Journal of Dental Sciences 20 (2025) 2307–2314

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Original Article

Performance of artificial intelligence chatbots in National dental licensing examination

Chad Chan-Chia Lin [a,b,c†], Jui-Sheng Sun [d,e†], Chin-Hao Chang [f], Yu-Han Chang [g], Jenny Zwei-Chieng Chang [a,h*]

[a] School of Dentistry, College of Medicine, National Taiwan University, Taipei, Taiwan
[b] Department of Dentistry, National Taiwan University Hospital Yunlin Branch, Yunlin County, Taiwan
[c] Department of Dentistry, Da Chien Health Medical System, Miaoli County, Taiwan
[d] Department of Orthopedic Surgery, Landseed International Hospital, Taoyuan City, Taiwan
[e] Department of Orthopedic Surgery, National Taiwan University Hospital, Taipei, Taiwan
[f] Department of Medical Research, National Taiwan University Hospital, Taipei, Taiwan
[g] Clinical Trial Center, National Taiwan University Hospital, Taipei, Taiwan
[h] Department of Dentistry, National Taiwan University Hospital, Taipei, Taiwan

Received 24 April 2025; Final revision received 14 May 2025
Available online 27 May 2025

KEYWORDS
Artificial intelligence;
Dentistry;
Large language models;
Dental education;
Chatbot

Abstract *Background/purpose:* The Taiwan dental board exams comprehensively assess dental candidates across twenty distinct subjects, spanning foundational knowledge to clinical fields, using multiple-choice single-answer exams with a minimum passing score of 60 %. This study assesses the performance of artificial intelligence (AI)-powered chatbots (specifically ChatGPT3.5, Gemini, and Claude2), categorized as Large Language Models (LLMs), on these exams from 2021 to 2023.

Materials and methods: A total of 2699 multiple-choice questions spanning eight subjects in basic dentistry and twelve in clinical dentistry were analyzed. Questions involving images and tables were excluded. Statistical analyses were conducted using McNemar's test. Furthermore, annual results of LLMs were compared with the qualification rates of human candidates to provide additional context.

Results: Claude2 demonstrated the highest overall accuracy (54.89 %) on the Taiwan national dental licensing examinations, outperforming ChatGPT3.5 (49.33 %) and Gemini (44.63 %), with statistically significant differences in performance across models. In the basic dentistry domain, Claude2 scored 59.73 %, followed by ChatGPT3.5 (54.87 %) and Gemini (47.35 %). Notably, Claude2 excelled in biochemistry (73.81 %) and oral microbiology (88.89 %), while

* Corresponding author. School of Dentistry, College of Medicine, National Taiwan University and Department of Dentistry, National Taiwan University Hospital, No 1 Chang-De Street, Taipei 10048, Taiwan.
E-mail address: jennyzc@ms3.hinet.net (J. Zwei-Chieng Chang).
† Contribute equally as the first authors in this work.

https://doi.org/10.1016/j.jds.2025.05.012

ChatGPT3.5 also performed strongly in oral microbiology (80.56 %). In the clinical dentistry domain, Claude2 led with a score of 52.45 %, surpassing ChatGPT3.5 (46.54 %) and Gemini (43.26 %), and showed strong results in dental public health (65.81 %). Despite these achievements, none of the LLMs attained passing scores overall.

Conclusion: None of the models achieved passing scores, highlighting their strengths in foundational knowledge but limitations in clinical reasoning.

Introduction

Artificial intelligence (AI) has seen rapid advancements, with Natural Language Processing (NLP) emerging as a pivotal branch that enables computers to comprehend and generate human language. NLP drives innovations in customer support, language learning, and healthcare applications. Among its most groundbreaking developments are Large Language Models (LLMs), which analyze vast text datasets and leverage billions to trillions of parameters to grasp the intricacies of language. These LLMs, including AI-powered chatbots, are trained on extensive datasets using significant computational resources, enabling them to predict words and sentences with remarkable precision. This capability makes LLMs indispensable across a wide range of applications, significantly enhancing NLP's utility in real-world scenarios. Notable examples of LLMs include chatbots such as ChatGPT (OpenAI, San Francisco, USA), Gemini (Google Inc., Mountainview, USA), and Claude (Anthropic, San Francisco, USA).[1] These chatbots are revolutionizing fields like education, healthcare, business, and creative industries, showcasing their adaptability and transformative impact in the digital era.

LLMs like ChatGPT are being rapidly adopted in medicine, transforming patient education, clinical support, and academic research.[1] For patient communication, ChatGPT has been shown to effectively answer cardiology queries,[2] explain radiology report findings,[3] and translate medical information to improve patient understanding.[4] In clinical workflows, LLMs have supported tasks such as advising on breast cancer management during tumor boards,[5] drafting patient discharge summaries,[6] and generating radiology templates.[3] In education and research, LLMs have demonstrated capabilities like solving complex medical physics problems,[7] achieving passing scores on licensing exams,[8] and generating study materials.[9] They are also used for brainstorming research topics and drafting publications.[10,11] These findings emphasize the transformative role of LLMs in modern medicine, highlighting their potential to improve health literacy, streamline administrative tasks, and support academic and clinical advancements. However, achieving their safe and responsible integration requires rigorous evaluation and careful consideration of ethical implications.

The rapid growth of research focusing on large LLMs is evident, as a PubMed search using "ChatGPT" as a keyword now yields over 5000 publications. This significant number underscores the increasing academic interest in evaluating and understanding the applications and performance of these advanced AI systems across diverse domains. Concerns have been raised about inaccuracies in professional content, biased outputs, and the dissemination of misinformation when using LLMs. Some educational institutions have developed guidelines to regulate the use of LLMs in report and paper writing; moreover, the increase in online licensing exams since the COVID-19 pandemic has raised apprehension about the potential misuse of these models for unethical practices, such as cheating.[12] To mitigate the risk of inaccurate medical advice or the spread of false information, evaluating LLMs' performance on standardized test has emerged as a practical metric. This method has been applied across various medical specialties to assess the reliability and effectiveness of these models.[13]

Studies have highlighted the potential of LLMs in medical licensing exams, with ChatGPT4 achieving passing-level accuracy on the Japanese Medical Licensing Examination (82.7 % in essential knowledge and 77.2 % in general clinical questions)[12] and exceeding 80 % accuracy on the 2023 Peruvian National Licensing Medical Examination.[14] ChatGPT also scored over 60 % in the United States Medical Licensing Examination (USMLE), aligning with third-year medical student performance.[15] However, for the Japanese Dental Society of Anesthesiology Basic Competency Examination, LLMs underperformed (<60 % accuracy), likely due to limited online information, lack of tailored prompt engineering, and ambiguous question phrasing.[16] This highlights the need for adaptation in specialized contexts. Integrating AI technologies into dental education is essential for modernizing curricula and raising clinician awareness of potential benefits and challenges. Potential LLM applications in dentistry include teledentistry, clinical decision-making, administrative tasks, patient education, and training programs.[17] However, there is a notable lack of research evaluating the performance of different LLMs in dental examinations.[16,18–20]

The Taiwan national dental licensing examinations are a two-stage assessment designed to evaluate the clinical competence of dental candidates. Successful completion of both stages is necessary for licensure and independent practice in Taiwan. The first-stage exam, basic dentistry (Stage 1), is open to those who have completed the first four years of foundational dental coursework. It includes two written exams: Dentistry-I and Dentistry-II. Dentistry-I covers basic topics like oral anatomy, dental morphology, oral histology, embryology, and biochemistry. Dentistry-II focuses on oral pathology, dental materials, oral microbiology, and dental pharmacology. The second-stage exam, clinical dentistry (Stage 2), requires candidates to have

completed clinical internships and passed Stage 1. This stage, typically taken after the sixth year of dental school, comprises four written exams: Dentistry-III to VI. Dentistry-III includes endodontics, operative dentistry, and periodontology. Dentistry-IV covers oral and maxillofacial surgery and dental radiology. Dentistry-V addresses prosthodontics, including complete and partial dentures, fixed prosthodontics, and occlusion. Dentistry-VI encompasses orthodontics, pediatric dentistry, and dental public health. Each exam consists of 80 multiple-choice questions, with a minimum passing score of 60 %. Candidates who fail the second-stage examination within six years must retake the first-stage examination, ensuring the rigor and comprehensive nature of the examination process. Given its comprehensive scope, the Taiwan national dental licensing examinations offer an ideal framework for evaluating the performance of LLMs in the dental field. The aim of this study was to compare the performance of AI-powered chatbots, ChatGPT (GPT3.5), Gemini, and Claude2 on the Taiwan dental board exams from 2021 to 2023.

Materials and methods

The Taiwan national dental licensing examinations are conducted in two stages, with all questions presented in a multiple-choice single-answer format. Stage 1 includes two test papers, each comprising 80 questions that cover eight subjects. This stage is administered twice a year, typically during February and July to align with winter and summer vacations. Stage 2 consists of four test papers, each containing 80 questions across 12 subjects, and follows the same biannual schedule. Each national examination includes a total of 480 questions, amounting to 960 questions annually. For this study, all exams conducted from 2021 to 2023 were analyzed, resulting in the systematic collection of 2880 multiple-choice questions.

To account for the role of prompt engineering in influencing generative LLM outputs, input formats of the data sets were standardized. Questions containing images and tables were excluded to ensure the focus remained on the LLM's ability to generate responses based on narrative medical knowledge without requiring complex parsing. As illustrated in Fig. 1, the data collection process identified 905 questions from basic dentistry (Stage 1) and 1794 from clinical dentistry (Stage 2). These questions were used to evaluate three AI language models (ChatGPT3.5, Gemini, and Claude2) covering all eight subjects in basic dentistry and 12 subjects in clinical dentistry. To ensure consistency in testing, each question was reformatted to start with a direct inquiry, followed by the question text, with single-choice answers listed individually on separate lines. This standardized approach streamlined the testing process and ensured uniform evaluation across all models. An example of a formatted question alongside an LLM-generated response is provided in Fig. 2.

The performance of ChatGPT3.5, Gemini, and Claude2 was evaluated across the entire dataset. Accuracy for each model was directly calculated, with statistical analysis performed using McNemar's test for paired data. All computations were conducted using SPSS version 15.0 (IBM Corporation, Armonk, NY, USA), and results were reported as mean ± standard deviation (SD). A *P*-value of less than 0.05 was considered statistically significant. Furthermore, the annual results of LLMs for both stages of the exams were compared with the qualification (passing) rates of human candidates to provide additional context. As this study relied exclusively on publicly available data sources, no institutional review board approval was required.

Results

Performance of ChatGPT3.5, Gemini, Claude2 on Taiwan national dental licensing examinations

Fig. 3 and Table 1 present the performance of the LLMs on the Taiwan national dental licensing examinations. Overall,

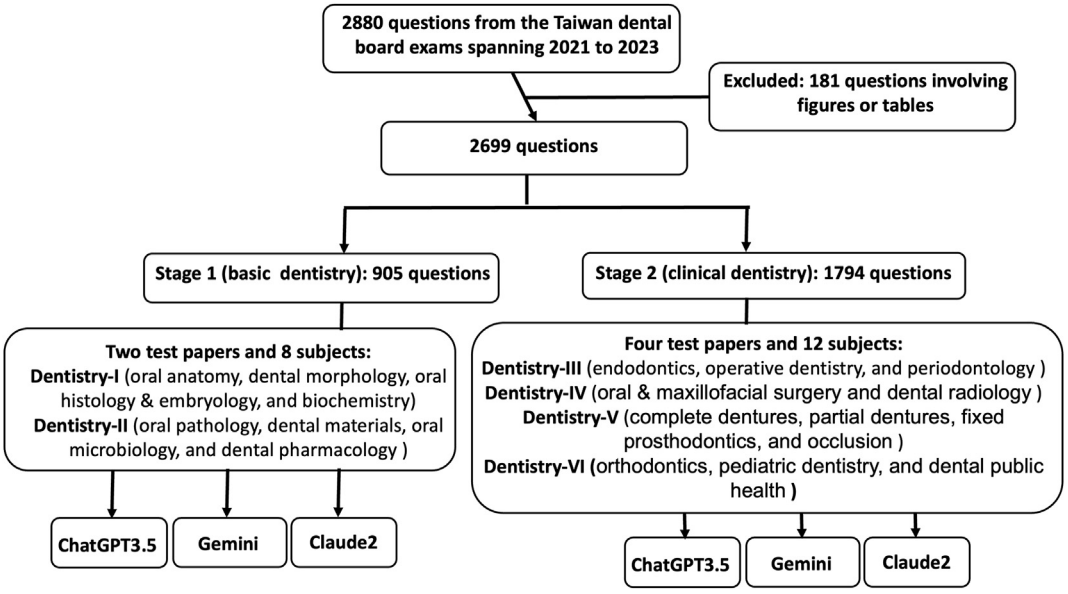

Figure 1 Flowchart of study design and setting.

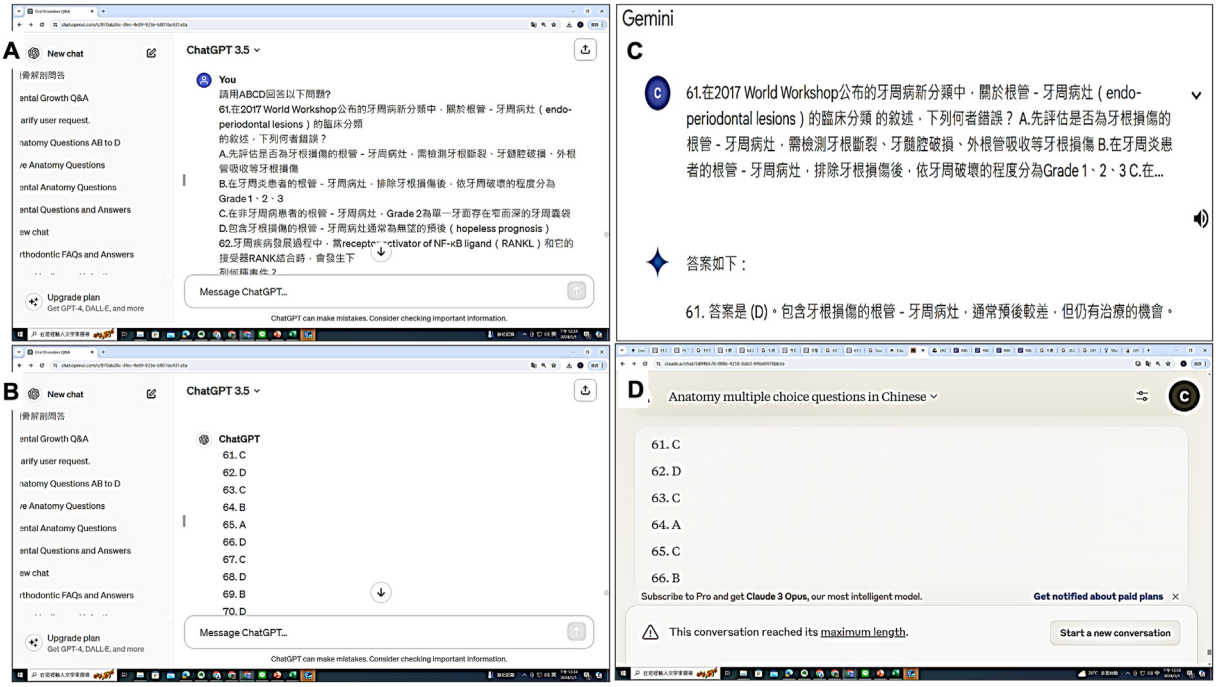

Figure 2 Examples of question posed to each large language model (LLM), including the response from Chat Generative Pre-trained Transformer 3.5 (ChatGPT3.5, A and B), Gemini (C), and Claude2 (D).

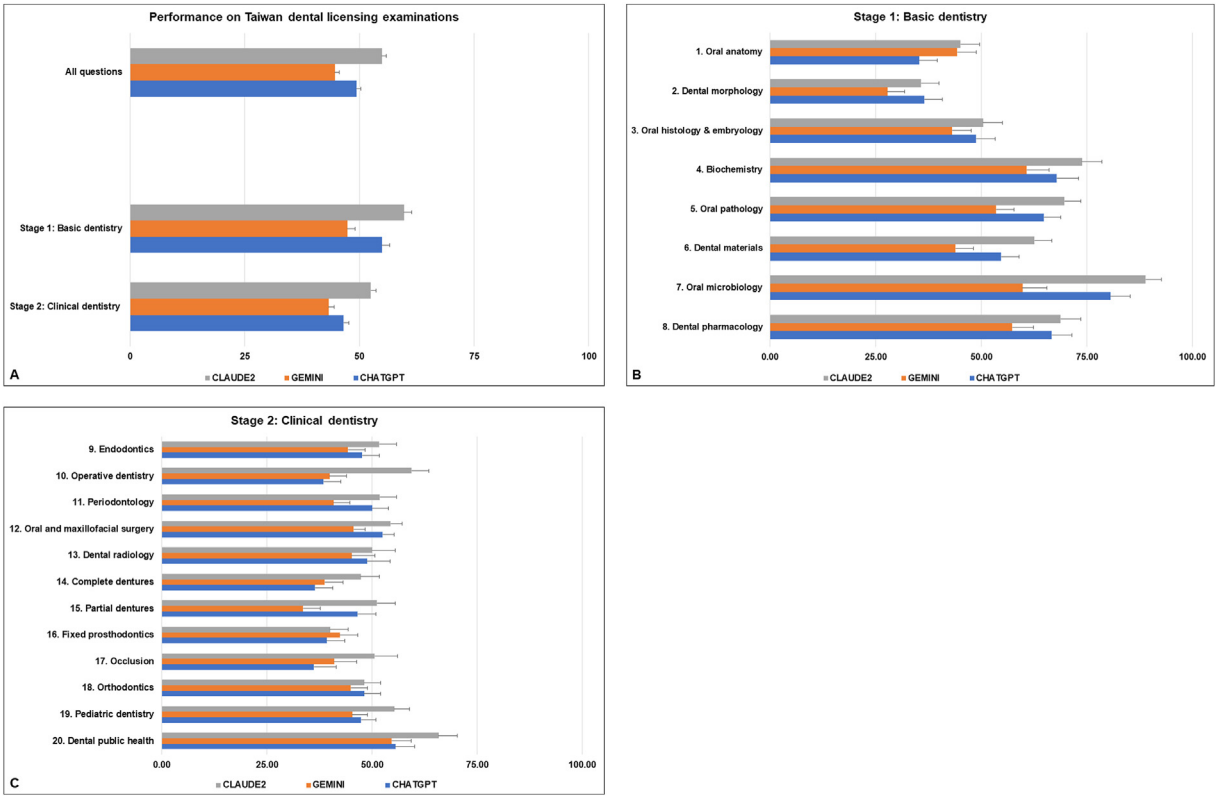

Figure 3 Performance of ChatGPT3.5, Gemini, Claude2 on the Taiwan dental licensing examinations (A) Performance on Taiwan dental licensing examinations; Stage 2: Clinical dentistry; Stage 1: Basic dentistry. (B) Stage 1: Basic dentistry. (C) Stage 2: Clinical dentistry.

Table 1 Performance analysis of large language models (LLMs) on questions from the Taiwan dental board exams.

		Question type	Number of questions	Performance (correct rate: %)			*P* value		
				ChatGPT3.5	Gemini	Claude2	ChatGPT3.5 vs. Gemini	ChatGPT3.5 vs. Claude2	Gemini vs. Claude2
				Mean ± standard deviation					
		All questions	2699	49.33 ± 0.963	44.63 ± 0.957	54.89 ± 0.958	<0.0001[a]	<0.0001[a]	<0.0001[a]
		Stage 1. Basic dentistry	905	54.87 ± 1.655	47.35 ± 1.661	59.73 ± 1.631	0.0001[a]	0.0098[a]	<0.0001[a]
		Stage 2. Clinical dentistry	1794	46.54 ± 1.178	43.26 ± 1.17	52.45 ± 1.179	0.019[a]	<0.0001[a]	<0.0001[a]
Stage 1	Dentistry-I	1. Oral anatomy	122	35.25 ± 4.325	44.26 ± 4.497	45.08 ± 4.505	0.1011	0.0578	0.8759
		2. Dental morphology	126	36.51 ± 4.289	27.78 ± 3.99	35.71 ± 4.269	0.1011	0.8886	0.1573
		3. Oral histology & embryology	123	48.78 ± 4.507	43.09 ± 4.465	50.41 ± 4.508	0.2743	0.7518	0.1699
		4. Biochemistry	84	67.86 ± 5.096	60.71 ± 5.329	73.81 ± 4.797	0.3304	0.2971	0.0411[a]
	Dentistry-II	5. Oral pathology	143	64.79 ± 4.008	53.52 ± 4.185	69.72 ± 3.856	0.0209[a]	0.2858	0.001[a]
		6. Dental materials	139	54.68 ± 4.222	43.88 ± 4.209	62.59 ± 4.104	0.0253[a]	0.1235	0.0008[a]
		7. Oral microbiology	72	80.56 ± 4.664	59.72 ± 5.78	88.89 ± 3.704	0.0018[a]	0.1088	<0.0001[a]
		8. Dental pharmacology	96	66.67 ± 4.811	57.29 ± 5.049	68.75 ± 4.731	0.1172	0.7055	0.063
Stage 2	Dentistry-III	9. Endodontics	149	47.65 ± 4.092	44.3 ± 4.069	51.68 ± 4.094	0.4751	0.3545	0.1235
		10. Operative dentistry	143	38.46 ± 4.068	39.86 ± 4.094	59.44 ± 4.106	0.7576	<0.0001[a]	0.0001[a]
		11. Periodontology	164	50 ± 3.904	40.85 ± 3.838	51.83 ± 3.902	0.0508	0.6803	0.0143[a]
	Dentistry-IV	12. Oral & maxillofacial surgery	318	52.52 ± 2.8	45.6 ± 2.793	54.4 ± 2.793	0.0278[a]	0.5708	0.0065[a]
		13. Dental radiology	82	48.78 ± 5.52	45.12 ± 5.495	50 ± 5.522	0.6015	0.8415	0.4927
	Dentistry-V	14. Complete dentures	129	36.43 ± 4.237	38.76 ± 4.29	47.29 ± 4.396	0.6547	0.0522	0.1308
		15. Partial dentures	131	46.56 ± 4.358	33.59 ± 4.126	51.15 ± 4.367	0.0269[a]	0.4308	0.0016[a]
		16. Fixed prosthodontics	130	39.23 ± 4.282	42.31 ± 4.333	40 ± 4.297	0.5862	0.884	0.6473
		17. Occlusion	83	36.14 ± 5.273	40.96 ± 5.398	50.6 ± 5.488	0.4652	0.0285[a]	0.1306
	Dentistry-VI	18. Orthodontics	158	48.1 ± 3.975	44.94 ± 3.957	48.1 ± 3.975	0.5351	1	0.5472
		19. Pediatric dentistry	190	47.37 ± 3.622	45.26 ± 3.611	55.26 ± 3.607	0.6115	0.071[a]	0.0241[a]
		20. Dental public health	117	55.56 ± 4.594	54.7 ± 4.602	65.81 ± 4.385	0.8658	0.0396[a]	0.0326[a]

[a] A *P*-value of less than 0.05 was considered statistically significant.

Claude2 achieved the highest accuracy (54.89 %), outperforming ChatGPT3.5 (49.33 %) and Gemini (44.63 %). In the Stage 1 examination, Claude2 achieved a correct rate of 59.73 %, compared to 54.87 % for ChatGPT3.5 and 47.35 % for Gemini. For the Stage 2 examination, Claude2 scored 52.45 %, ChatGPT3.5 46.54 %, and Gemini 43.26 %. None of the models passed the board exam.

In the Stage 1 examination, all three LLMs demonstrated notable accuracy in biochemistry, with Claude2 achieving the highest score of 73.81 %. Claude2 and ChatGPT3.5 also showed strong performance in oral pathology, oral microbiology, and dental pharmacology. In oral microbiology, Claude2 excelled with an accuracy of 88.89 %, while ChatGPT3.5 also performed well with 80.56 %. However, only Claude2 exceeded 60 % accuracy in dental materials, achieving a score of 62.59 %. In the Stage 2 examination, Claude2 achieved the highest accuracy in dental public health at 65.81 %, while none of the models reached 60 % accuracy across other subjects. Statistically significant differences were observed among the models in overall accuracy, performance across examination stages, and subject-specific results.

Yearly performance comparison of ChatGPT3.5, Gemini, and Claude2 against human qualification rates on the Taiwan national dental licensing examinations

Table 2 summarizes the yearly performance (2021—2023) of LLMs on the Taiwan dental board exams, compared to human qualification rates. In the February 2022 Stage 1 (basic dentistry) exam, where the human pass rate was 49.57 %, Claude2 achieved a passing score, indicating ranking in the top 50.43 % of examinees (percentile rank of 50.43 (PR50.43) or higher). Conversely, in the July 2022 Stage 2 (clinical dentistry) exam, where the human pass rate was 92.72 %, all LLMs failed to meet the passing threshold, placing them in the bottom 7.28 % (PR7.28 or lower), highlighting a significant gap in their performance on clinically focused assessments.

Discussion

LLMs have become integral in healthcare, aiding patients in seeking health information and supporting professionals with research and clinical decision-making. Despite these benefits, significant risks persist, including the potential for inaccurate recommendations to patients and misinformation for clinicians. Such concerns are particularly relevant in academic settings where AI-driven training and testing systems are under development. The performance of LLMs in clinical board-style examinations has been explored to some extent, but their strengths and limitations remain insufficiently understood. Systematic review and meta-analyses show that LLMs achieve an overall accuracy of 61 % on medical examinations, with an average accuracy of 51 % on USMLE.[21] While most studies focus primarily on ChatGPT, comparative analyses involving multiple LLMs are relatively rare. Previous research has demonstrated the ability of individual LLMs to pass certain medical licensing exams; however, limited attention has been given to evaluating their relative performance or their application in dental contexts. This study evaluated the performance of three leading LLMs (OpenAI's GPT3.5, Google's Gemini, and Anthropic's Claude2) on the Taiwan national dental licensing examinations. A total of 2699 multiple-choice questions were posed to each LLM, making it one of the most extensive analyses of dental examinations. The evaluation covered 20 distinct subjects, including oral anatomy, dental morphology, oral histology and embryology, biochemistry, oral pathology, dental materials, oral microbiology, dental pharmacology, endodontics, operative dentistry, periodontology, oral and maxillofacial surgery, dental radiology, complete dentures, partial dentures, fixed prosthodontics, occlusion, orthodontics, pediatric dentistry, and dental public health. The results indicate that Claude2 achieved the highest accuracy (54.89 %),

Table 2 Yearly performance (2021—2023) of large language models (LLMs) on questions from the Taiwan dental board exams compared to human qualification rates.

Performance (correct rate)	ChatGPT3.5	Gemini	Claude2	Human qualification rate
2021 February stage 1	58.07 %	55.83 %	61.39 %	n/a
2021 February stage 2	45.68 %	47.19 %	54.23 %	72.81 %
2021 July stage 1	58.70 %	53.38 %	68.45 %	n/a
2021 July stage 2	50.07 %	48.00 %	58.00 %	93.21 %
2022 February stage 1	53.67 %	41.43 %	63.14 %	49.57 %
2022 February stage 2	48.93 %	39.17 %	57.51 %	52.00 %
2022 July stage 1	58.37 %	50.51 %	59.20 %	59.82 %
2022 July stage 2	40.03 %	48.83 %	50.87 %	92.72 %
2023 February stage 1	54.56 %	46.74 %	55.61 %	49.82 %
2023 February stage 2	42.34 %	37.47 %	45.36 %	37.50 %
2023 July stage 1	55.08 %	47.49 %	63.44 %	46.93 %
2023 July stage 2	45.02 %	37.22 %	45.70 %	88.21 %

n/a: Information not available on the website of the Ministry of Examination.

outperforming ChatGPT3.5 (49.33 %) and Gemini (44.63 %). However, none of the models met the passing criteria for the board exam.

Brozović et al. evaluated Bing Chat's ability to answer 532 multiple-choice questions covering a range of dental disciplines, including dental pre-clinics, operative dentistry and endodontics, oral surgery, periodontology, pediatric dentistry, prosthodontics, oral medicine and implant dentistry. These questions were drawn from exams for 2nd to 6th-year students at the Osijek Faculty of Dental Medicine and Health, where Bing Chat achieved a score of 71.99 %, exceeding the 60 % passing threshold.[18] Danesh et al. assessed ChatGPT's capabilities using 143 text-based multiple-choice dental board questions from the Integrated National Board Dental Examination Bootcamp, ITDOnline, and board-style questions provided by the Joint Commission on National Dental Examinations. ChatGPT3.5 achieved an average accuracy of 61.3 %, while ChatGPT4 significantly outperformed it with a score of 76.9 %.[20] Despite these encouraging results, Brozović et al. focused solely on Bing Chat without comparing other LLMs and did not provide a subject-specific analysis, while Danesh et al. compared only ChatGPT versions, similarly neglecting other LLMs and detailed breakdowns by specific dental topics. Fujimoto et al. evaluated the performance of three LLMs (ChatGPT4, Gemini, and Claude3), using 295 text-based multiple-choice questions from the 2020 to 2022 Japanese Dental Society of Anesthesiology Board Certification Examination. ChatGPT4 achieved an accuracy of 51.2 %, Claude3 scored 47.4 %, both significantly outperforming Gemini's 30.3 %.[16] These findings align closely with our study, where Claude and ChatGPT outperformed Gemini, despite using different versions. Fujimoto et al.'s study concentrated exclusively on dental anesthesiology and featured two types of questions: selecting 1 to 3 correct answers from five options and selecting all correct answers. In contrast, the Taiwan national dental licensing examination used in our research comprised only multiple-choice single-answer questions. Research has indicated that LLMs often encounter greater difficulty to answer questions requiring multiple correct responses than single-answer ones, although the difference has not consistently reached statistical significance.[18]

Sabri et al. compared the accuracy of ChatGPT (GPT4 and GPT3.5) and Gemini against periodontal residents using 1312 multiple-choice questions from the 2020–2023 American Academy of Periodontology in-service exams. GPT4 achieved the highest accuracy (79.57 %), followed by Gemini (72.86 %), with both surpassing all human resident groups. GPT3.5 (64.93 %), however, only outperformed first-year residents (63.48 %), while second-year and third-year residents scored 66.25 % and 69.06 %, respectively.[19] Sabri et al. further evaluated the LLMs' performance through sub-analyses of their proficiency across ten key exam sections: embryology and anatomy, biochemistry and physiology, microbiology and immunology, periodontal etiology and pathogenesis, pharmacology and therapeutics, biostatistics and experimental design and data analysis, diagnosis, treatment planning and prognosis, therapy, and oral pathology and oral medicine. Consistent with our findings, LLMs exhibited strong performance in biochemistry, oral pathology, oral microbiology, and dental pharmacology. However, while Sabri et al. observed satisfactory results in embryology and anatomy, our study highlighted weaker outcomes in oral anatomy and oral histology and embryology. These observations may indicate that LLMs tend to excel in tasks requiring lower-order cognitive skills but encounter challenges when addressing more complex higher-order cognitive demands.[22]

The primary limitation of this study is the absence of a human control group, which prevents direct performance comparisons between the LLMs and actual examinees. Instead, qualification (pass) rates from corresponding years of the dental licensure examinations were used for contextual reference (Table 2). Another limitation is the evaluation of only three LLMs, despite the rapid evolution and diversity of available models.[21] Our findings showed that LLMs achieved lower pass rates compared to those reported in other dental examination studies,[18–20] though they closely aligned with the results of Fujimoto et al.[16] This discrepancy may arise from differences in the models or versions analyzed. For instance, this study did not include the latest version of GPT4 or other notable models such as Bing Chat. GPT4 consistently outperforms GPT3.5 in multiple studies,[12,19,20] yet its subscription-based access may limit its widespread use by student. Furthermore, GPT4's tendency to provide longer incorrect responses than concise, correct ones increase risks of misinformation.[20] Additionally, the difficulty level of the exams and the broad scope of topics covered may have contributed to the LLMs' suboptimal performance. The Taiwan national dental licensing examinations encompass a wide range of subjects, requiring both foundational knowledge and clinical reasoning, which pose significant challenges for current LLMs. Language barriers and suboptimal prompt engineering further compound these challenges, underscoring the need for improved AI capabilities and refined application strategies in this context.

Within the limitations of this study, the results indicate that Claude2 (54.89 %) outperformed ChatGPT3.5 (49.33 %), while Gemini (44.63 %) demonstrated the weakest performance on the Taiwan national dental licensing examinations. Statistically significant differences were observed among the three AI models in overall performance, as well as across Stage 1 (basic dentistry) and Stage 2 (clinical dentistry) exams, and individual test subjects. Although Claude2 (59.73 %) came close to passing the first-stage examination, none of the AI models achieved a passing score in either Stage 1 or Stage 2. All three LLMs performed better in foundational dental sciences compared to clinical sciences, highlighting their relative proficiency in knowledge-based tasks but limitations in applied clinical reasoning. These findings emphasize the importance of dental students critically evaluating AI-generated outputs and avoiding reliance on these tools without verification. Expanding future research to include a wider range of LLMs could offer greater insights and support the development of AI as a reliable educational resource in dentistry.

LLMs can serve as interactive tools for dental students, providing immediate feedback on practice questions and offering explanations to reinforce conceptual understanding. LLMs can assist educators in generating draft teaching materials, quiz questions, and case scenarios for simulation-based learning. With appropriate oversight, LLMs may help identify gaps in student knowledge or assist

in creating personalized study plans based on student performance analytics. Despite current limitations in clinical judgment, we propose that LLMs could complement traditional dental education by enhancing accessibility to learning resources and promoting active engagement, particularly in foundational science subjects where their performance was stronger.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of supporting data

The datasets used and/or analyzed in the current study are available from the corresponding author on reasonable request.

Funding

None.

Competing interests

The authors declare no conflict of interest.

Acknowledgement

All works were performed at National Taiwan University, National Taiwan University Hospital (Taipei City, Taiwan). The Authors would like to express their thanks to the staff of National Taiwan University Hospital-Statistical Consulting Unit (NTUH-SCU) for statistical consultation and analyses.

References

1. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med* 2024;177:210—20.
2. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA* 2023;329:842—4.
3. Grewal H, Dhillon G, Monga V, et al. Radiology gets chatty: the ChatGPT saga unfolds. *Cureus* 2023;15:e40135.
4. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. *Lancet Digit Health* 2023;5:e179—81.
5. Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer* 2023;9:44.
6. Singh S, Djalilian A, Ali MJ. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. *Semin Ophthalmol* 2023;38:503—7.
7. Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Front Oncol* 2023;13:1219326.
8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198.
9. Cross J, Robinson R, Devaraju S, et al. Transforming medical education: assessing the integration of ChatGPT into faculty workflows at a Caribbean medical school. *Cureus* 2023;15:e41399.
10. Brameier DT, Alnasser AA, Carnino JM, Bhashyam AR, von Keudell AG, Weaver MJ. Artificial intelligence in orthopaedic surgery: can a large language model "write" a believable orthopaedic journal article? *J Bone Joint Surg Am* 2023;105:1388—92.
11. Gupta R, Herzog I, Weisberger J, Chao J, Chaiyasate K, Lee ES. Utilization of ChatGPT for plastic surgery research: friend or foe? *J Plast Reconstr Aesthetic Surg* 2023;80:145—7.
12. Ishida K, Hanada E. Potential of ChatGPT to pass the Japanese medical and healthcare professional national licenses: a literature review. *Cureus* 2024;16:e66324.
13. März M, Himmelbauer M, Boldt K, Oksche A. Legal aspects of generative artificial intelligence and large language models in examinations and theses. *GMS J Med Educ* 2024;41:Doc47.
14. Torres-Zegarra BC, Rios-Garcia W, Nana-Cordova AM, et al. Performance of ChatGPT, Bard, Claude, and bing on the Peruvian national licensing medical examination: a cross-sectional study. *J Educ Eval Health Prof* 2023;20:30.
15. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312.
16. Fujimoto M, Kuroda H, Katayama T, et al. Evaluating large language models in dental anesthesiology: a comparative analysis of ChatGPT-4, Claude 3 Opus, and Gemini 1.0 on the Japanese dental society of anesthesiology board certification exam. *Cureus* 2024;16:e70302.
17. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. *J Esthetic Restor Dent* 2023;35:1098—102.
18. Brozović J, Mikulić B, Tomas M, Juzbašić M, Blašković M. Assessing the performance of Bing Chat artificial intelligence: dental exams, clinical guidelines, and patients' frequent questions. *J Dent* 2024;144:104927.
19. Sabri H, Saleh MHA, Hazrati P, et al. Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education. *J Periodontal Res* 2025;60:121—33.
20. Danesh A, Pazouki H, Danesh K, Danesh F, Danesh A. The performance of artificial intelligence language models in board-style dental knowledge assessment: a preliminary study on ChatGPT. *J Am Dent Assoc* 2023;154:970—4.
21. Waldock WJ, Zhang J, Guni A, Nabeel A, Darzi A, Ashrafian H. The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: systematic review and meta-analysis. *J Med Internet Res* 2024;26:e56532.
22. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology* 2023;307:e230582.