan University Hospital, No. 1, Chang-Te Street, Taipei 10048, Taiwan.

E-mail addresses: andysun7702@yahoo.com.tw (A. Sun), cpchiang@ntu.edu.tw (C.-P. Chiang).

[†] These two authors contributed equally to this work.

https://doi.org/10.1016/j.jds.2025.07.011

later and the subsequent ARs were also recorded. The initial and the subsequent overall ARs were compared to assess whether the three LLMs had the self-learning ability over time.

Results: The initial overall ARs for ChatGPT-4, Gemini, and DeepSeek-V3 were 52.1 %, 57.2 %, and 69.6 %, respectively, indicating that DeepSeek-V3 outperforms ChatGPT-4 and Gemini. However, Gemini showed significant improvement in performance one week and three weeks later, but the ChatGPT-4 and DeepSeek-V3 showed no significant improvement in performance over time. Among the 9 different subjects of dental technology, Gemini showed notable progress in several subjects, ChatGPT-4 showed limited improvement, and DeepSeek-V3 remained stable overall.

Conclusion: Without external prompts, Gemini demonstrates self-learning potential. DeepSeek-V3 shows stable performance but limited learning ability, while ChatGPT-4 exhibits minimal learning. For the improvement in self-learning ability over time, Gemini outperforms ChatGPT-4 and DeepSeek-V3.

Introduction

Artificial Intelligence (AI) refers to computer systems designed to perform tasks that usually require human intelligence, such as problem-solving, learning, and decision-making. Large language models (LLMs) are a type of AI trained on vast amounts of text data. They can understand and generate human-like language, enabling applications like chatbots, writing assistants, and medical question answering.[1—7]

ChatGPT is a LLM with an AI chatbot developed by OpenAI. It is based on the GPT (Generative Pre-trained Transformer) architecture and can understand and generate human-like text. The main functions of ChatGPT include understanding and generating human-like text, answering questions in various subjects, assisting with writing, editing, and summarizing content, supporting learning and tutoring in education, helping with coding, brainstorming, and idea generation. Thus, ChatGPT is now widely used in the education, customer service, content creation, and research. ChatGTP-3.5 was the first LLM available to the public in late November 2022. ChatGPT-4 (OpenAI Global, San Francisco, CA, USA, released on March 14, 2023), based on ChatGPT-3.5 LLM, 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—5]

Gemini is a LLM developed by Google DeepMind (Mountain View, CA, USA, released on December 19, 2024). It combines language understanding with advanced reasoning and is designed for tasks like answering questions, explaining concepts, and solving problems across subjects.[5] Deepseek is a newer LLM created by a Chinese research team. It focuses on academic and professional use, especially in science, technology, engineering, and mathematics fields. Deepseek can solve mathematic problems, generate code, and support research-related questions. Moreover, it excels at tasks that require professional knowledge and logical reasoning, such as science, mathematics, and programming.[6,7]

In the dental field, the majority of the previous studies have explored different LLMs' (such as ChatGPT, Gemini, and DeepSeek) performance on answering the exam questions selected from the national dental licensing examinations, national dental hygienist licensing examinations, and national dental technician licensing examinations. The cumulative results from these previous studies confirm the LLMs' potential in understanding and generating relevant responses and thus these LLMs may be used as the educational support tools for dental school students, although the performance of these LLMs has not yet reached the ideal condition.[8—15] Up to date, only one study used ChatGPt-4o, OpenAI o1, and Claude 3.5 Sonnet to assess their performance on answering exam questions from the Japanese national dental technician licensing examination.[15] However, there is still no research specifically evaluating the performance of ChatGPT-4, Gemini, and DeepSeek on the exam questions from the Taiwan national dental technician licensing examination (TNDTLE).

Therefore, this study used three different LLMs, ChatGPT-4, Gemini 2.5 Flash (hereinafter referred to as Gemini), and ChatStream x DeepSeek-V3 (hereinafter referred to as DeepSeek-V3), to answer the 194 text-based exam questions selected from the 2023 TNDTLE and the initial accuracy rates (ARs) were recorded. The same process was performed one, two, and three weeks later and the subsequent ARs for the three LLMs were also recorded. We tried to understand whether these three LLMs could answer the exam questions correctly and what their exact ARs were. In addition, the initial and the subsequent overall ARs were compared to assess whether the three LLMs had the self-learning ability to improve their performance on correctly answering the exam questions over time.

Materials and methods

This study primarily investigated the performance of three different LLMs, ChatGPT-4, Gemini, and DeepSeek-V3, in correctly answering the exam questions selected from the 2023 TNDTLE. The whole TNDTLE consisted of four parts: Exam 1 included three subjects: oral anatomy and physiology, dental morphology, and dental materials science; Exam 2 consisted of a single subject: fixed prosthodontics technology; Exam 3 included two subjects: complete denture technology and removable partial denture technology; and Exam 4 covered three subjects: orthodontic technology, pediatric dental technology, and dental technology regulations and ethics. Of the four exams, each exam contained 50 multiple-choice questions (MCQs) with a total of 200 MCQs in the whole annual TNDTLE. Moreover, each MCQ had four answer options and only one of them was correct.

In this study, 194 text-based exam questions without images or tables were selected. We then tested the efficacies of the three LLMs (ChatGPT-4, Gemini, and DeepSeek-V3) on answering the selected 194 exam questions and the initial ARs were recorded. Then, the same process using the same set of exam questions was repeated one week, two weeks, and three weeks later, and the subsequent ARs for each LLM were also recorded. The main aim of this study was to understand whether the three LLMs could answer the exam questions correctly and what their exact ARs were. In addition, the initial and the subsequent overall ARs were compared to assess whether the three LLMs had the self-learning ability of improving their performance on correctly answering the exam questions over a three-week period.

Statistical analysis

We used the chi-square test in the IBM SPSS statistical software (Chicago, IL, USA) to analyze whether the ARs of the three LLMs in answering the same set of exam questions showed a statistically significant increase over different time points (one week, two weeks, and three weeks later) compared to their initial ARs. Furthermore, we conducted a comparative analysis among the three different LLMs to determine whether one LLM demonstrated a significantly higher AR than each of the other two LLMs when answering the same set of exam questions.

Results

This study evaluated the performance of three LLMs (ChatGPT-4, Gemini, and DeepSeek-V3) on correctly answering the 194 exam questions from the 2023 TNDTLE at four different time points: initially, one week later, two weeks later, and three weeks later.

Initial accuracy rates (ARs) and changes of AR over time for the three LLMs

For ChatGPT-4, the initial overall AR was 52.1 % (101/194). The AR slightly increased to 56.2 % after one week and remained at 56.2 % after two weeks. However, after three weeks, the AR dropped back to the initial rate of 52.1 %. Statistical analysis showed that there was no significant difference between the initial overall AR and the subsequent overall ARs at any three later time points. These findings indicate that, in the absence of additional guidance or prompts, the ChatGPT-4 exhibits only minimal self-learning ability after one week, and this self-learning ability does not improve further over time (Table 1).

For Gemini, the initial overall AR was 57.2 %. The overall ARs after one week, two weeks, and three weeks were 67.0 %, 66.0 %, and 67.0 %, respectively. Both the subsequent overall ARs at one week and three weeks later showed a statistically significant increase compared to the initial overall AR. Although the subsequent overall AR after two weeks also increased, it did not reach a statistically significant level when compared to the initial overall AR. These results indicate that Gemini demonstrates self-learning ability without additional guidance or prompts, with the most notable improvement observed after one week. However, the self-learning effect did not continue to improve over time at two weeks and three weeks later (Table 1).

For DeepSeek-V3, the initial overall AR was 69.6 %, while the ARs after one week, two weeks, and three weeks later were 72.7 %, 72.2 %, and 70.1 %, respectively. Although

Table 1 The initial, post-one week, post-two week, and post-three week overall accuracy rates (ARs) for the three large language models (LLMs) of ChatGPT-4, Gemini, and DeepSeek-V3 in answering the 194 text-based exam questions selected from the 2023 Taiwan national dental technician licensing examination as well as the comparisons of the overall ARs between any two of the three LLMs of the ChatGPT-4, Gemini, and DeepSeek-V3.

LLM	Number of question with correct answer (AR)			
	Initial	Post-one week	Post-two weeks	Post-three weeks
ChatGPT-4 (n = 194)	101 (52.1 %)	109 (56.2 %)	109 (56.2 %)	101 (52.1 %)
Gemini (n = 194)	111 (57.2 %)	130 (67.0 %)	128 (66.0 %)	130 (67.0 %)
**P*-value	0.308	0.028	0.048	0.003
DeepSeek-V3 (n = 194)	135 (69.6 %)	141 (72.7 %)	140 (72.2 %)	136 (70.1 %)
**P*-value	<0.001	<0.001	<0.001	<0.001
***P*-value	0.011	0.224	0.187	0.512

*P-value Comparison of the overall ARs between the ChatGPT-4 and Gemini or DeepSeek-V3 by chi-square test.

**P-value Comparison of the overall ARs between Gemini and DeepSeek-V3 by chi-square test.

there was a slight increase in the subsequent overall AR over time, none of these increases were statistically significant compared to the initial overall AR. These results suggest that DeepSeek-V3, without any guidance or prompts, does not show self-learning ability (Table 1).

Comparison of the overall accuracy rates among the three LLMs

The overall ARs of the three LLMs at the initial stage, one week later, two weeks later, and three weeks later were compared pairwise, and the results are summarized in Table 1. In the comparison between Gemini and ChatGPT-4, the initial overall ARs of the Gemini (57.2 %) and ChatGPT-4 (52.1 %) showed no statistically significant difference ($P = 0.308$). However, at one week, two weeks, and three weeks later, the Gemini consistently demonstrated significantly higher overall ARs compared to ChatGPT-4 (all the P-values <0.05) (Table 1).

In the comparison between DeepSeek-V3 and ChatGPT-4, DeepSeek-V3 consistently exhibited significantly higher overall ARs than ChatGPT-4 at all four time points (initial and one week, two weeks, and three weeks later). The differences in the overall AR between the two LLMs at each time point were statistically significant, with all the P-values <0.001 (Table 1).

When comparing the overall ARs between DeepSeek-V3 and Gemini, DeepSeek-V3's initial overall AR (69.6 %) was significantly higher than the Gemini's initial overall AR (57.2 %) ($P = 0.011$). However, at one week, two weeks, and three weeks later, there were no statistically significant differences in the subsequent overall ARs between DeepSeek-V3 and Gemini (all the P-values >0.05) (Table 1).

Accuracy rates of the three LLMs across nine different subjects of dental technology

Accuracy rates of ChatGPT-4 across nine different subjects of dental technology

For the initial ARs of the 9 subjects of dental technology, ChatGPT-4 acquired the highest AR in the subject 9 (70.0 %). This was followed by the relatively higher ARs in the subjects 7 (68.4 %) and 1 (60 %). The ARs for the subjects 2 (57.9 %), 3 (57.9 %), and 5 (52.0 %) were slightly lower. The lower ARs were found in the subjects 4 (44.9 %) and 8 (44.4 %). The lowest AR was found in the subject 6 (29.2 %). Regarding improvement over time, the subjects that showed a notable increase in the ARs (≥10 % improvement) included the subject 8 (improving from 44.4 % to 66.7 %), the subject 9 (improving from 70.0 % to 80.0 %). Conversely, the subjects that demonstrated a significant decline in the ARs (≥10 % decrease) over time was the subject 2 (dropping from 57.9 % to 43.4 %). Overall, for ChatGPT-4, the ARs across most subjects tended to show only a slight improvement over time (Table 2).

Accuracy rates of Gemini across nine different subjects of dental technology

For the initial ARs of the 9 subjects, Gemini obtained the highest AR in the subject 9 (70.0 %). This was followed by the relatively higher ARs in the subjects 1 (60.0 %), 5 (68 %), 7 (63.2 %), and 8 (66.7 %). The ARs for the subjects 2 (57.9 %) and 6 (50 %) were slightly lower. The lowest ARs were found in the subjects 3 (47.4 %) and 4 (49.0 %). Regarding improvement over time, the subjects that showed a notable increase in the ARs (≥10 % improvement) included the subject 1 (improving from 60 % to 90 %, and eventually to 100 %), the subject 3 (improving from 47.4 % to 57.9 %, and then to 63.2 %), the subject 4 (improving from 49.0 % to 59.2 %, and then to 63.3 %), the subject 6 (improving from 50 % to 70.8 %, and then to 75 %), the subject 7 (improving from 63.2 % to 73.7 %), the subject 8 (improving from 66.7 % to 77.8 %, and eventually to 100 %), and the subject 9 (improving from 70 % to 80.0 %, and eventually to 90 %). Conversely, the subjects that demonstrated a significant decline in the ARs (≥10 % decrease) over time were the subject 2 (dropping from 57.9 % to 42.1 %) and the subject 5 (dropping from 68.0 % to 56.0 %). Overall, for the Gemini, the ARs across most subjects tended to show more noticeable improvement over time (Table 3).

Accuracy rates of DeepSeek-V3 across nine different subjects of dental technology

For the initial ARs of the 9 subjects, DeepSeek-V3 achieved the highest AR in the subject 1 (90 %). This was followed by the relatively higher ARs in the subjects 4 (75.5 %), 8 (77.8 %), and 9 (70 %). The ARs for the subjects 2 (68.4 %), 3 (63.2 %), 5 (68.0 %), and 7 (68.4 %) were slightly lower. The lowest AR was found in the subject 6 (54.2 %) (Table 4). Regarding improvement over time, the subjects that showed a notable increase in the ARs (≥10 % improvement) included the subject 1 (improving from 90 % to 100 %), the subject 6 (improving from 54.2 % to 66.7 %), the subject 7 (improving from 68.4 % to 89.5 %), and the subject 9 (improving from 70 % to 85 %, and then to 90 %). Conversely, the subjects that demonstrated a significant decline in the ARs (≥10 % decrease) over time were the subject 1 (dropping from 90 % to 80 %) and the subject 8 (dropping from 77.8 % to 55.6 %). For the remaining subjects, the ARs either slightly increased or slightly decreased over time, but the differences were all less than 10 % (Table 4).

Accuracy rates of ChatGPT-4, Gemini, DeepSeek-V3 for 4 different basic subjects of dental technology and for 5 different clinical subjects of dental technology

If we further divided the 194 text-based exam questions into the 68 text-based exam questions of 4 basic subjects of dental technology (oral anatomy and physiology, dental morphology, dental materials science, and dental technology regulations and ethics) and the 126 text-based exam questions of 5 clinical subjects of dental technology (fixed prosthodontics technology, complete denture technology, removable partial denture technology, orthodontic technology, and pediatric dental technology), and subsequently compared the initial, post-one week, post-two week, and post-three week overall ARs for the three LLMs of ChatGPT-4, Gemini, and DeepSeek-V3 in answering the 68 text-based exam questions of 4 basic subjects of dental technology and the 126 text-based exam questions of 5 clinical subjects of

Table 2 The initial, post-one week, post-two week, and post-three week accuracy rates (ARs) for ChatGPT-4 in answering the 194 text-based exam questions selected from the Taiwan national dental technician licensing examination according to 9 different subjects of dental technology.

Subject	Number of question with correct or incorrect answer (Correct or incorrect rate)			
	Initial	Post-one week	Post-two weeks	Post-three weeks
1. Oral anatomy and physiology (n = 10)				
Correct	6 (60.0 %)	6 (60.0 %)	6 (60.0 %)	6 (60.0 %)
Incorrect	4 (40.0 %)	4 (40.0 %)	4 (40.0 %)	4 (40.0 %)
2. Dental morphology (n = 19)				
Correct	11 (57.9 %)	11 (57.9 %)	11 (57.9 %)	9 (43.4 %)
Incorrect	8 (42.1 %)	8 (42.1 %)	8 (42.1 %)	10 (52.6 %)
3. Dental materials science (n = 19)				
Correct	11 (57.9 %)	11 (57.9 %)	11 (57.9 %)	11 (57.9 %)
Incorrect	8 (42.1 %)	8 (42.1 %)	8 (42.1 %)	8 (42.1 %)
4. Fixed prosthodontics technology (n = 49)				
Correct	22 (44.9 %)	22 (44.9 %)	22 (44.9 %)	23 (46.9 %)
Incorrect	27 (55.1 %)	27 (55.1 %)	27 (55.1 %)	26 (53.1 %)
5. Complete denture technology (n = 25)				
Correct	13 (52.0 %)	15 (60.0 %)	15 (60.0 %)	13 (52.0 %)
Incorrect	12 (48.0 %)	10 (40.0 %)	10 (40.0 %)	12 (48.0 %)
6. Removable partial denture technology (n = 24)				
Correct	7 (29.2 %)	8 (33.3 %)	8 (33.3 %)	8 (33.3 %)
Incorrect	17 (70.8 %)	16 (66.7 %)	16 (66.7 %)	16 (66.7 %)
7. Orthodontic technology (n = 19)				
Correct	13 (68.4 %)	14 (73.7 %)	14 (73.7 %)	14 (73.7 %)
Incorrect	6 (31.6 %)	5 (26.3 %)	5 (26.3 %)	5 (26.3 %)
8. Pediatric dental technology (n = 9)				
Correct	4 (44.4 %)	6 (66.7 %)	6 (66.7 %)	4 (44.4 %)
Incorrect	5 (55.6 %)	3 (33.3 %)	3 (33.3 %)	5 (55.6 %)
9. Dental technology regulations and ethics (n = 20)				
Correct	14 (70.0 %)	16 (80.0 %)	16 (80.0 %)	13 (65.0 %)
Incorrect	6 (30.0 %)	4 (20.0 %)	4 (20.0 %)	7 (35.0 %)
Total (n = 194)				
Correct	101 (52.1 %)	109 (56.2 %)	109 (56.2 %)	101 (52.1 %)
Incorrect	93 (47.9 %)	85 (43.8 %)	85 (43.8 %)	93 (47.9 %)

dental technology, we found that for all three LLMs of ChatGPT-4, Gemini, and DeepSeek-V3, the initial, post-one week, post-two week, and post-three week overall ARs for the 4 basic subjects of dental technology were always higher than the overall ARs for the 5 clinical subjects of dental technology. However, the differences in the overall ARs between the 4 basic subjects and the 5 clinical subjects of dental technology did not reach the statistically significant levels (all the *P*-values >0.05) for the three LLMs of ChatGPT-4, Gemini, and DeepSeek-V3 (Table 5).

Discussion

This study assessed the performance and potential self-learning abilities of three LLMs (including ChatGPT-4, Gemini, and DeepSeek-V3) on correctly answering the 194 questions from the 2023 TNDTLE over a three-week period. Overall, our results showed considerable variation in the initial ARs among the three LLMs and different capacities for accuracy improvement over time.

Among the overall ARs of the three LLMs, DeepSeek-V3 demonstrated the highest initial overall AR (69.6 %) which significantly outperformed both Gemini (57.2 %) and ChatGPT-4 (52.1 %). While the DeepSeek-V3's AR increased slightly in the following three weeks, these changes were not statistically significant, indicating that its accuracy remained relatively stable. Moreover, this finding suggests that although DeepSeek-V3 has superior baseline performance, it lacks substantial self-improvement capability without targeted guidance or new input.

Regarding the overall ARs of **Gemini,** on the other hand, Gemini exhibited a moderate initial overall AR, but demonstrated statistically significant improvements at both one and three weeks (67.0 % at both time points), suggesting a modest but detectable form of self-learning. However, the Gemini's performance plateaued thereafter, with no significant gain at two weeks. These findings imply that Gemini may be responsive to internal data updates or background adjustments by its developers, although no explicit retraining was applied. In contrast to DeepSeek-V3, Gemini showed greater fluctuations and improvement in its

Table 3 The initial, post-one week, post-two week, and post-three week accuracy rates (ARs) for Gemini in answering the 194 text-based exam questions selected from the 2023 Taiwan national dental technician licensing examination according to 9 different subjects of dental technology.

Subject	Number of question with correct or incorrect answer (Correct or incorrect rate)			
	Initial	Post-one week	Post-two weeks	Post-three weeks
1. Oral anatomy and physiology (n = 10)				
Correct	6 (60.0 %)	10 (100.0 %)	9 (90.0 %)	9 (90.0 %)
Incorrect	4 (40.0 %)	0 (0.0 %)	1 (10.0 %)	1 (10.0 %)
2. Dental morphology (n = 19)				
Correct	11 (57.9 %)	10 (52.6 %)	8 (42.1 %)	10 (52.6 %)
Incorrect	8 (42.1 %)	9 (47.4 %)	11 (57.9 %)	9 (47.4 %)
3. Dental materials science (n = 19)				
Correct	9 (47.4 %)	11 (57.9 %)	12 (63.2 %)	12 (63.2 %)
Incorrect	10 (52.6 %)	8 (42.1 %)	7 (36.8 %)	7 (36.8 %)
4. Fixed prosthodontics technology (n = 49)				
Correct	24 (49.0 %)	29 (59.2 %)	31 (63.3 %)	28 (57.1 %)
Incorrect	25 (51.0 %)	20 (40.8 %)	18 (36.7 %)	21 (42.9 %)
5. Complete denture technology (n = 25)				
Correct	17 (68.0 %)	16 (64.0 %)	14 (56.0 %)	15 (60.0 %)
Incorrect	8 (32.0 %)	9 (36.0 %)	11 (44.0 %)	10 (40.0 %)
6. Removable partial denture technology (n = 24)				
Correct	12 (50.0 %)	17 (70.8 %)	17 (70.8 %)	18 (75.0 %)
Incorrect	12 (50.0 %)	7 (29.2 %)	7 (29.2 %)	6 (25.0 %)
7. Orthodontic technology (n = 19)				
Correct	12 (63.2 %)	11 (57.9 %)	12 (63.2 %)	14 (73.7 %)
Incorrect	7 (36.8 %)	8 (42.1 %)	7 (36.8 %)	5 (26.3 %)
8. Pediatric dental technology (n = 9)				
Correct	6 (66.7 %)	8 (88.9 %)	9 (100.0 %)	7 (77.8 %)
Incorrect	3 (33.3 %)	1 (11.1 %)	0 (0.0 %)	2 (22.2 %)
9. Dental technology regulations and ethics (n = 20)				
Correct	14 (70.0 %)	18 (90.0 %)	16 (80.0 %)	17 (85.0 %)
Incorrect	6 (30.0 %)	2 (10.0 %)	4 (20.0 %)	3 (15.0 %)
Total (n = 194)				
Correct	111 (57.2 %)	130 (67.0 %)	128 (66.0 %)	130 (67.0 %)
Incorrect	83 (42.8 %)	64 (33.0 %)	66 (34.0 %)	64 (33.0 %)

subject-specific ARs, indicating a potentially more dynamic response to stored information or feedback mechanisms.

ChatGPT-4, while widely known for its versatility, showed limited improvement. Its overall AR increased marginally to 56.2 % after one and two weeks but returned to 52.1 % by the third week. No statistically significant changes were observed at any time point. These results indicate that, in its current form, ChatGPT-4 does not exhibit measurable autonomous learning across repeated trials with the same questions. Its performance stability also suggests a reliance on its existing knowledge base without background updates or adaptations during the study period of three weeks.

In this study, the whole TNDTLE consisted of 194 text-based exam questions for testing of 4 basic subjects of dental technology (oral anatomy and physiology, dental morphology, dental materials science, and dental technology regulations and ethics) and 5 clinical subjects of dental technology (fixed prosthodontics technology, complete denture technology, removable partial denture technology, orthodontic technology, and pediatric dental technology). The subject-wise analysis revealed differences in performance across the nine dental technology subjects. All the three LLMs tended to perform better on the **basic subjects** than on the **clinical subjects**. Although the overall ARs for the 4 basic subjects were consistently higher than the overall ARs for the 5 clinical subjects, the differences did not reach statistically significant level. These results may reflect the structured and fact-based nature of the basic science content, which aligns more closely with the LLMs' pre-trained data, compared to the more delicate reasoning often required for solving the questions of clinical topics.

In subject-specific improvements, Gemini showed the most remarkable gains over time in several subjects, such as oral anatomy and physiology, removable partial denture technology, pediatric dental technology, and dental technology regulations and ethics, with some ARs improving by 20—40 %. ChatGPT-4 showed only marginal gains or fluctuations, while DeepSeek-V3 remained the most consistent, with minor variations across the 9 subjects.

Several different famous LLMs were used to answer the exam questions selected from national dental licensing examinations, national dental hygienist licensing

Table 4 The initial, post-one week, post-two week, and post-three week accuracy rates (ARs) for DeepSeek-V3 in answering the 194 text-based exam questions selected from the 2023 Taiwan national dental technician licensing examination according to 9 different subjects of dental technology.

Subject	Number of question with correct or incorrect answer (Correct or incorrect rate)			
	Initial	Post-one week	Post-two weeks	Post-three weeks
1. Oral anatomy and physiology (n = 10)				
Correct	9 (90.0 %)	8 (80.0 %)	10 (100.0 %)	10 (100.0 %)
Incorrect	1 (10.0 %)	2 (20.0 %)	0 (0.0 %)	0 (0.0 %)
2. Dental morphology (n = 19)				
Correct	13 (68.4 %)	13 (68.4 %)	11 (57.9 %)	12 (63.2 %)
Incorrect	6 (31.6 %)	6 (31.6 %)	8 (42.1 %)	7 (36.8 %)
3. Dental materials science (n = 19)				
Correct	12 (63.2 %)	12 (63.2 %)	13 (68.4 %)	12 (63.2 %)
Incorrect	7 (36.8 %)	7 (36.8 %)	6 (31.6 %)	7 (36.8 %)
4. Fixed prosthodontics technology (n = 49)				
Correct	37 (75.5 %)	36 (73.5 %)	36 (73.5 %)	34 (69.4 %)
Incorrect	12 (24.5 %)	13 (26.5 %)	13 (26.5 %)	15 (30.6 %)
5. Complete denture technology (n = 25)				
Correct	17 (68.0 %)	19 (76.0 %)	18 (72.0 %)	16 (64.0 %)
Incorrect	8 (32.0 %)	6 (24.0 %)	7 (28.0 %)	9 (36.0 %)
6. Removable partial denture technology (n = 24)				
Correct	13 (54.2 %)	13 (54.2 %)	16 (66.7 %)	15 (62.5 %)
Incorrect	11 (45.8 %)	11 (45.8 %)	8 (33.3 %)	9 (37.5 %)
7. Orthodontic technology (n = 19)				
Correct	13 (68.4 %)	17 (89.5 %)	14 (73.7 %)	13 (68.4 %)
Incorrect	6 (31.6 %)	2 (10.5 %)	5 (26.3 %)	6 (31.6 %)
8. Pediatric dental technology (n = 9)				
Correct	7 (77.8 %)	5 (55.6 %)	5 (55.6 %)	7 (77.8 %)
Incorrect	2 (22.2 %)	4 (44.4 %)	4 (44.4 %)	2 (22.2 %)
9. Dental technology regulations and ethics (n = 20)				
Correct	14 (70.0 %)	18 (90.0 %)	17 (85.0 %)	17 (85.0 %)
Incorrect	6 (30.0 %)	2 (10.0 %)	3 (15.0 %)	3 (15.0 %)
Total (n = 194)				
Correct	135 (69.6 %)	141 (72.7 %)	140 (72.2 %)	136 (70.1 %)
Incorrect	59 (30.4 %)	53 (27.3 %)	54 (27.8 %)	58 (29.9 %)

examinations, national dental technician licensing examinations, and dental specialty examinations to explore their potential as the educational support tools. For the exam questions in the national dental licensing examination, Lin et al.[8] tested the performance of ChatGPT-3.5, Gemini, and Claude2 on correctly answering the 2699 text-based exam questions (spanning 8 subjects in basic dentistry and 12 in clinical dentistry) selected from the Taiwan national dental licensing examinations. They found that Claude2 acquires the highest overall AR (54.89 %), outperforming ChatGPT-3.5 (49.33 %) and Gemini (44.63 %). In the basic dentistry domain, Claude2 scores 59.73 %, followed by ChatGPT-3.5 (54.87 %) and Gemini (47.35 %). In the clinical dentistry domain, Claude2 obtains an AR of 52.45 %, surpassing ChatGPT-3.5 (46.54 %) and Gemini (43.26 %). However, none of the three LLMs attain passing scores of 60 %.[8]

Wu et al.[9] evaluated the performance of ChatGPT-4, ChatGPT-4o without a prompt, and ChatGPT-4o with a prompt (answering the questions referring to two oral pathology textbooks) in answering the 280 oral pathology MCQs selected from the Taiwan national dental licensing examinations from 2015 to 2024. They discovered that ChatGPT-4o with a prompt achieves the highest overall AR of 90.0 %, slightly outperforming ChatGPT-4o without a prompt (88.6 %) and significantly exceeding ChatGPT-4 (79.6 %, $P < 0.001$). If the 39 image-based questions, 39 case-based questions, and 92 odd-one-out questions are selected for testing. There is a significant difference in the AR of odd-one-out questions between ChatGPT-4 (77.2 %) and ChatGPT-4o without a prompt (91.3 %, $P = 0.015$) or ChatGPT-4o with a prompt (92.4 %, $P = 0.008$). However, there is no significant difference in the AR among three different models when answering the image-based and case-based questions.[9]

Mine et al.[10] used 4 multimodal LLMs, ChatGPT-4o, OpenAI o1, Claude 3.5 Sonnet (Sonnet), and Gemini 2.0 Flash Thinking Experimental (Gemini) to answer the 353 exam questions from the 2024 Japanese national dental licensing examination, including the 204 text-only and 149 visually-based questions. A zero-shot approach was used without prompt engineering. They showed that OpenAI o1 achieves the highest overall AR (81.9 %), followed by Sonnet (71.7 %), Gemini (66.6 %), and ChatGPT-4o (65.7 %). Moreover, all the 4 LLMs perform significantly better on the text-

Table 5 Comparisons of the initial, post-one week, post-two week, and post-three week overall accuracy rates (ARs) for the three large language models (LLMs) of ChatGPT-4, Gemini, and DeepSeek-V3 in answering the 68 text-based exam questions of 4 basic subjects of dental technology (oral anatomy and physiology, dental morphology, dental materials science, and dental technology regulations and ethics) and the 126 text-based exam questions of 5 clinical subjects of dental technology (fixed prosthodontics technology, complete denture technology, removable partial denture technology, orthodontic technology, and pediatric dental technology) selected from the 2023 Taiwan national dental technician licensing examination.

LLM	Number of question with correct answer (AR)			
	Initial	Post-one week	Post-two weeks	Post-three weeks
ChatGPT-4				
Basic subjects (n = 68)	42 (61.8 %)	44 (64.7 %)	44 (64.7 %)	39 (57.4 %)
Clinical subjects (n = 126)	59 (46.8 %)	65 (51.6 %)	65 (51.6 %)	62 (49.2 %)
**P*-value	0.066	0.108	0.108	0.351
Gemini				
Basic subjects (n = 68)	40 (58.8 %)	49 (72.1 %)	45 (66.2 %)	48 (70.6 %)
Clinical subjects (n = 126)	71 (56.4 %)	81 (64.3 %)	83 (65.9 %)	82 (65.1 %)
**P*-value	0.857	0.348	0.907	0.536
DeepSeek-V3				
Basic subjects (n = 68)	48 (70.6 %)	51 (75.0 %)	51 (75.0 %)	51 (75.0 %)
Clinical subjects (n = 126)	87 (69.1 %)	90 (71.4 %)	89 (70.6 %)	85 (67.5 %)
**P*-value	0.953	0.716	0.632	0.352

**P*-value Comparison of the initial, post-one week, post-two week, and post-three week overall ARs between the 68 exam questions of 4 basic subjects of dental technology and the 126 exam questions of 5 clinical subjects of dental technology.

only questions (ARs of 79.9—92.2 %) than on the visually-based questions (ARs of 45.6—67.8 %).[10]

Morishita et al.[11] assessed the efficacy of ChatGPT-4V with image recognition capabilities on answering the 160 image-based questions from the Japanese national dental licensing examination to explore its potential as an educational support tool for the dental students. They demonstrated that the overall AR of ChatGPT-4V for the 160 image-based questions is 35.0 %. Moreover, the ARs are 57.1 % for the 7 compulsory questions, 43.6 % for the 55 general questions, and 28.6 % for the 98 clinical practical questions. In specialties like dental anesthesiology and endodontics, ChatGPT-4V achieves the ARs above 70 %, while the ARs for orthodontics and oral surgery are below 40 %. A higher number of images in questions is correlated with lower accuracy, suggesting an impact of the number of images on obtaining the correct responses.[11]

Fukuda et al.[12] also compared the performance of two LLMs with the image recognition capabilities, ChatGPT-4V and Gemini Pro, on 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. In addition, across dental specialties, the ARs of ChatGPT-4V are generally higher than those of Gemini Pro, with some areas showing equal accuracy. Furthermore, they also observed the fact 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.[12]

For the exam questions in the dental specialty examination, Tosun and Yilmaz[13] evaluated the performance of 7 AI-based chatbots (ChatGPT-3.5, ChatGPT-4, Gemini, Gemini Advanced, Claude AI, Microsoft Copilot, and Smodin AI) on correctly answering the 128 multiple-choice prosthodontics questions from the dental specialty examination in Turkey. They discovered that ChatGPT-4 achieves the highest AR (75.8 %), while Gemini acquires the lowest AR (46.1 %). ChatGPT-4 and ChatGPT-3.5 show significantly higher ARs in the knowledge-based questions than in the case-based questions ($P < 0.05$). For the case-based questions, both Gemini and Gemini Advanced have the lowest AR (36.4 %), while other chatbots achieve a mean AR of 45.5 %. In the knowledge-based questions, ChatGPT-4 performs the best (78.6 %) and Gemini the worst (47 %).[13]

Diniz-Freitas and Diz-Dios[7] evaluated the performance of DeepSeek-R1 on diagnosing oral diseases and conditions using the text-based case descriptions from the New England Journal of Medicine's "Image Challenge." They found that DeepSeek-R1 achieves a diagnostic accuracy of 91.6 %, slightly outperforming ChatGPT-4o (88.9 %) and significantly exceeding the 47.8 % accuracy of the journal' readers.[7]

For the exam questions in the national dental hygienist licensing examination, Yamaguchi et al.[14] tested the efficacy of 4 LLMs (ChatGPT-3.5, ChatGPT-4, Google's Bard, and Microsoft's Bing Chat) on answering the 73 text-based questions from the 32nd Japanese national dental hygienist licensing examination. They demonstrated that ChatGPT-4 achieves the highest AR (75.3 %), followed by Microsoft's Bing Chat (68.5 %), Google's Bard (66.7 %), and ChatGPT-3.5 (63.0 %). There are no statistically significant differences in the acquired ARs among the 4 LLMs. Moreover, the performance varies across different question categories, with all models excelling in the "Disease mechanism and promotion of recovery process" category (100 % AR). ChatGPT-4 generally outperforms other LLMs, especially in the multi-answer questions.[14]

For the exam questions in the national dental technician licensing examination, Mine et al.[15] used 3 LLMs, ChatGPT-4o, OpenAI o1, and Claude 3.5 Sonnet, to answer the 240 exam questions from the Japanese national dental technician licensing examination (JNDTLE), including the 139 text-only questions and the 101 visually-based questions. The overall ARs for a total of the 240 questions, the 139 text-only questions, and the 101 visually-based questions are 58.3 %, 68.3 %, and 44.6 %, respectively for ChatGPT-4o, 67.5 %, 79.1 %, and 51.5 %, respectively for OpenAI o1, 64.6 %, 71.2 %, and 55.4 %, respectively for Claude 3.5 Sonnet. These results suggest that OpenAI o1 has better performance than Claude 3.5 Sonnet and ChatGPT-4o. In addition, OpenAI o1 has the significantly higher overall ARs than those of the ChatGPT-4o on answering a total of the 240 questions and the 139 text-only questions, respectively (the *P*-values are 0.019 and 0.017, respectively). This finding also indicates that ChatGPT-4o has a significantly higher overall AR (68.3 %) on answering the 139 text-only questions from JNDTLE than the average overall AR (54.2 %) of ChatGPT-4 on answering the 194 text-only questions from TNDTLE in this study ($P = 0.012$ by chi-square test).[15]

In summary, this study highlights that while all three LLMs (ChatGPT-4, Gemini, and DeepSeek-V3) can provide moderately accurate responses to dental technology exam questions, they differ in baseline performance and responsiveness over time. DeepSeek-V3 offers the highest initial accuracy but limited adaptive capacity. Gemini shows evidence of short-term improvement, suggesting limited self-learning potential. ChatGPT-4 remains consistent but demonstrates minimal change. These findings emphasize the importance of continual evaluation of LLMs in educational and professional settings and suggest that improvements in the medical or dental domain-specific performance may require targeted fine-tuning or integration of domain-specific updates rather than relying solely on generalized AI training.

Declaration of competing interest

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

Acknowledgments

None.

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. Takagi S, Watan T, Erabi A. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination. *JMIR Med Educ* 2023;9:e48002.
5. Alhur A. Redefining healthcare with artificial intelligence (AI): the contribution of ChatGPT, Gemini, and Co-pilot. *Cureus* 2024;16:e57795.
6. Normile D. Chinese firm's large language model makes a splash. *Science* 2025;387:238.
7. Diniz-Freitas M, Diz-Dios P. DeepSeek: another step forward in the diagnosis of oral lesions. *J Dent Sci* 2025;20:1904—7.
8. Lin CCC, Sun JS, Chang CH, Chang YH, Chang JZC. Performance of artificial intelligence chatbots in national dental licensing examination. *J Dent Sci* 2025;20:2307—14.
9. 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.
10. Mine Y, Okazaki S, Taji T, Kawaguchi H, Kakimoto N, Murayama T. Benchmarking multimodal large language models on the dental licensing examination: challenges with clinical image interpretation. *J Dent Sci* 2025;20:2427—35.
11. 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.
12. 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.
13. 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.
14. Yamaguchi S, Morishita M, Fukuda H, et al. Evaluating the efficacy of leading large language models in the Japanese national dental hygienist examination: a comparative analysis of ChatGPT, Bard, and Bing Chat. *J Dent Sci* 2024;19:2262—7.
15. Mine Y, Taji T, Okazaki S, et al. Analyzing the performance of multimodal large language models on visually-based questions in the Japanese National Examination for Dental Technicians. *J Dent Sci* 2025;20:2460—6.