Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Journal of Dental Sciences

Original Article

Clinical decision support of advanced large language models in endodontic disease

Jiahe Li [a,b*†], Xian He [a,c†], Yong Wang [d], Yanan Liu [e], Jie Liu [a,b], Mingxiao Liu [a,c], Tianyu Huang [a,c], Zeyu Huang [f]

[a] State Key Laboratory of Oral Diseases, National Clinical Research Center for Oral Diseases, West China College of Stomatology, Sichuan University, Chengdu, China
[b] Department of Cariology and Endodontics, West China Hospital of Stomatology, Sichuan University, Chengdu, China
[c] Department of Orthodontics, West China Hospital of Stomatology, Sichuan University, Chengdu, China
[d] Department of Pediatric Dentistry, Beijing Stomatological Hospital & School of Stomatology, Capital Medical University, Beijing, China
[e] Beijing Tiantan Hospital, Capital Medical University, Beijing, China
[f] School & Hospital of Stomatology, Wuhan University, Wuhan, China

Received 12 April 2025; Final revision received 7 June 2025
Available online 20 June 2025

KEYWORDS
Artificial intelligence;
Large language models;
Endodontics;
Clinical decision-making;
Clinical decision support

Abstract *Background/purpose:* Large language models (LLMs) exhibit significant potential for clinical decision support, yet their application in endodontic disease remains underexplored.
Materials and methods: This study assessed the decision-making capabilities of three advanced LLMs (GPT-4o, Claude 3.5, and Grok2) in specialized endodontic contexts. A question bank of 421 multiple-choice questions was constructed across 27 core endodontic topics, including theory, procedures, and 35 complex cases. The three LLMs were tested using standardized prompts, with performance evaluated via topic-stratified accuracy analysis.
Results: Claude 3.5 achieved the highest overall accuracy (73.39 %), followed by Grok2 (66.27 %) and GPT-4o (46.32 %). Grok2 excelled in complex case analysis (69.57 %). The models performed strongly in theoretical domains (e.g., clinical examination, structural function, pharmacology) but showed limitations in complex scenarios and procedural techniques.
Conclusion: LLMs hold promise as endodontic decision support tools, though domain-specific refinement is essential for effective clinical application.

* Corresponding author. State Key Laboratory of Oral Diseases, West China College of Stomatology, Sichuan University, No. 14, 3rd Section, Renmin South Road, Chengdu 610065, China.
E-mail addresses: jiaheli@stu.scu.edu.cn, dentist.lijiahe@gmail.com (J. Li).

† These authors contributed equally to this work.

https://doi.org/10.1016/j.jds.2025.06.007

Introduction

With the rapid development of artificial intelligence technology, large language models (LLMs) are gradually expanding their applications in healthcare, demonstrating immense potential to support clinical decision-making.[1—3] Although LLMs have demonstrated impressive capabilities in medical knowledge assessment, research on their application in dentistry, particularly in specialized fields such as endodontics, remains relatively scarce.

Endodontic diseases, including pulpitis, periapical periodontitis, and dental trauma, have high global incidence rates, causing significant pain and discomfort for patients.[4] Accurate diagnosis and timely treatment are crucial for tooth preservation, symptom relief, and prevention of complications. However, the diverse clinical presentations of endodontic diseases, complex diagnostic criteria, and treatment selection considerations that encompass multiple factors present challenges for clinicians.[5]

Recent research indicates that LLMs such as GPT-4, Claude, and Grok have demonstrated remarkable capabilities in the medical domain, passing the United States Medical Licensing Examination and performing excellently in various medical specialty knowledge assessments.[6,7,8] However, the performance of these models in highly specialized fields such as endodontics has not been systematically evaluated. Whether LLMs can comprehend the complex pathophysiology of endodontic diseases, master diagnostic criteria, and provide treatment recommendations consistent with clinical guidelines remains an urgent question for exploration.

This study aimed to systematically evaluate the capabilities of three advanced large language models—GPT-4o, Claude 3.5 Grok2—in supporting professional knowledge and clinical decision-making in endodontics. Through a carefully designed set of professional questions covering multiple dimensions including endodontic disease diagnosis, treatment planning, medication use, and clinical operational techniques, we expect to comprehensively examine the strengths and limitations of these models. The results will provide empirical reference for endodontic clinicians regarding the potential value of AI-assisted tools and guidance for future optimization directions of LLMs in specialized dental applications.

Materials and methods

Research design

This study employed a cross-sectional design, evaluating LLM performance through a constructed set of endodontic professional questions. The research process included three main phases: question bank construction, model testing, and results analysis.

Question bank construction

The research team consisted of three endodontic specialists and two research assistants. The expert team identified 27 core topics (Table 1) based on the latest version of "Clinical Guidelines for Endodontics" and high-quality endodontic research literature published in the past five years. These topics covered theoretical foundations, diagnostic methods, treatment techniques, pharmaceutical applications, and ethical regulations in endodontics. For each topic, the expert team designed 5—20 multiple-choice questions (MCQs), each with 5 potential answers. The questions ensured reasonable distribution of difficulty levels and coverage of key decision points in clinical practice. The final comprehensive question bank contained 421 questions. To ensure objectivity in assessment, all questions were based on existing clinical guidelines and evidence-based medicine, with each question having only one recognized best answer. Additionally, the expert team designed 35 comprehensive case analysis questions requiring the models to recommend diagnoses and treatment plans based on clinical information.

Model selection and testing

This study selected three representative large language models: GPT-4o (OpenAI), Claude 3.5 Sonnet (Anthropic), and Grok2 (xAI). Testing followed a standardized process, using identical prompts for each model: "Please answer the following endodontic multiple-choice question, providing only the letter of the option you consider correct without explanation." For comprehensive case analyses, the prompt was adjusted to: "Please analyze the following endodontic clinical case and provide only the letter of the option you consider correct without explanation." Testing was completed without human intervention, with model responses automatically recorded for analysis.

Human expert control groups

To provide clinical context for the LLM performance evaluation, we recruited three control groups of human participants: senior dental students (n = 10), endodontic residents (n = 10), and experienced endodontic specialists with >5 years of clinical experience (n = 10). All participants completed the same 421-question assessment under standardized conditions. Senior dental students were in their final year of dental school with completed endodontic coursework. Endodontic residents were in their second or

Table 1 Question design.

Chapter	Subject	Counts
Chapter 1	Diagnose	20
Chapter 2	Emergency	24
Chapter 3	Clinical examination	11
Chapter 4	Treatment planning	12
Chapter 5	Preparation for treatment	21
Chapter 6	Armamentarium and sterilization	13
Chapter 7	Cavity preparation	10
Chapter 8	Cleaning and shaping	12
Chapter 9	Obturation	17
Chapter 10	Records and legal responsibilities	12
Chapter 11	Structure and functions	16
Chapter 12	Pathobiology	15
Chapter 13	Microbiology	20
Chapter 14	Instruments and materials	20
Chapter 15	Pulp reaction	19
Chapter 16	Traumatic injuries	17
Chapter 17	Endodontic and periodontic	10
Chapter 18	Pharmacology	19
Chapter 19	Endodontic microsurgery	15
Chapter 20	Management of pain and anxiety	19
Chapter 21	Tooth whitening	6
Chapter 22	Restoration of the endodontically treated tooth	10
Chapter 23	Pediatric endodontics	14
Chapter 24	Geriatric endodontics	17
Chapter 25	Nonsurgical endodontic retreatment	12
Chapter 26	Digital technologies in endodontic practice	5
Chapter 27	Case	35
Total		421

third year of specialty training. Specialist endodontists were board-certified with minimum 5 years of independent practice experience. Participants completed the assessment online using the same question format as the LLMs, with a maximum time limit of 12 h to simulate realistic clinical decision-making conditions.

Data analysis

The primary outcome measures were overall accuracy and topic-specific accuracy for each model. Accuracy was defined as the number of correctly answered questions divided by the total number of questions in that category. The research team further analyzed performance differences between theoretical knowledge and clinical application dimensions.

Results

Overall performance

In the systematic evaluation of 421 endodontic professional questions, the three large language models demonstrated significant performance differences. As shown in Table 2, Claude 3.5 led with an overall accuracy of 73.39 %, followed by Grok2 (66.27 %) and GPT-4o (46.32 %), revealing capability disparities among current mainstream LLMs in processing specialized medical knowledge. Comprehensive case analysis results presented a different pattern from the multiple-choice assessment. Among 35 complex clinical cases, Grok2 performed best (69.57 %), surpassing Claude 3.5 (62.86 %) and GPT-4o (57.14 %). Comprehensive case analysis required models to integrate multidimensional clinical information, weigh different treatment factors, and formulate individualized plans, representing a capability test closer to actual clinical decision-making. Grok2's advantage in these tasks suggests that different models may employ different internal knowledge representation and reasoning mechanisms, with certain architectures potentially better suited for processing clinical scenarios requiring complex information integration.

Table 2 Accuracy of three large language models and human experts.

	Correct	Fault	Accuracy
ChatGPT-4o	195	226	46.32 %
Grok2	279	142	66.27 %
Claude3.5	309	112	73.39 %
Senior dental students	221	200	52.56 %
Endodontic residents	318	103	75.62 %
Specialist endodontists	377	44	89.48 %

Comparison with human performance

Human expert groups demonstrated the following overall accuracies: senior dental students (52.56 %), endodontic residents (75.62 %), and specialist endodontists (89.48 %). Claude 3.5's performance was comparable to endodontic residents, falling just 2.23 percentage points below their performance level, while substantially exceeding senior dental student performance by 20.83 percentage points. Grok2 performed at an intermediate level between senior dental students and endodontic residents. GPT-4o performed below the level of senior dental students, with a 6.24 percentage point gap.

Topic-specific performance analysis

The models' performance across 27 professional topics revealed clear "areas of strength" and "weak points" (Figs. 1–3). All three models excelled in four core knowledge domains: (a) pulp anatomy and physiology (average accuracy 86.79 %); (b) basic theory of pulpitis (average accuracy 82.54 %); (c) clinical examination and diagnostic methods (average accuracy 79.31 %); and (d) pharmaceutical applications (average accuracy 77.62 %). Conversely, all models performed notably poorly in four highly specialized application areas: (a) pediatric endodontic treatment (average accuracy 41.28 %), where GPT-4o achieved 0 % accuracy; (b) pulp regeneration techniques (average accuracy 47.56 %); (c) instrument sterilization and infection control (average accuracy 49.74 %); and (d) legal regulations and ethical issues (average accuracy 53.17 %).

Detailed analysis revealed unique "preference patterns" for each model. Claude 3.5 led in 21 of 27 topics, demonstrating broad-spectrum knowledge advantages, particularly excelling in theoretical foundations: clinical examination (100 %), structure and function (100 %), pharmacology (94.74 %), treatment planning (91.67 %), and geriatric endodontics (88.24 %). Grok2 excelled in practice-related topics, particularly in traumatic injuries (88.24 %), pulp reactions (84.21 %), tooth whitening (83.33 %), and digital technology applications (80 %). Although GPT-4o performed weakest overall, it demonstrated relative strengths in specific domains including instruments and materials (91.67 %), diagnostics (84.62 %), and restorative treatment (80 %).

In dental trauma knowledge assessment, Grok2 performed best in this domain, achieving 88.24 % accuracy. In-depth analysis showed that Grok2 performed well in dental trauma classification and initial assessment, while also demonstrating notable capability in managing complex trauma and developing long-term follow-up plans. In emergency dental trauma scenarios, the excellent performance of Grok2 and Claude 3.5 could provide immediate decision support for clinicians. Traditional doctors facing uncertain cases might need to consult guidelines or colleagues, whereas LLMs can provide professional preliminary diagnostic advice within seconds, reducing emergency processing time. This is particularly valuable when primary

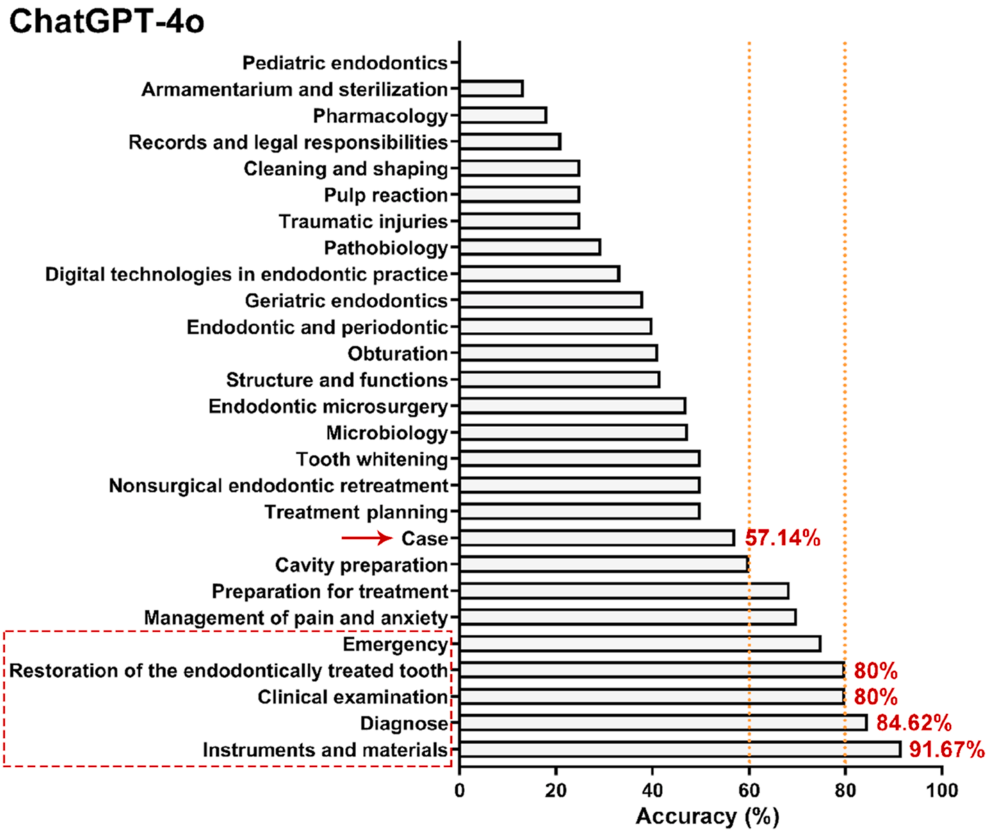

Figure 1 Accuracy rates of GPT-4o across various endodontic topics.

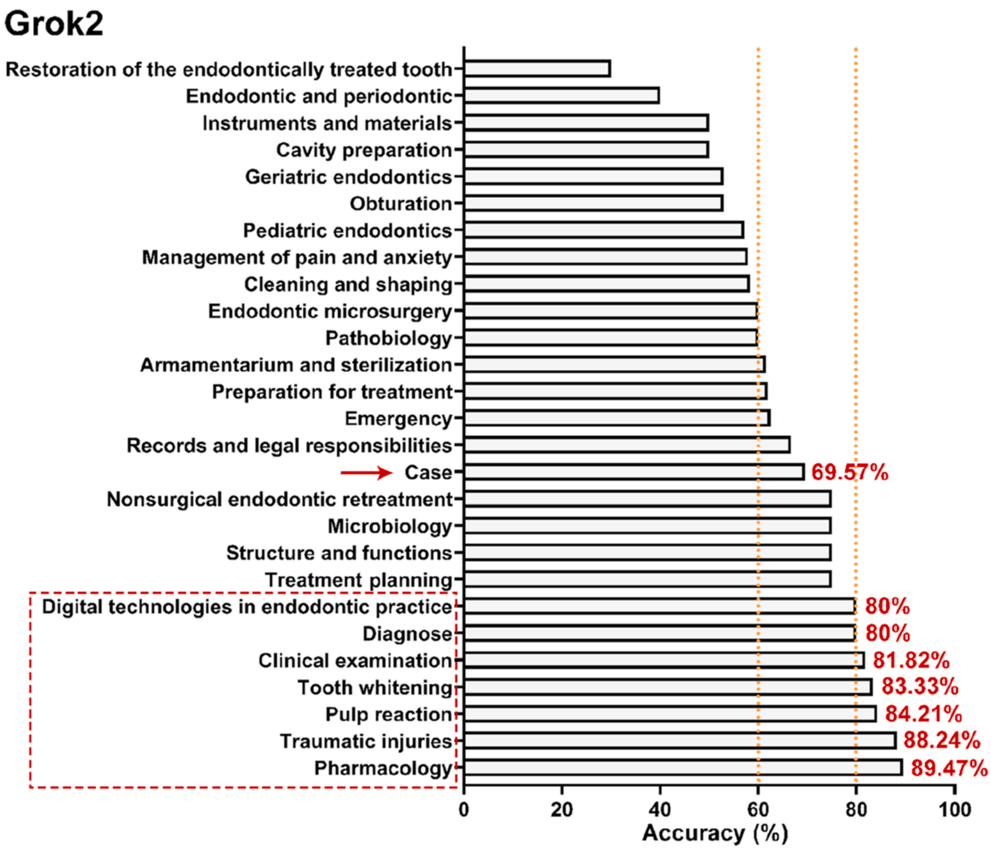

Figure 2 Accuracy rates of Grok2 across various endodontic topics.

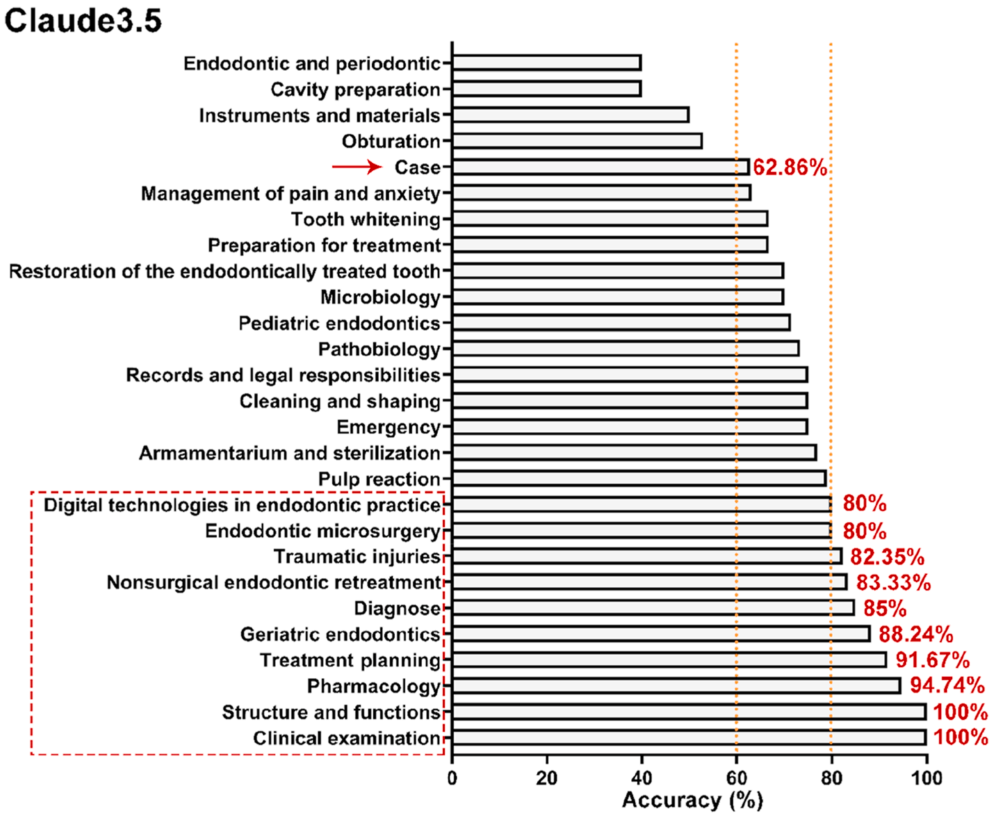

Figure 3 Accuracy rates of Claude 3.5 across various endodontic topics.

healthcare institutions and non-specialists handle dental trauma.[9–11]

Comparison between theoretical knowledge and clinical application

Further analysis showed that all models achieved significantly higher average accuracy on theoretical knowledge questions (73.61 %) compared to clinical application questions (56.29 %). This "theory-practice gap" was most pronounced in GPT-4o, while Claude 3.5 demonstrated the most balanced performance. In-depth analysis revealed that within clinical application questions, complex scenarios involving multi-step decision processes represented the weakest area for models, with an average accuracy of only 45.86 %. Particularly when questions involved patient characteristics (such as age, systemic conditions), anatomical variations, and clinical limiting factors, all three models showed significantly decreased accuracy, indicating substantial limitations in current LLMs' ability to integrate multidimensional clinical information to formulate individualized treatment plans.

Discussion

This study systematically evaluated the performance of three advanced large language models in the specialized field of endodontics. Results indicate these models have developed certain capabilities in clinical treatment decision-making and comprehensive case analysis, but still demonstrate significant limitations in complex clinical decision support. Particularly in highly specialized areas such as pediatric endodontic treatment, regenerative techniques, and infection control, accuracy rates below 50 % indicate that existing models have not yet achieved clinical decision support capabilities in highly specialized branches of endodontics.

The performance patterns observed in this endodontic evaluation are consistent with findings from LLM assessments across other medical specialties. Recent comparative studies have demonstrated similar model hierarchies and performance disparities across different domains. For instance, evaluation on the Japanese Medical Licensing Examination revealed that GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro showed varying strengths across different medical specialties, with distinct performance patterns in clinical versus theoretical knowledge areas.[12] Similarly, comprehensive comparisons on National Board of Medical Examiners sample questions showed that GPT-4 consistently outperformed other LLMs across multiple medical specialties, achieving 100 % accuracy while GPT-3.5 scored 82.2 %, Claude scored 84.7 %, and Bard scored 75.5 %.[13] These cross-specialty findings support our observation that different LLM architectures exhibit distinct cognitive strengths, with some models excelling in structured knowledge domains while others demonstrate superior performance in complex clinical reasoning tasks. The consistency of these patterns across diverse medical fields suggests that the performance characteristics we observed in endodontics reflect fundamental architectural differences among LLMs rather than domain-specific limitations.

Significant performance differences exist among the three models, with Claude 3.5 leading in most theoretical knowledge topics while Grok2 demonstrated superior performance in clinical treatment decisions. Notably, all models exhibited a "capability gap" between theoretical knowledge and clinical application, indicating they still face challenges in translating abstract knowledge into specific clinical decisions. In conclusion, large language models demonstrate potential as decision support tools in endodontics, but require further domain-specific optimization and rigorous clinical validation to truly meet the precision requirements of endodontic treatment.

First, all evaluated models performed significantly better on theoretical knowledge questions than on clinical application questions. This "theory-practice gap" phenomenon has also been observed in previous research in other medical specialties.[14] This disparity likely reflects limitations in existing LLM training data—medical textbooks and theoretical knowledge are relatively abundant in training corpora, while contextualized data on actual clinical decision-making processes are comparatively scarce. Endodontics, as a technique-intensive specialty, often requires integration of multiple factors in clinical decision-making. Our research found that when questions simultaneously incorporated multiple clinical variables, all models' performance decreased significantly, indicating limitations in LLMs' handling of multidimensional clinical decisions.

Second, the significant performance differences among the three evaluated models warrant in-depth exploration. Claude 3.5 led in most theoretical knowledge and standardized diagnostic processes, potentially benefiting from broader medical literature training data; Grok2 excelled in clinical treatment decisions and comprehensive case analysis, suggesting it may possess optimized reasoning capabilities; while GPT-4o's relative advantages in communication and special case management may reflect its strengths in contextual understanding and non-standard situation handling. Existing LLMs may employ different internal knowledge representation and reasoning mechanisms. Particularly noteworthy is Grok2's superior performance in comprehensive case analysis despite Claude 3.5's lead in multiple-choice questions, suggesting Grok2 may possess stronger information integration capabilities and clinical thinking simulation abilities.

Additionally, this study found significant performance disparities among LLMs across different branches of endodontics. All models performed poorly in pediatric endodontics, pulp regeneration techniques, and infection control—precisely the highly specialized and most challenging areas in clinical practice. Pediatric endodontic treatment involves special anatomical structures, behavioral management, and growth and development considerations; regenerative techniques represent frontier development areas with limited and rapidly changing literature. In contrast, LLMs performed excellently in structured knowledge domains such as root canal anatomy, basic pathology, and standardized diagnostic processes. This unbalanced performance pattern reflects training data bias, resulting in relatively deficient capabilities in areas requiring experience accumulation and contextual judgment.

Further analysis of the specific reasons behind poor model performance in certain topics reveals several underlying factors. In pediatric endodontics, where all models struggled (average accuracy 41.28 %, with GPT-4o achieving 0 %), the poor performance likely stems from the unique complexity of this subspecialty that combines anatomical variations, behavioral management considerations, and age-specific treatment protocols that are inadequately represented in training data. The 0 % accuracy of GPT-4o suggests potential gaps in training data coverage for this highly specialized area. Similarly, the consistently poor performance in pulp regeneration techniques (average accuracy 47.56 %) can be attributed to this being a rapidly evolving field with limited established protocols and frequent updates to clinical guidelines. LLMs trained on historical data may not capture the latest developments in regenerative endodontics, leading to outdated or incomplete knowledge representation. The weakness in instrument sterilization and infection control (average accuracy 49.74 %) is particularly concerning given the critical importance of these protocols. This poor performance likely reflects the procedural and protocol-heavy nature of this domain, where models struggle to apply step-by-step procedural knowledge that requires understanding of sequential dependencies and safety-critical decision points. In contrast, the models' excellent performance in theoretical domains such as pulp anatomy and physiology (average accuracy 86.79 %) and basic pathology suggests that well-established, textbook-based knowledge is effectively captured in training data. The superior performance in these areas indicates that LLMs excel when processing structured, foundational knowledge that has remained relatively stable over time and is abundantly documented in medical literature.

This study has several limitations: First, while we included human expert comparison groups, the sample size was limited to 10 participants per group; second, the assessment was primarily based on multiple-choice questions, which may not comprehensively reflect model performance in open-ended clinical reasoning; third, it failed to evaluate model performance under multimodal input conditions (such as imaging data). Future research should consider larger human control groups, multimodal inputs, and direct clinical outcome studies to further explore the impact of LLM-assisted decision-making on actual diagnostic and therapeutic outcomes.

The comparison with human experts provides important clinical context for interpreting LLM performance. Claude 3.5's accuracy level (73.39 %) approaching that of endodontic residents (75.62 %) suggests significant potential utility as a clinical decision support tool, particularly for general dentists managing endodontic cases. The model's performance substantially exceeded that of senior dental students (52.56 %), indicating that advanced LLMs have already surpassed entry-level dental knowledge in endodontics. However, the 16.09 percentage point gap between Claude 3.5 and specialist endodontists (89.48 %) highlights the continued superiority of extensive clinical experience and specialized training. The similar theory-practice gaps observed in both LLMs and human participants highlight the inherent challenges in translating theoretical knowledge to clinical decision-making, regardless of whether the decision-maker is artificial or human. This parallel suggests that current LLM limitations may reflect fundamental challenges in medical decision-making rather than purely technological constraints.

Declaration of competing interest

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

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw Open* 2023;6:e2343689.
2. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open* 2024;7:e2440969.
3. Sandmann S, Riepenhausen S, Plagwitz L, et al. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nat Commun* 2024;15:2050.
4. Aminoshariae A, Kulild JC, Mickel A, et al. Association between systemic diseases and endodontic outcome: a systematic review. *J Endod* 2017;43:514—9.
5. McCabe PS, Dummer PMH. Pulp canal obliteration: an endodontic diagnosis and treatment challenge. *Int Endod J* 2012; 45:177—97.
6. 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.
7. Khan AA, Khan AR, Munshi S, et al. Assessing the performance of ChatGPT in medical ethical decision-making: a comparative study with USMLE-based scenarios. *J Med Ethics* 2025 (in press).
8. Li J, He X, Liu M. Digital Inequities in Dental Education: Challenges in the Age of Generative Artificial Intelligence. *J Dent Educ* 2025:e13925.
9. Ferreira MC, Batista AM, Marques LS, et al. Retrospective evaluation of tooth injuries and associated factors at a hospital emergency ward. *BMC Oral Health* 2015;15:1—6.
10. Loureiro RM, Naves EA, Zanello RF, et al. Dental emergencies: a practical guide. *Radiographics* 2019;39:1782—95.
11. Andersson L. Epidemiology of traumatic dental injuries. *J Endod* 2013;39.
12. Ozaki Y, Hara Y, Yano T, et al. Performance of advanced large language models on Japanese medical licensing examination: a comparative study. *medRxiv* 2024 (in press).
13. Abbas M, Razak AA, Zaidi STR, et al. Comparing the performance of popular large language models on the National Board of Medical Examiners sample questions. *Cureus* 2024;16: e57851.
14. Greenway K, Butt G, Walthall H. What is a theory-practice gap? An exploration of the concept. *Nurse Educ Pract* 2019;34:1—6.