Scientific Research Report

# Accuracy of Large Language Models for Infective Endocarditis Prophylaxis in Dental Procedures

*Paak Rewthamrongsris* [a,#], *Jirayu Burapacheep* [b,#], *Vorapat Trachoo* [c], *Thantrira Porntaveetus* [d*]

[a] *Department of Anatomy, Faculty of Dentistry, Chulalongkorn University, Bangkok, Thailand*
[b] *Stanford University, Stanford, California, USA*
[c] *Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Chulalongkorn University, Bangkok, Thailand*
[d] *Center of Excellence in Genomics and Precision Dentistry, Clinical Research Center, Geriatric Dentistry and Special Patients Care International Program, Department of Physiology, Faculty of Dentistry, Chulalongkorn University, Bangkok, Thailand*

ARTICLE INFO

*Article history:*
Received 25 July 2024
Received in revised form
22 September 2024
Accepted 24 September 2024
Available online 12 October 2024

*Key words:*
Artificial intelligence
ChatGPT
AHA guidelines
Gemini
Claude

ABSTRACT

*Purpose:* Infective endocarditis (IE) is a serious, life-threatening condition requiring antibiotic prophylaxis for high-risk individuals undergoing invasive dental procedures. As LLMs are rapidly adopted by dental professionals for their efficiency and accessibility, assessing their accuracy in answering critical questions about antibiotic prophylaxis for IE prevention is crucial.
*Methods:* Twenty-eight true/false questions based on the 2021 American Heart Association (AHA) guidelines for IE were posed to 7 popular LLMs. Each model underwent five independent runs per question using two prompt strategies: a pre-prompt as an experienced dentist and without a pre-prompt. Inter-model comparisons utilised the Kruskal–Wallis test, followed by post-hoc pairwise comparisons using Prism 10 software.
*Results:* Significant differences in accuracy were observed among the LLMs. All LLMs had a narrower confidence interval with a pre-prompt, and most, except Claude 3 Opus, showed improved performance. GPT-4o had the highest accuracy (80% with a pre-prompt, 78.57% without), followed by Gemini 1.5 Pro (78.57% and 77.86%) and Claude 3 Opus (75.71% and 77.14%). Gemini 1.5 Flash had the lowest accuracy (68.57% and 63.57%). Without a pre-prompt, Gemini 1.5 Flash's accuracy was significantly lower than Claude 3 Opus, Gemini 1.5 Pro, and GPT-4o. With a pre-prompt, Gemini 1.5 Flash and Claude 3.5 were significantly less accurate than Gemini 1.5 Pro and GPT-4o. None of the LLMs met the commonly used benchmark scores. All models provided both correct and incorrect answers randomly, except Claude 3.5 Sonnet with a pre-prompt, which consistently gave incorrect answers to eight questions across five runs.
*Conclusion:* LLMs like GPT-4o show promise for retrieving AHA-IE guideline information, achieving up to 80% accuracy. However, complex medical questions may still pose a challenge. Pre-prompts offer a potential solution, and domain-specific training is essential for optimizing LLM performance in healthcare, especially with the emergence of models with increased token limits.



## Introduction

Large Language Models (LLMs) like Chat- GPT,[1] Claude,[2] and Gemini,[3] represent a rapidly evolving class of artificial intelligence (AI) systems capable of processing and generating human-like text. Trained on extensive datasets, these models exhibit impressive contextual comprehension and capabilities across diverse domains, including healthcare. LLMs hold

* *Corresponding author*. Center of Excellence in Genomics and Precision Dentistry, Faculty of Dentistry, Chulalongkorn University, Bangkok, 10330, Thailand.
E-mail address: thantrira.p@chula.ac.th (T. Porntaveetus).
Paak Rewthamrongsris: http://orcid.org/0009-0004-3678-8554
Jirayu Burapacheep: http://orcid.org/0009-0004-9238-4415
Vorapat Trachoo: http://orcid.org/0000-0002-1478-1122
Thantrira Porntaveetus: http://orcid.org/0000-0003-0145-9801

# Equal contribution.

https://doi.org/10.1016/j.identj.2024.09.033

significant potential to transform dentistry by aiding in differential diagnoses, automating clinical documentation, and advancing research through literature review and idea generation.[4,5] The emergence of AI in dentistry has gained significant traction in answering dentistry-related questions, particularly in areas like dental trauma, oral and maxillofacial surgery, and pediatric dentistry.[6,7] LLMs offer a promising avenue for addressing human variability and supporting dental professionals in rapidly retrieving essential information, a process significantly faster than manual reading, thus promoting adherence to best practices.

Infective endocarditis (IE) is a serious and life-threatening condition, with an in-hospital mortality rate of up to 30%.[8] It can lead to long-term complications such as progressive deterioration of cardiac function that may require cardiac valve replacement. The American Heart Association (AHA) underscores the importance of antibiotic prophylaxis to prevent IE in high-risk individuals undergoing dental procedures, which can introduce bacteria into the bloodstream. The 2021 AHA Scientific Statement on IE prevention updates the 2007 guidelines by removing the recommendation for antibiotic prophylaxis in moderate-risk patients and emphasizing the role of oral hygiene and gingival disease as risk factors. The statement also details characteristics of dental procedures that may increase the risk of bacteremia. Additionally, the 2021 guidelines express concerns about using clindamycin as an alternative to amoxicillin due to the increased risk of *Clostridioides difficile* infection and its serious outcomes, and introduces doxycycline as an option for patients with penicillin allergies.[9-11]

While systematic reviews and meta-analyses suggest that antibiotics may be beneficial in reducing the incidence of IE following invasive dental procedures in high-risk individuals,[12,13] evidence from randomised controlled trials remains limited, and conflicting results from observational studies contribute to ongoing controversy surrounding their use. Despite the ongoing debate, the AHA guidelines remain one of the most widely referenced standards for IE prevention in high-risk individuals. Evaluating the accuracy of LLMs in providing recommendations aligned with these guidelines is especially essential for patients with specific cardiac conditions, where precise, evidence-based guidance is essential for ensuring patient safety.[14] This study aimed to assess the accuracy of seven recent proprietary LLMs in recommending antibiotic prophylaxis for dental procedures as a preventive measure against IE, in alignment with AHA guidelines. The study seeks to illustrate both the capabilities and limitations of these AI tools in supporting clinical decision-making and enhancing treatment outcomes in dental practice.

## Materials and methods

### *Dataset construction*

We developed a set of 28 binaries (true or false) regarding antibiotic prophylaxis before dental procedures to prevent viridans group streptococcal infective endocarditis in JavaScript Object Notation (JSON) format. The questions were formulated based on the 2021 AHA statement on prevention of IE.[11] To ensure content validity, the questions and their corresponding answers underwent expert review by an oral and maxillofacial surgeon who hold a degree in both medicine and dentistry. The questions in this study were selected to address key aspects of IE prophylaxis in dental procedures. They encompass various facets of antibiotic prophylaxis, including indications, contraindications, dosage regimens, antibiotic choices, and timing considerations, reflecting the complex nature of clinical decision-making in this area (Table 1). The sample size was based on previous studies assessing LLMs in healthcare.[6,7]

### *Model selection and querying*

We selected and evaluated seven LLMs, comprising GPT 3.5 Turbo, GPT-4o (OpenAI), Claude 3 Sonnet, Claude 3 Opus, Claude 3.5 Sonnet (Anthropic), and Gemini 1.5 Flash, Gemini 1.5 Pro (Google). To ensure a standardised evaluation framework, we employed a consistent prompting strategy across all models. The system pre-prompt was set to "You are an experienced dentist." The term "an experienced dentist" was interpreted differently by each model. The GPT and Gemini models attempted to role-play, using their knowledge to emulate a professional. However, the Claude models acknowledged their limitations as AI assistants without real-world experience and focused on providing information in a straightforward manner (Supplementary Data S1). For each question, we posed the user prompt, "Is the following statement true or false?" followed by the question text (Supplementary Table 1). To account for the inherent stochasticity in LLM outputs, we conducted five independent runs for each model on every question. This approach enabled more robust assessment of model performance by capturing the variability in responses, allowing for a reliable comparison of different LLMs. Additionally, we conducted another set of evaluations without using system pre-prompt. This allowed us to gain insights into how LLM responses.

All experiments were conducted on June 24, 2024 (evaluations with pre-prompt) and July 11, 2024 (evaluations without pre-prompt). We leveraged the respective APIs of each model provider for consistent and programmatic interaction, ensuring uniformity in data collection across all LLMs evaluated.

### *Response parsing*

We performed a systematic parsing on LLM responses to account for the nuanced and potentially verbose nature of their outputs. Affirmative answers, including "true," "generally true," or "partially true," were classified as true. Negative answers, such as "false" or equivalent negations, were classified as false. Responses indicating uncertainty or inability to answer were categorised as incorrect.

### *Statistical analysis*

Model predictions were evaluated against predetermined correct answers and categorised as correct or incorrect. Statistical analysis was conducted using Prism10 version 10.1.0 (GraphPad Software). We computed the percentage of average accuracy across all running using the formula (mean of correct answers / total number of questions) $\times$ 100, standard deviation (SD), and

**Table 1 – Questions used to evaluate accuracy of responses from various LLMs.**

| | |
|---|---|
| Q1. Repaired congenital heart disease with remaining defects is one of the reasons for using antibiotics before dental procedures. | Q15. Amoxicillin is the first choice for oral antibiotic prophylaxis for dental procedures. |
| Q2. Coronary artery stents are one of the reasons for using antibiotics before dental procedures. | Q16. The recommended single dose of amoxicillin for adults is 1 gram. |
| Q3. Patients with transcatheter prosthetic valve placement does not require antibiotics prophylaxis before a dental procedure. | Q17. For children the recommended single dose of amoxicillin for antibiotic prophylaxis is 50 mg/kg. |
| Q4. Patients with pacemaker devices do not require antibiotic prophylaxis for a dental procedure. | Q18. Azithromycin or clarithromycin are suggested for patients unable to take oral medication. |
| Q5. Patients with peripheral vascular grafts require antibiotic prophylaxis for dental procedures. | Q19. The recommended single dose of cefazolin for adults unable to take oral medication is 1 gram. |
| Q6. Patients with complete closure of septal defects are recommended to receive antibiotic prophylaxis for dental procedures. | Q20. The use of cephalexin cefazolin or ceftriaxone is not recommended for patients with a history of anaphylaxis angioedema or urticarial with penicillin or ampicillin. |
| Q7. Patients with previous relapse or recurrent infective endocarditis should not be prescribed antibiotics before invasive dental procedure. | Q21. Patients allergic but not anaphylaxis to penicillin should not receive cephalosporins due to the high risk of cross-reactivity. |
| Q8. Antibiotic prophylaxis should be prescribed to at-risk patients before any dental procedure. | Q22. Alternative antibiotics should not be used for each dental procedure if repeated procedures are required in a short period. |
| Q9. Viridans group streptococci are the main group of microorganisms that cause infective endocarditis from oral sources. | Q23. In patients receiving a short course (7–10 days) of oral antibiotic therapy before a dental procedure it is preferable to select a different class of antibiotics. |
| Q10. Viridans group streptococci infective endocarditis is much more likely to be caused by daily oral routine activities than by dental procedures. | Q24. The administration of antibiotics for prophylaxis should always occur before the dental procedure and not afterwards. |
| Q11. Clindamycin is no longer recommended for antibiotic prophylaxis for a dental procedure. | Q25. Single dose of clindamycin can cause serious complications including death from a C. difficile infection. |
| Q12. Maintaining good oral health and having regular dental care are more important for preventing Viridans group streptococci infective endocarditis than taking antibiotics before a dental procedure. | Q26. In patients undergoing multiple sequential dental appointments it is preferable to delay the next procedure for 10 days after the last dose of antibiotic therapy. |
| Q13. Cephalexin is an alternative for patients allergic to penicillin or ampicillin. | Q27. In patients who are receiving parenteral antimicrobial therapy for IE or other infections and require a dental procedure the same parenteral antibiotic may be continued through the dental procedure. |
| Q14. Doxycycline can be used for patients who are unable to tolerate penicillin cephalosporin or macrolide antibiotics. | Q28. Patients with prosthetic joint implants antibiotic prophylaxis are recommended before any dental procedure. |

95% confidence intervals for each model's performance. Intermodel comparisons were performed using the Kruskal–Wallis test, followed by post-hoc pair wise comparisons. The performance comparison between system-prompted and nonsystem-prompted models was determined using the Mann-Whitney U test. Statistical significance was set at $p < .05$.

## Results

Significant differences in accuracy were observed among the LLMs ($p = .0007$ with pre-prompt and $p = .0002$ without pre-prompt). With a pre-prompt, GPT-4o demonstrated the highest accuracy (80%), followed by Gemini 1.5 Pro (78.57%), and Claude 3 Opus and GPT 3.5 Turbo (both at 75.71%). Gemini 1.5 Flash exhibited the lowest accuracy (68.57%). Statistical analysis indicated that Gemini 1.5 Flash performed significantly worse than GPT-4o and Gemini 1.5 Pro ($p < .01$ for both). Similarly, Claude 3.5 Sonnet displayed significantly lower accuracy compared to GPT-4o ($p = .0264$) and Gemini 1.5 Pro ($p = .0312$) (Figure 1A-B, Table 2). The models exhibited varying 95% confidence intervals of mean accuracy, with GPT 3.5 Turbo and Gemini 1.5 Flash displaying the widest intervals (Table 2). Notably, Gemini 1.5 Pro and Claude 3.5 Sonnet consistently provided 22 and 20 correct answers, respectively, across all iterations. Further, Claude 3.5 Sonnet consistently produced the same incorrect response to specific questions. Among the LLMs evaluated, only one response (question 22, GPT-4o, second iteration) was categorised as "unable to answer" and treated as incorrect (Supplementary Table 1).

Without a pre-prompt, GPT-4o exhibited the highest accuracy (75.57%), followed by Gemini 1.5 Pro (77.86%), and Claude 3 Opus (77.14%). Gemini 1.5 Flash exhibited the lowest accuracy (63.57%). These findings are consistent with the results observed when using a pre-prompt. Statistical analysis indicated that Gemini 1.5 Flash performed significantly worse than GPT-4o and Gemini 1.5 Pro ($p < .01$ for both) and Claude 3 Opus ($p = .0158$). Notably, in contrast to the pre-prompt condition, Claude 3.5 Sonnet did not show significantly lower accuracy compared to GPT-4o and Gemini 1.5 Pro. GPT-4o and GPT 3.5 Turbo displayed the widest confidence intervals without a pre-prompt (Figure 2A-B, Table 3).

All models, except for Claude 3 Opus, demonstrated lower accuracy when evaluated without a system prompt compared to the pre-prompt condition, though these differences were not statistically significant (Figure 2C). Interestingly, all LLMs exhibited wider confidence intervals of mean accuracy (higher standard deviation) compared to those with a pre-prompt. (Figure 1C, Table 2, 3). Details of each LLM's responses, both with and without a pre-prompt, are provided in Supplementary Tables 1 and 2.

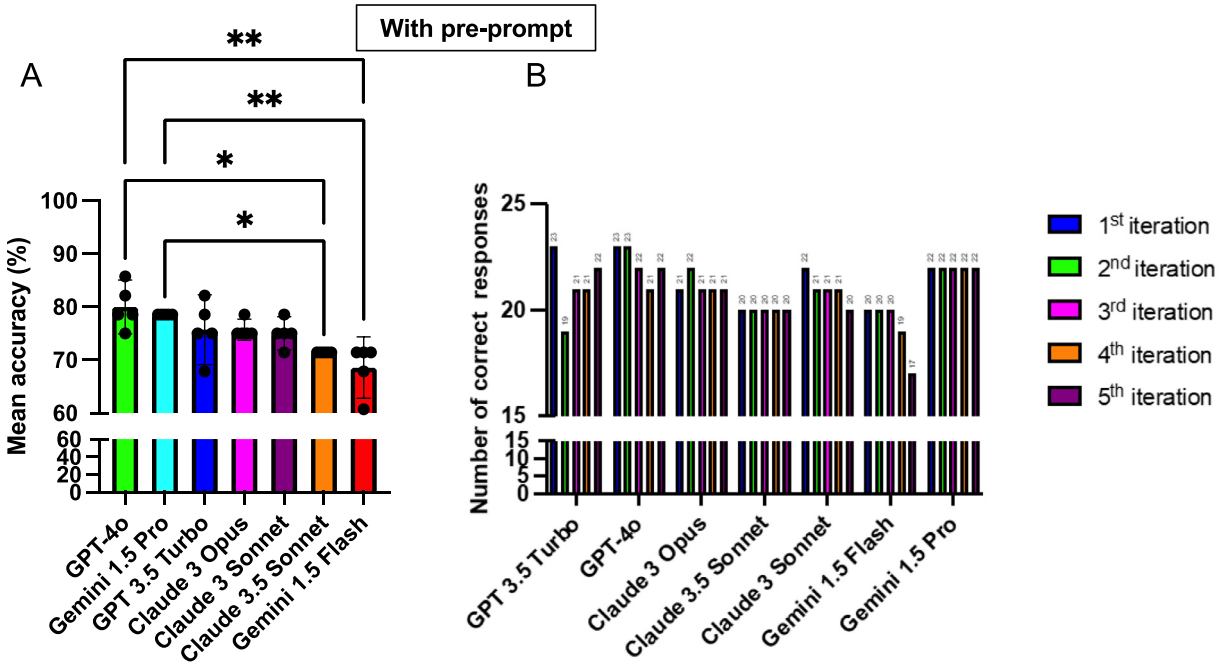


**Fig. 1 – The accuracy of LLMs tested with pre-prompts. (A) Comparison of mean accuracy (%) across LLMs. (B) Consistency of correct answers provided by LLMs over five runs. Asterisks indicate statistically significant between models. *$p < .05$, ** $p < .01$.**

## Discussion

The AHA guideline for infective endocarditis (IE) prevention is the gold standard in dental care for high-risk patients, but its complexity often hinders consistent adherence.[15,16] With varying practices in antibiotic prescription among dental professionals, recent updates to the AHA-IE guidelines, and the rapid advancement of LLMs, these models present a promising opportunity. They can serve as accessible tools to quickly retrieve details, provide answers, and outline protocols for clinicians.

The 2008 UK National Institute for Health and Care Excellence (NICE) guideline, last updated in 2016, did not recommend routine antibiotic prophylaxis for dental procedures.[17] However, recent research has provided a more nuanced understanding of the issue. Notably, a 2023 study in England analyzing IE hospital admissions[18] and a 2022 US study focusing on individuals with commercial/Medicare-supplemental coverage[19] demonstrated a significant reduction in IE incidence among high-risk individuals who received antibiotic prophylaxis before invasive dental procedures, particularly tooth extractions and oral surgery. These findings highlight the evolving evidence regarding antibiotic prophylaxis for IE in high-risk individuals. While recent studies suggest a potential benefit of prophylaxis, national guidelines, such as those from NICE, evaluate various factors, including local epidemiological data, cost-effectiveness, and patient outcomes, when updating their recommendations. This comprehensive approach ensures that guidelines remain relevant and applicable to the healthcare context in the United Kingdom. In contrast, the 2021 AHA guidelines[11] and the 2023 European Society of Cardiology (ESC) guidelines,[20] have updated their positions to recommend antibiotic prophylaxis for high-risk individuals prior to invasive dental procedures. These guidelines incorporate a comprehensive understanding of both patient-specific and procedure-specific risk factors for IE.

**Table 2 – The accuracy performance of LLMs with pre-prompts.**

| Models | Mean accuracy (%) | Confidence interval | | Standard deviation |
|---|---|---|---|---|
| | | Upper limit | Lower limit | |
| GPT 3.5 Turbo | 75.71 | 82.29 | 69.14 | 1.48 |
| GPT 4o | 80 | 85.06 | 74.94 | 0.84 |
| Claude 3 Opus | 75.71 | 77.70 | 73.73 | 0.45 |
| Claude 3.5 Sonnet | 71.43 | 71.43 | 71.43 | 0 |
| Claude 3 Sonnet | 75 | 78.14 | 71.86 | 0.71 |
| Gemini 1.5 Flash | 68.57 | 74.35 | 62.79 | 1.30 |
| Gemini 1.5 Pro | 78.57 | 78.57 | 78.57 | 0 |

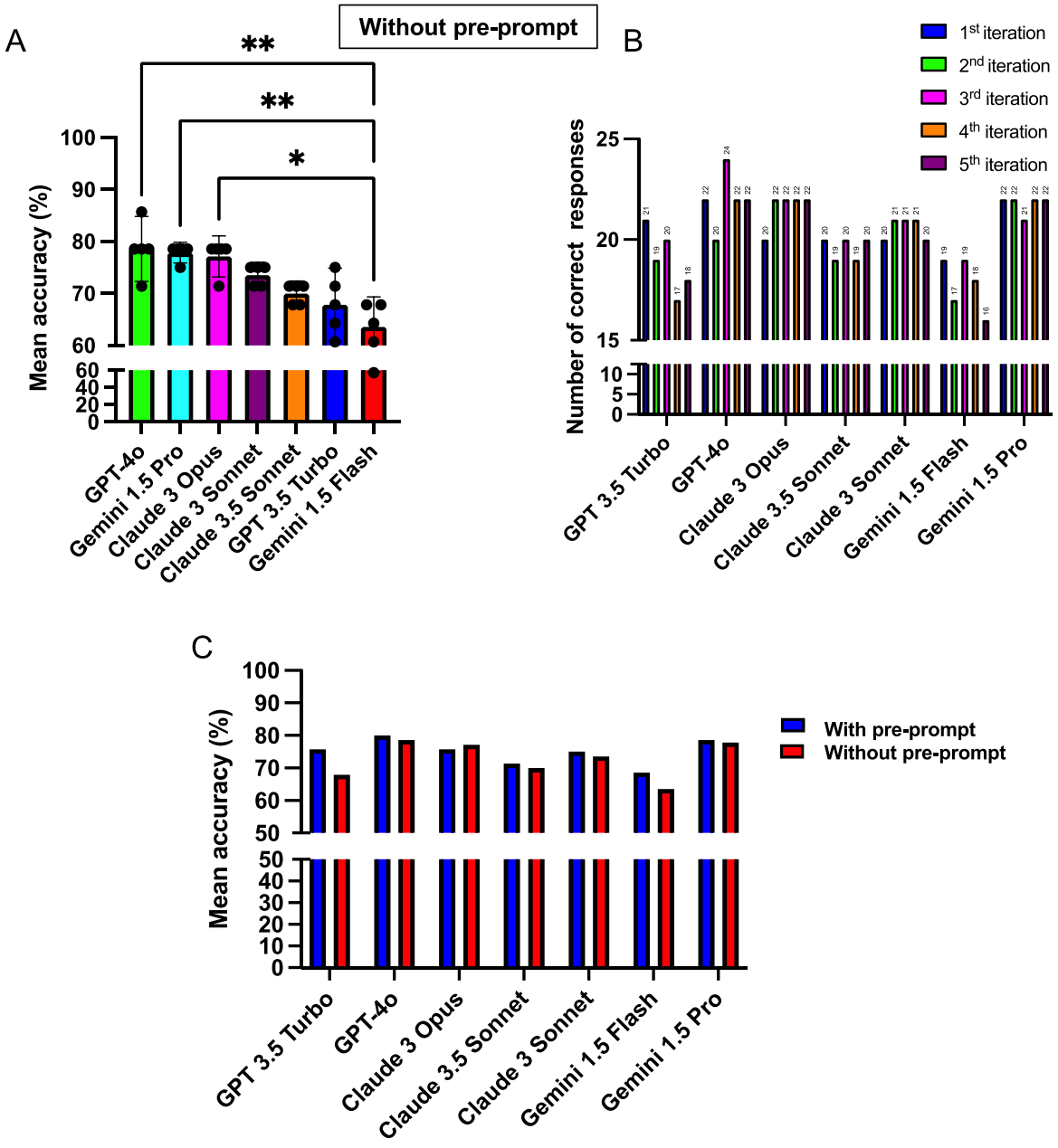


**Fig. 2 – The accuracy of LLMs tested without pre-prompts and comparison with those with pre-prompts. (A) Comparison of mean accuracy (%) across LLMs. (B) Consistency of correct answers provided by LLMs over five runs. (C) Comparison of mean accuracy between with and without pre-prompts within the same model. Asterisks indicate statistically significant between models. $^{*}p < .05$, $^{**}p < .01$.**

**Table 3 – The accuracy performance of LLMs without pre-prompts.**

| Models | Mean accuracy (%) | Confidence interval | | Standard deviation |
|---|---|---|---|---|
| | | Upper limit | Lower limit | |
| GPT 3.5 Turbo | 67.86 | 74.87 | 60.85 | 5.647 |
| GPT 4o | 78.57 | 84.84 | 72.30 | 5.051 |
| Claude 3 Opus | 77.14 | 81.11 | 73.18 | 3.194 |
| Claude 3.5 Sonnet | 70.00 | 72.43 | 67.57 | 1.956 |
| Claude 3 Sonnet | 73.57 | 76.00 | 71.14 | 1.956 |
| Gemini 1.5 Flash | 63.57 | 69.35 | 57.79 | 4.657 |
| Gemini 1.5 Pro | 77.86 | 79.84 | 75.87 | 1.597 |

**Table 4 – Comparison of benchmark scores of LLMs from official online sources.[24-26]**

| Models | MMLU (%) | GPQA (%) | DROP (%) |
|---|---|---|---|
| GPT-4o | 88.7 | 53.6 | 83.4 |
| Claude3.5 Sonnet | 88.3 | 59.4 | 87.1 |
| Claude3 Opus | 85.7 | 50.4 | 83.1 |
| Gemini1.5 Flash | 78.9 | 41.5 | N/A |
| Gemini1.5 Pro | 85.9 | 46.2 | 78.9 |

MMLU: Massive Multitask Language Understanding, GPQA: General-Purpose Question Answering, DROP: Discrete Reasoning Over Paragraphs.

LLMs have been benchmarked in understanding tasks in the medical field and expanded to evaluate tasks requiring specialised knowledge.[21-23] Recent benchmarks like Massive Multitask Language Understanding (MMLU) assess LLMs across various domains, including classification, extraction, and reasoning. General-Purpose Question Answering (GPQA) specifically evaluates question-answering performance, while Discrete Reasoning Over Paragraphs (DROP) focuses on tasks requiring more complex reasoning. However, benchmarks specifically evaluating LLMs in the dental field are very limited. The LLMs selected for this study achieved MMLU benchmark scores of 78.9% to 88.7%, DROP scores of 78.9% to 87.1%, and GPQA scores of 41.5% to 59.4% (accessed July 23, 2024) (Table 4),[24-26] highlighting the disparity in performance across different models. Our study found that these LLMs achieved an accuracy of 68.57% to 80% with pre-prompt and 63.57% to 78.57% without pre-prompt, suggesting that they may not yet be optimally trained to answer medical-dental specific questions. Furthermore, general benchmarks, while valuable for broad assessments, may not accurately predict performance in highly specialised fields like dentistry.

The knowledge cutoff dates of the LLMs can significantly influence their performance,[27-29] resulting in varying levels of up-to-date information across models. While GPT-4o and Claude 3 Opus acknowledged the 2021 AHA guidelines in their responses, inconsistencies in applying this updated information across different questions were observed. This highlights the need for regular model updates and consistent application of knowledge in answering medical-dental questions. It is crucial for users to verify information provided by LLMs before applying it to clinical scenarios.

A notable finding was the presence of internal contradictions within some model responses. For example, while GPT-4o and Claude 3 Opus correctly stated in one question that clindamycin is no longer recommended for IE prophylaxis as per the 2021 AHA guidelines, they inconsistently recommend clindamycin in responses to subsequent questions. Such contradictions pose a significant challenge for practical applications, as they could lead to confusion or misinformation in clinical settings. This inconsistency underscores the need for careful interpretation of LLM outputs and highlights a key area for improvement in future model developments.

Another challenge observed pertained to the models' handling of negative questions. They frequently struggled to discern the presence of "not" in the queries, resulting in incorrect responses that misclassified the nature of True/False scenarios. Even when correctly identifying negative questions, instances of inaccuracies in their responses were noted. Moreover, despite referencing reputable sources, their answers occasionally included hallucinated or inadequately reasoned content. To enhance query accuracy, it is recommended to avoid using negative framing in questions posed to LLMs.

A possible way to improve the performance of LLMs is through Prompt Engineering (PE). PE focuses on creating and refining prompts to better utilise LLMs' extensive textual data sets. By carefully designing prompts, one can guide the model's behavior, thereby enhancing the quality of its responses.[30] This approach has been particularly effective in academic writing, where PE provides the necessary structure and context for the model to generate accurate and relevant outputs.[31] In this study, the absence of system pre-prompts generally resulted in lower accuracy and a higher range of deviation compared to system-prompted performance, indicating the benefit of pre-prompts in setting the models to specialise in the context of the questions' input. Interestingly, Claude 3 Opus demonstrated a higher percentage of mean accuracy with non-system prompts, which could be attributed to the stochastic nature of LLMs. Another interesting finding was that GPT 3.5 Turbo responded with an average of 40% non-reasoning, providing only true or false responses.

This study identified several limitations, including instances where LLM responses were incorrectly classified due to the inclusion of additional accurate information. For example, in question 17, both Claude 3.5 Sonnet and Claude 3 Sonnet correctly provided the appropriate dosage but incorrectly labeled the statement false by adding that the dose for children should not exceed 2 grams. This highlights the need for a nuanced understanding of specialised fields when evaluating LLM responses. Furthermore, users should carefully review responses in their entirety, as answers and reasoning within LLM outputs may occasionally contradict each other, even when citing reputable references. Healthcare professionals must exercise caution and maintain a critical awareness of these models' limitations, as even minor inaccuracies or outdated information can have significant implications for patient care and safety.

## Conclusion

Among the 7 LLMs evaluated, GPT-4o demonstrated the highest accuracy in answering questions related to the 2021 AHA-IE guidelines, particularly when provided with a pre-prompt. This suggests that specialist guidance can significantly improve LLM performance and reduce variability. Continuous benchmarking of LLM-provided information against current, peer-reviewed sources remains crucial to ensure the highest standards of patient care and safety in medical practice. While LLMs offer promising potential, they should be viewed as supplementary tools, requiring careful evaluation and verification by healthcare professionals.

## Conflict of interest

None disclosed.

## Data availability statement

The data supporting the findings of this study are available within the article and its supplementary materials.

## Acknowledgements

We acknowledge the use of ChatGPT for checking English grammar in the preparation of this manuscript.

## Funding

This work was supported by Thailand Science Research and Innovation Fund Chulalongkorn University, Health Systems Research Institute (66-101), National Research Council of Thailand (N42A650229), and Faculty of Dentistry (DRF68_007), Chulalongkorn University.

## Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.identj.2024.09.033.

## REFERENCES

1. OpenAI. ChatGPT. 2022.
2. Anthropic. Claude. 2023.
3. Google. Google gemini. 2024.
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930–40. doi: 10.1038/s41591-023-02448-8.
5. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med 2023;3 (1):141. doi: 10.1038/s43856-023-00370-1.
6. Ozden I, Gokyar M, Ozden ME, Sazak Ovecoglu H. Assessment of artificial intelligence applications in responding to dental trauma. Dent Traumatol 2024. doi: 10.1111/edt.12965.
7. Rokhshad R, Zhang P, Mohammad-Rahimi H, Pitchika V, Entezari N, Schwendicke F. Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: a pilot study. J Dentistry 2024;144:104938. doi: 10.1016/j.jdent.2024.104938.
8. Muñoz P, Kestler M, De Alarcon A, et al. Current epidemiology and outcome of infective endocarditis: a multicenter, prospective, cohort study. Medicine 2015;94(43):e1816. doi: 10.1097/md.0000000000001816.
9. Lockhart PB, Bolger A, Baddour LM. The 2021 American Heart Association Statement on prevention of infective endocarditis: what's new? J am Dent Assoc 2021;152(11):880–2. doi: 10.1016/j.adaj.2021.08.001.
10. American Academy of Pediatric Dentistry. Antibiotic prophylaxis for dental patients at risk for infection. The Reference Manual of Pediatric Dentistry. Chicago, IL; 2022. p. 500–6.
11. Wilson WR, Gewitz M, Lockhart PB, et al. Prevention of viridans group streptococcal infective endocarditis: a scientific statement from the American Heart Association. Circulation. 2021;143(20):e963–e78. doi: 10.1161/CIR.0000000000000096.
12. Sperotto F, France K, Gobbo M, et al. Antibiotic prophylaxis and infective endocarditis incidence following invasive dental procedures: a systematic review and meta-analysis. JAMA Cardiol 2024;9(7):599–610. doi: 10.1001/jamacardio.2024.0873.
13. Lean SSH, Jou E, Ho JSY, Jou EGL. Prophylactic antibiotic use for infective endocarditis: a systematic review and meta-analysis. BMJ Open 2023;13(8):e077026. doi: 10.1136/bmjopen-2023-077026.
14. Mylonakis E, Calderwood SB. Infective endocarditis in adults. New E J Med 2001;345(18):1318–30. doi: 10.1056/NEJMra010082.
15. Jain P, Stevenson T, Sheppard A, et al. Antibiotic prophylaxis for infective endocarditis: knowledge and implementation of American Heart Association Guidelines among dentists and dental hygienists in Alberta, Canada. J Am Dent Assoc. 2015;146(10):743–50. doi: 10.1016/j.adaj.2015.03.021.
16. Lockhart PB, Hanson NB, Ristic H, Menezes AR, Baddour L. Acceptance among and impact on dental practitioners and patients of American Heart Association recommendations for antibiotic prophylaxis. J Am Dent Assoc 2013;144(9):1030–5. doi: 10.14219/jada.archive.2013.0230.
17. Richey R, Wray D, Stokes T. Prophylaxis against infective endocarditis: summary of NICE guidance. Last updated: July 2016. BMJ 2008;336(7647):770–1. doi: 10.1136/bmj.39510.423148.AD.
18. Thornhill MH, Crum A, Campbell R, et al. Temporal association between invasive procedures and infective endocarditis. Heart 2023;109(3):223–31. doi: 10.1136/heartjnl-2022-321519.
19. Thornhill MH, Gibson TB, Yoon F, et al. Antibiotic prophylaxis against infective endocarditis before invasive dental procedures. J Am Coll Cardiol 2022;80(11):1029–41. doi: 10.1016/j.jacc.2022.06.030.
20. Delgado V, Ajmone Marsan N, de Waha S, et al. 2023 ESC Guidelines for the management of endocarditis: Developed by the task force on the management of endocarditis of the European Society of Cardiology (ESC) Endorsed by the European Association for Cardio-Thoracic Surgery (EACTS) and the European Association of Nuclear Medicine (EANM). Eur Heart J 2023;44(39):3948–4042. doi: 10.1093/eurheartj/ehad193.
21. Jahan I, Laskar MTR, Peng C, Huang JX. A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks. Comput Biol Med 2024;171:108189. doi: 10.1016/j.compbiomed.2024.108189.
22. Xu J, Lu L, Peng X, et al. Data set and benchmark (MedGPTEval) to evaluate responses from large language models in medicine: evaluation development and validation. JMIR Med Inform 2024;12:e57674. doi: 10.2196/57674.
23. Chau RCW, Thu KM, Yu OY, Hsung RT-C, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. Int Dent J 2024;74(3):616–21. doi: 10.1016/j.identj.2023.12.007.
24. G. DeepMind Gemini: AI model technical report. 2023. Available from: https://www.deepmind.com/gemini. Accessed 23 July 2024.
25. OpenAI. Hello GPT-4o 2024. Available from: https://openai.com/index/hello-gpt-4o/. Accessed 23 July 2024.
26. Anthropic. Claude 3.5: A Sonnet for the next generation of AI 2024. Available from: https://www.anthropic.com/news/claude-3-5-sonnet. Accessed 30 June 2024.
27. OpenAI. ChatGPT: GPT-4 technical report. 2023. Available from: https://www.openai.com/research/chatgpt. Accessed 30 June 2024.
28. Anthropic. Learn about Claude 2024. Available from: https://docs.anthropic.com/en/docs/about-claude/models. Accessed 30 June 2024.