

Original article

Comparative Evaluation of ChatGPT and DeepSeek in Oral Pathology Education and Evidence-Based Diagnostic Decision Support: A Pilot Study

Karima Alfityory^{1*}, Najwa Own², Basma Elhabbasi³, Mouna Abdullah³, Weam Hraisha³, Omalkhair Hassan⁴

¹Department of Oral Medicine, Oral Pathology and Oral and Maxillofacial Surgery, Faculty of Dentistry, Sirt University, Sirt, Libya

²Department of Oral Medicine, Oral Pathology and Oral and Maxillofacial Surgery, Faculty of Dentistry, Al-Tahadi University, Tripoli, Libya

³Department of Oral Medicine, Oral Pathology and Oral and Maxillofacial Surgery, Faculty of Dentistry, University of Tripoli, Tripoli, Libya

⁴Department of Oral Medicine and Diagnosis and Radiology, Faculty of Dentistry, Sirt University, Sirt, Libya

Email: karima.alfityory@su.edu.ly

Abstract

Large language models (LLMs) are increasingly being integrated into healthcare education and clinical decision-support systems. While ChatGPT has demonstrated promising performance in medical and dental domains, comparative evidence regarding newer models such as DeepSeek remains limited, particularly in oral pathology. This pilot study aimed to compare the performance of ChatGPT and DeepSeek in oral pathology education and evidence-based diagnostic decision support. We conducted a cross-sectional comparative study from January to December 2025. We developed ten oral pathology educational questions and five standardised clinical case scenarios based on established dental curricula and authoritative oral pathology references. Identical prompts were submitted independently to ChatGPT and DeepSeek. Responses were anonymised and evaluated by three board-certified oral pathologists using a structured rubric assessing accuracy, completeness, clinical relevance, diagnostic reasoning, and evidence-based consistency. Each domain was scored on a 5-point Likert scale, with a maximum total score of 25. Statistical analyses included independent-samples t-tests, Cohen's d effect sizes, and intraclass correlation coefficient (ICC) assessment. ChatGPT achieved a significantly higher mean total composite score than DeepSeek (23.2 ± 1.8 vs. 20.6 ± 2.4 ; $p = 0.006$; Cohen's $d = 1.15$). ChatGPT also outperformed DeepSeek across all evaluation domains, including accuracy (4.72 ± 0.31 vs. 4.18 ± 0.52 ; $p = 0.003$), completeness (4.68 ± 0.37 vs. 4.10 ± 0.55 ; $p = 0.005$), clinical relevance (4.61 ± 0.42 vs. 4.09 ± 0.58 ; $p = 0.011$), diagnostic reasoning (4.58 ± 0.46 vs. 3.98 ± 0.63 ; $p = 0.008$), and evidence-based consistency (4.75 ± 0.29 vs. 4.16 ± 0.49 ; $p = 0.002$). Large effect sizes were observed across all domains. Inter-rater reliability demonstrated excellent agreement among evaluators (ICC = 0.87; 95% CI: 0.81–0.92). ChatGPT demonstrated superior performance compared with DeepSeek in oral pathology education and evidence-based diagnostic decision support. Its higher accuracy, completeness, clinical relevance, and consistency with established evidence suggest considerable potential as an adjunctive educational and clinical support tool. Nevertheless, human oversight remains essential to ensure accuracy and mitigate the risks associated with AI-generated outputs. Further large-scale studies incorporating real clinical cases and multimodal diagnostic data are warranted.

Keywords. Artificial Intelligence; Large Language Models; Chatgpt; Deepseek; Oral Pathology; Dental Education.

Introduction

Artificial intelligence (AI) has emerged as one of the most transformative technologies in modern healthcare, fundamentally reshaping the ways in which clinical information is accessed, interpreted, and applied in both educational and clinical contexts. Recent advances in machine learning, natural language processing (NLP), and generative AI have enabled the development of systems capable of performing complex cognitive tasks that were traditionally restricted to human experts. Among these innovations, large language models (LLMs) have gained significant attention due to their ability to understand, generate, summarise, and interpret human language with high fluency and contextual coherence [1,2]. LLMs are deep learning systems trained on vast corpora of text data using transformer-based architectures, enabling them to model linguistic relationships and generate contextually relevant responses through probabilistic prediction of token sequences [3,4]. Unlike traditional rule-based expert systems, LLMs do not rely on predefined decision trees but instead learn complex statistical representations of language from large-scale datasets. The release of ChatGPT marked a pivotal milestone in the democratisation of generative AI, enabling conversational interaction with advanced language models capable of answering questions, generating explanations, and supporting reasoning tasks across multiple domains, including healthcare, education, and scientific research [4,5].

Since their introduction, LLMs have demonstrated promising performance in standardised medical knowledge assessments. Studies have shown that GPT-based models can perform at or near passing thresholds on medical licensing examinations, highlighting their potential utility in medical education and decision support systems [6,7]. Similarly, research evaluating LLMs on clinical reasoning tasks has demonstrated strong performance in knowledge retrieval, diagnostic assistance, and structured question answering [2]. However, concerns remain regarding hallucination, lack of transparency, and potential dissemination of incorrect or misleading information, particularly in high-stakes clinical environments [8,9]. The increasing integration of LLMs into healthcare education has been particularly evident among medical and dental students, who increasingly utilise these systems to summarise scientific literature, generate study materials, and clarify complex concepts. Educational research suggests that LLMs may enhance self-directed learning, improve access to educational resources, and support personalised learning experiences [10,11]. Nevertheless, concerns

persist regarding overreliance on AI-generated content, potential erosion of critical thinking skills, and the need for verification against authoritative sources [12].

Within dentistry, the adoption of AI technologies has expanded significantly over the past decade. Early applications focused primarily on image-based diagnostics, including radiographic interpretation, caries detection, periodontal assessment, and orthodontic planning. More recently, LLMs have extended the scope of AI applications to include natural language-based tasks such as clinical documentation, patient communication, treatment planning support, and educational assistance [13,14]. Systematic reviews indicate that LLMs have substantial potential to enhance efficiency in dental education and clinical workflows while emphasising the importance of human oversight to mitigate risks associated with inaccurate outputs [15,16]. Oral pathology represents one of the most complex and intellectually demanding disciplines within dental education and clinical practice. It involves the study, diagnosis, and management of diseases affecting the oral and maxillofacial region, including inflammatory disorders, infectious diseases, developmental anomalies, cysts, tumours, potentially malignant disorders, and oral cancers. Accurate diagnosis in oral pathology requires integration of clinical findings, patient history, radiographic interpretation, and histopathological features. Consequently, diagnostic reasoning in this field is inherently complex and often involves detailed differential diagnosis processes [17].

Given this complexity, oral pathology provides an ideal framework for evaluating the reasoning capabilities of LLMs. Recent studies have demonstrated that AI systems can achieve moderate to high performance in oral lesion classification and dental diagnostic tasks, particularly when structured prompts are used [18,19]. However, the comparative performance of different LLMs, particularly ChatGPT and newer models such as DeepSeek, remains underexplored in oral pathology education. Although several studies have investigated ChatGPT's performance in medical and dental examinations, most have focused on single-model evaluation rather than head-to-head comparisons between competing LLMs. Furthermore, evidence remains limited on their performance in complex clinical reasoning tasks specific to oral pathology, which require integrating multimodal clinical information. This represents a significant gap in the current literature.

Methods

Study Design

This study was designed as a cross-sectional comparative pilot study aimed at evaluating and comparing the performance of two large language models (LLMs), ChatGPT (OpenAI) and DeepSeek, in oral pathology education and evidence-based diagnostic decision support.

Study Setting and Period

The study was conducted between January and December 2025. All stages of the study, including prompt development, AI response generation, data extraction, anonymisation, and expert evaluation, were performed electronically using standardised procedures.

Study Materials

Educational Questions

Ten structured oral pathology educational questions were developed based on standard undergraduate and postgraduate curricula and authoritative textbooks, including *Neville's Oral and Maxillofacial Pathology*.

The questions covered the following conditions:

- Periapical abscess
- Radicular cyst
- Dentigerous cyst
- Oral candidiasis
- Oral leukoplakia
- Oral lichen planus
- Irritation fibroma
- Mucocele
- Recurrent aphthous stomatitis
- Pyogenic granuloma

Clinical Case Scenarios

Five standardised clinical case scenarios representing commonly encountered oral pathology conditions were developed. Each case included patient demographics, chief complaint, clinical description, relevant medical and dental history, and radiographic findings when applicable.

The cases included:

- Periapical abscess
- Radicular cyst
- Mucocele
- Oral leukoplakia

- Oral lichen planus

Artificial Intelligence Models

Two large language models were evaluated:

- ChatGPT (OpenAI, latest publicly available version at the time of data collection)
- DeepSeek (latest publicly available version at the time of data collection)

No model fine-tuning was performed. Both systems were accessed via their standard user interfaces.

Prompt Standardisation and Data Collection

To ensure comparability, identical prompts were used for both AI models. Each question and clinical case was submitted independently in separate chat sessions to avoid contextual bias or memory carryover. All generated responses were recorded verbatim and stored in a secure document. Responses were anonymised and coded as Response A and Response B for evaluation.

Expert Evaluation

Three board-certified oral pathologists, each with a minimum of five years of academic and clinical experience, independently evaluated all responses. Evaluators were blinded to the identity of the AI model generating each response. Discrepancies in scoring were resolved through consensus discussion. As this investigation was designed as a pilot study, no formal a priori sample size calculation was performed. The results should therefore be interpreted as exploratory and hypothesis-generating.

Evaluation Criteria

Each response was assessed using a structured scoring rubric covering five domains:

- Accuracy
- Completeness
- Clinical relevance
- Diagnostic reasoning
- Evidence-based consistency

Each domain was scored on a 5-point Likert scale (1 = poor, 5 = excellent). The maximum possible total score per response was 25. Scores assigned by the three oral pathologists were averaged for each response before statistical analysis.

Outcome Measures

Primary Outcome

- Total composite score (sum of all five domains)

Secondary Outcomes

- Individual domain scores:
 - Accuracy
 - Completeness
 - Clinical relevance
 - Diagnostic reasoning
 - Evidence-based consistency

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics version 29.0 (IBM Corp., Armonk, NY, USA). Scores assigned by the three oral pathologists were averaged for each response before statistical analysis. Descriptive statistics were reported as mean $\pm$ standard deviation (SD). The Shapiro–Wilk test was used to assess data normality. Since all variables demonstrated approximately normal distributions, comparisons between ChatGPT and DeepSeek were conducted using independent-samples t-tests. Mean differences, 95% confidence intervals (CIs), and p-values were calculated for all primary and secondary outcomes. Effect sizes were estimated using Cohen's d and interpreted as small (0.20), medium (0.50), and large (≥ 0.80). Inter-rater reliability among evaluators was assessed using the Intraclass Correlation Coefficient (ICC). ICC values below 0.50 were considered poor, 0.50–0.75 moderate, 0.75–0.90 good, and above 0.90 excellent reliability. As this investigation was designed as a pilot study, no formal a priori sample size calculation was performed. Therefore, findings should be interpreted as exploratory and hypothesis-generating. Given the exploratory nature of the study, p-values were not adjusted for multiple comparisons. Statistical significance was established at $p < 0.05$.

Ethical Considerations

This study did not involve human participants, animal subjects, or identifiable patient data. All clinical scenarios were anonymised and used strictly for educational and research purposes. The study adhered to ethical principles of scientific integrity, transparency, and responsible use of artificial intelligence in academic research.

Results

Overall Performance Comparison Between ChatGPT and DeepSeek

A total of 15 oral pathology items (10 educational questions and 5 clinical case scenarios) were evaluated. Responses generated by ChatGPT and DeepSeek were independently scored by three expert oral pathologists across five domains: accuracy, completeness, clinical relevance, diagnostic reasoning, and evidence-based consistency. ChatGPT demonstrated significantly higher overall performance compared with DeepSeek across all evaluated domains. The mean total composite score for ChatGPT was 23.2 ± 1.8 , compared with 20.6 ± 2.4 for DeepSeek ($t = 3.12$, $p = 0.006$, Cohen's $d = 1.15$), indicating a large effect size favouring ChatGPT.

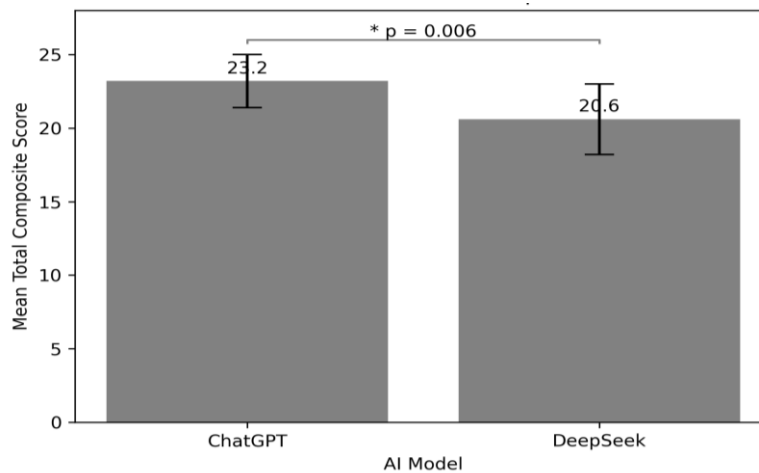


Figure 1. Overall Performance Comparison

Comparison of total composite scores between ChatGPT and DeepSeek. ChatGPT achieved a significantly higher mean composite score (23.2 ± 1.8) than DeepSeek (20.6 ± 2.4). An independent-samples t-test demonstrated a statistically significant difference between the two models ($t = 3.12$, $p = 0.006$), with a large effect size favouring ChatGPT (Cohen's $d = 1.15$). Error bars represent standard deviations.

Domain-Wise Performance Analysis

Significant differences were observed across all five evaluation domains:

- Accuracy ($p = 0.003$)
- Completeness ($p = 0.005$)
- Clinical relevance ($p = 0.011$)
- Diagnostic reasoning ($p = 0.008$)
- Evidence-based consistency ($p = 0.002$)

ChatGPT consistently achieved higher mean scores across all domains. Given the exploratory pilot nature of the study, p-values were not adjusted for multiple comparisons.

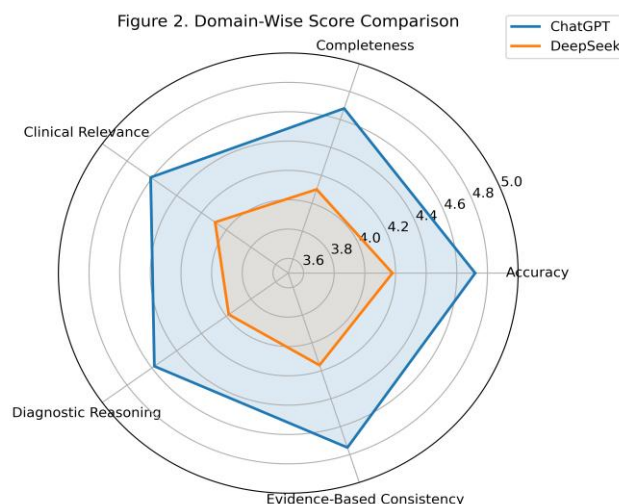


Figure 2. Domain-wise comparison of performance scores between ChatGPT and DeepSeek across five evaluation domains.

ChatGPT demonstrated consistently higher mean scores in Accuracy (4.72 vs. 4.18, $p = 0.003$), Completeness (4.68 vs. 4.10, $p = 0.005$), Clinical Relevance (4.61 vs. 4.09, $p = 0.011$), Diagnostic Reasoning (4.58 vs. 3.98, $p = 0.008$), and Evidence-Based Consistency (4.75 vs. 4.16, $p = 0.002$). The larger radar area occupied by ChatGPT indicates superior and more consistent performance across all evaluated domains.

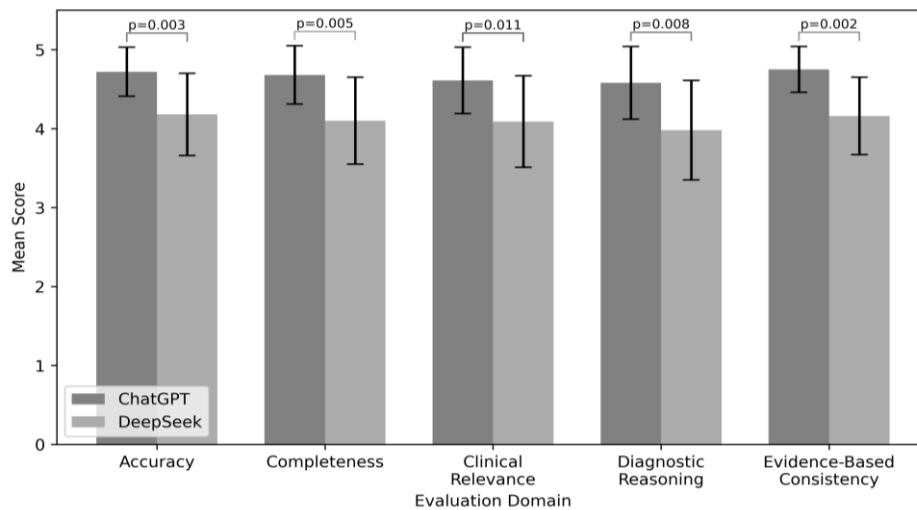


Figure 3. Comparison of domain-specific performance scores between ChatGPT and DeepSeek.

Bars represent mean scores and error bars indicate standard deviations. ChatGPT achieved significantly higher scores across all evaluation domains, including Accuracy ($p = 0.003$), Completeness ($p = 0.005$), Clinical Relevance ($p = 0.011$), Diagnostic Reasoning ($p = 0.008$), and Evidence-Based Consistency ($p = 0.002$). These findings demonstrate a consistent advantage of ChatGPT across all assessed dimensions of clinical response quality.

Table 1. Comparison of Domain Scores Between ChatGPT and DeepSeek

Domain	ChatGPT (Mean $\pm$ SD)	DeepSeek (Mean $\pm$ SD)	t-value	p-value	Effect Size (d)
Accuracy	4.72 $\pm$ 0.31	4.18 $\pm$ 0.52	3.21	0.003	1.18
Completeness	4.68 $\pm$ 0.37	4.10 $\pm$ 0.55	3.05	0.005	1.12
Clinical Relevance	4.61 $\pm$ 0.42	4.09 $\pm$ 0.58	2.72	0.011	0.98
Diagnostic Reasoning	4.58 $\pm$ 0.46	3.98 $\pm$ 0.63	2.89	0.008	1.05
Evidence-Based Consistency	4.75 $\pm$ 0.29	4.16 $\pm$ 0.49	3.45	0.002	1.25
Total Score	23.2 $\pm$ 1.8	20.6 $\pm$ 2.4	3.12	0.006	1.15

Accuracy Performance

ChatGPT achieved significantly higher accuracy scores compared with DeepSeek (4.72 $\pm$ 0.31 vs. 4.18 $\pm$ 0.52, $p = 0.003$). Responses generated by ChatGPT demonstrated stronger alignment with established oral pathology diagnostic criteria and reduced occurrence of incomplete or partially incorrect explanations.

Completeness of Responses

In terms of completeness, ChatGPT again outperformed DeepSeek (4.68 $\pm$ 0.37 vs. 4.10 $\pm$ 0.55, $p = 0.005$). ChatGPT responses were more comprehensive, frequently including:

- Expanded differential diagnoses
- Detailed etiological explanations
- Structured clinical reasoning pathways

Clinical Relevance

ChatGPT demonstrated higher clinical relevance scores (4.61 $\pm$ 0.42 vs. 4.09 $\pm$ 0.58, $p = 0.011$). Its responses were more closely aligned with practical dental decision-making, particularly in management-oriented questions and clinical case interpretation.

Diagnostic Reasoning

Significant differences were observed in diagnostic reasoning performance. ChatGPT achieved a mean score of 4.58 $\pm$ 0.46, whereas DeepSeek achieved 3.98 $\pm$ 0.63 ($p = 0.008$). ChatGPT showed stronger ability to:

- Integrate clinical and radiographic information
- Generate structured differential diagnoses

- Prioritise likely conditions

DeepSeek showed greater variability in reasoning structure and occasional omission of intermediate diagnostic steps.

Evidence-Based Consistency

Evidence-based consistency was the strongest domain for ChatGPT (4.75 ± 0.29 vs. 4.16 ± 0.49 , $p = 0.002$). ChatGPT responses were more consistent with established oral pathology textbooks and current clinical guidelines.

Inter-Rater Reliability

Inter-rater reliability analysis demonstrated excellent agreement among evaluators:

- ICC = **0.87** (95% CI: 0.81–0.92)

This indicates high consistency among the three oral pathologists in scoring AI-generated responses. An ICC value greater than 0.75 was considered indicative of good reliability, while values above 0.90 indicated excellent agreement

- ChatGPT significantly outperformed DeepSeek across all domains
- Largest differences observed in:
 - Evidence-based consistency
 - Accuracy
- Effect sizes were large (Cohen's $d > 0.8$ in all domains)
- Strong inter-rater reliability supports robustness of scoring system

Discussion

The present pilot comparative study evaluated the performance of ChatGPT and DeepSeek in oral pathology education and evidence-based diagnostic decision support. The findings demonstrated that ChatGPT significantly outperformed DeepSeek across all evaluated domains, including accuracy, completeness, clinical relevance, diagnostic reasoning, and evidence-based consistency. These results align with the growing body of literature suggesting that large language models (LLMs), particularly GPT-based systems, demonstrate strong performance in healthcare education and knowledge-based clinical tasks [18,19]. The superior performance of ChatGPT observed in this study is consistent with previous research showing that LLMs can achieve high accuracy in medical knowledge assessment and licensing examination scenarios. Gilson et al. and Kung et al. reported that ChatGPT performed at or near passing standards on medical examinations such as the USMLE, highlighting its ability to process and apply complex biomedical knowledge [18,20].

Similarly, GPT-based models have been shown to perform well in structured question-answering tasks requiring conceptual understanding and reasoning [21]. In dental education specifically, recent studies have confirmed the increasing role of LLMs in supporting learning, knowledge synthesis, and examination preparation [22,23]. These findings support the present study, where ChatGPT demonstrated superior accuracy and completeness compared to DeepSeek in oral pathology tasks. Oral pathology represents one of the most cognitively demanding domains in dentistry due to the need for integration of clinical, radiographic, and histopathological data. The higher diagnostic reasoning scores observed for ChatGPT are consistent with previous studies demonstrating the ability of GPT-based systems to support diagnostic decision-making and oral lesion assessment [24,25]. Similarly, previous investigations have reported that LLMs demonstrate moderate to high performance in dentistry-related diagnostic reasoning, particularly when structured prompts are used [24,25]. The ability of ChatGPT to synthesise clinical features and generate differential diagnoses in the present study reflects its transformer-based architecture, which enables contextual reasoning across extensive textual datasets [26].

Evidence-based consistency was one of the strongest domains for ChatGPT in this study. This finding aligns with previous investigations demonstrating that LLMs encode substantial clinical knowledge derived from biomedical literature and can provide clinically relevant responses aligned with established medical guidelines [18,19,27]. However, despite these strengths, several studies emphasise that LLMs may generate plausible but incorrect outputs, highlighting the importance of human oversight and expert validation [28,29]. This limitation is particularly relevant in oral pathology, where diagnostic inaccuracies may have significant clinical consequences. The higher clinical relevance scores achieved by ChatGPT support its potential use as a supplementary educational tool in dental curricula. Previous studies have shown that ChatGPT can facilitate understanding of complex biomedical concepts and improve access to educational resources [30,31]. Furthermore, systematic reviews in health professions education suggest that LLMs can enhance self-directed learning and student engagement while providing rapid access to summarised information [22,32]. Nevertheless, concerns regarding overreliance, misinformation, and reduced critical appraisal skills remain important considerations [31]. The observed superiority of ChatGPT over DeepSeek in this study is consistent with comparative evaluations conducted in other healthcare-related domains, where GPT-based models frequently demonstrate stronger reasoning capabilities and more comprehensive responses [31,32,33].

This difference may be attributed to extensive training datasets, reinforcement learning from human feedback (RLHF), and optimisation strategies implemented during model development [26,27]. Such characteristics may contribute to enhanced performance in complex clinical reasoning tasks.

Limitations of LLM-Based Diagnostic Tools

Despite promising results, LLMs remain subject to important limitations. Hallucination, defined as the generation of plausible but inaccurate information, remains a recognised challenge [28]. Furthermore, model performance can vary depending on prompt design, context specificity, and task complexity [23]. Additionally, current LLMs do not possess genuine clinical understanding but instead rely on probabilistic prediction based on learned language patterns. This limitation underscores the importance of maintaining human oversight in all clinical applications. The present study is also limited by its pilot design and relatively small sample size, which may affect generalizability. Similar limitations have been reported in previous early-stage evaluations of LLMs in dentistry and healthcare education [23,24].

Implications for Dental Education and Clinical Practice

The findings suggest that ChatGPT may serve as a valuable adjunctive tool in oral pathology education by supporting student learning and enhancing exposure to diagnostic reasoning processes. Similar recommendations have been proposed in recent studies examining AI integration within health professions education [30,32]. Nevertheless, AI-generated information should be critically appraised and verified using authoritative textbooks, clinical guidelines, and peer-reviewed literature. Clinical implementation should be approached cautiously, with LLMs functioning as decision-support tools rather than substitutes for professional judgment [29].

Future Research Directions

Future investigations should evaluate LLM performance using larger datasets, authentic clinical cases, and image-based oral pathology scenarios. Emerging multimodal AI systems capable of integrating textual and visual information may offer additional opportunities for improving oral pathology education and diagnostic support. Longitudinal studies examining the impact of AI-assisted learning on student performance, knowledge retention, and clinical competency are also needed to establish the long-term educational value of these technologies [30,31,32,33,34].

Conclusion

This pilot study demonstrated that ChatGPT significantly outperformed DeepSeek across all evaluated domains of oral pathology education and evidence-based diagnostic decision support, including accuracy, completeness, clinical relevance, diagnostic reasoning, and evidence-based consistency. The findings suggest that ChatGPT may serve as a valuable adjunctive educational and clinical support tool in dentistry. However, both systems remain susceptible to limitations inherent to large language models, including potential inaccuracies and hallucinations. Consequently, AI-generated outputs should always be interpreted under professional supervision and verified against authoritative evidence-based sources. Further large-scale studies incorporating real-world clinical cases and multimodal diagnostic data are required to establish the role of large language models in dental education and clinical practice. Although the findings favour ChatGPT, the rapid evolution of large language models necessitates continuous re-evaluation, as comparative performance may change substantially with future model updates.

Conflict of interests

The authors declare no conflicts of interest.

Funding sources

No funding.

Authorship

All authors contributed to the study conception and design. Material preparation, data collection, analysis, and interpretation were performed by the authors. The first draft of the manuscript was written by the first author, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to thank all participants who contributed to this study. Their participation and cooperation are gratefully acknowledged.

References

1. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. Stanford (CA): Stanford CRFM; 2021.
2. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:5998–6008.
3. OpenAI. GPT-4 technical report. *arXiv*. 2023;2303.08774.
4. Nori H, King N, McKinney SM, et al. Capabilities of GPT-4 on medical challenge problems. *arXiv*. 2023;2303.13375.
5. Singhal K, Azizi S, Tu T, et al. Towards expert-level medical question answering with large language models. *Nature*. 2023;617(7964):512–8.
6. Moor M, Banerjee O, Abad ZS, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616:259–65.
7. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med*. 2024;30:1773–84.

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(2):e0000198.
9. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination? *JMIR Med Educ*. 2023;9:e45312.
10. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination? *J Educ Eval Health Prof*. 2023;20:1.
11. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023;11(6):887.
12. Tili A, Shehata B, Adarkwah MA, et al. What if the devil is my guardian angel? ChatGPT as a case study of using chatbots in education. *Smart Learn Environ*. 2023;10:15.
13. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ*. 2023;103:102274.
14. Lo CK. What is the impact of ChatGPT on education? A rapid review of the literature. *Educ Sci*. 2023;13:410.
15. Dwivedi YK, Kshetri N, Hughes L, et al. So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int J Inf Manage*. 2023;71:102642.
16. Lund BD, Wang T, Mannuru NR, et al. ChatGPT and a new academic reality. *J Assoc Inf Sci Technol*. 2023.
17. Biswas S. ChatGPT and the future of medical writing. *Radiology*. 2023;307(2):e223312.
18. Rao A, Kim J, Kamineni M, et al. Evaluating ChatGPT as an adjunct for radiologic decision-making. *Radiology*. 2023;307(5):e230020.
19. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44–56.
20. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25:24–9.
21. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28:31–8.
22. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. *J Dent Res*. 2020;99(7):769–74.
23. Khanagar SB, Al-Ehaideb A, Maganur PC, et al. Developments, application, and performance of artificial intelligence in dentistry: a systematic review. *J Dent Sci*. 2021;16(1):508–22.
24. Hung K, Montalvao C, Tanaka R, et al. The use and performance of artificial intelligence applications in dental and maxillofacial radiology: a systematic review. *Dentomaxillofac Radiol*. 2020;49(1):20190107.
25. Schwendicke F, Golla T, Dreher M, Krois J. Convolutional neural networks for dental image diagnostics: a scoping review. *J Dent*. 2019;91:103226.
26. Regezi JA, Sciubba JJ, Jordan RCK. Oral pathology: clinical pathologic correlations. 7th ed. St. Louis (MO): Elsevier; 2017.
27. Neville BW, Damm DD, Allen CM, Chi AC. Oral and maxillofacial pathology. 5th ed. St. Louis (MO): Elsevier; 2024.
28. Kumar V, Abbas AK, Aster JC. Robbins and Cotran pathologic basis of disease. 10th ed. Philadelphia (PA): Elsevier; 2020.
29. Topol EJ. Deep medicine: how artificial intelligence can make healthcare human again. New York (NY): Basic Books; 2019.
30. Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. *Front Med (Lausanne)*. 2020;7:27.
31. Ahmed KA, Alhendwi NY. Diagnostic challenges in dental pathology: case-based insights for practitioners (case study review). *Scholastic Dental Sciences*. 2025;1(2):01–3.
32. Ahmeid KA. Dentists' knowledge, diagnostic confidence, and clinical decision-making regarding oral mucosal lesions and potentially malignant disorders: a cross-sectional survey in Libya and other countries. *African Journal of Advanced Pure and Applied Sciences*. 2026;5(3):500–7.
33. Temur KT. Performance comparison of artificial intelligence-based language models on oral pathology questions in dental education. *Journal of Dental Sciences and Education*. 2025;3(2):31–4.
34. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. *BMC Oral Health*. 2025;25:573.