Journal of Dental Sciences 20 (2025) 1904–1907

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Short Communication

DeepSeek: Another step forward in the diagnosis of oral lesions

Márcio Diniz-Freitas*, Pedro Diz-Dios

Medical-Surgical Dentistry Research Group (OMEQUI), Health Research Institute of Santiago de Compostela (IDIS), University of Santiago de Compostela (USC), Santiago de Compostela, Spain

Received 23 February 2025; Final revision received 25 February 2025
Available online 9 March 2025

KEYWORDS
Large language models;
ChatGPT4o;
DeepSeek;
Oral medicine

Abstract Artificial intelligence (AI) is increasingly being explored as a tool for medical diagnosis, particularly in fields with limited specialized training, such as oral medicine. This study evaluates the performance of DeepSeek-R1, an open-source large language model (LLM), in diagnosing oral diseases and conditions using text-based case descriptions from the New England Journal of Medicine's "Image Challenge." Results indicate that DeepSeek-R1 achieved a diagnostic accuracy of 91.6 %, slightly outperforming ChatGPT-4o (88.9 %) and significantly exceeding the 47.8 % accuracy of the journal' readers. While DeepSeek lacks direct image interpretation capabilities, it demonstrates high proficiency in textual diagnostic tasks. These findings suggest that DeepSeek-R1 could be a valuable aid for medical professionals in diagnosing oral diseases.

Introduction

Medical diagnosis is a challenge for which artificial intelligence (AI) can serve as a powerful supportive tool, even for diagnosing clinically complex cases.[1] The capacity of large language models (LLMs) to process and generate human-like text in medical and healthcare-related tasks has been, and continues to be, the subject of investigation by medical professionals and healthcare researchers.[2]

While ChatGPT 3.5 was the first LLM available to the public, other advanced models have since been developed, such as Claude, Gemini (Google), Copilot (Microsoft), Claude (Anthropc), Mistral (Mistral AI), Llama (Meta) and, more recently, DeepSeek (Deepsek). The Chinese-developed LLM, DeepSeek-R1, is generating significant interest among researchers as a cost-effective and open-source alternative to advanced reasoning models like OpenAI's. Due to its strong performance and affordability, DeepSeek-R1 is expected to encourage a broader adoption of LLMs among researchers, enabling their integration into daily scientific workflows without cost-related constraints.[3] DeepSeek has gained recognition for developing a series of

* Corresponding author. Special Care Dentistry Unit, School of Medicine and Dentistry, University of Santiago de Compostela. Calle entrerríos s/n, Santiago de Compostela -15782, Spain.
E-mail address: marcio.diniz@usc.es (M. Diniz-Freitas).

https://doi.org/10.1016/j.jds.2025.02.023

ambitious and highly efficient LLMs that, while similar to OpenAI's ChatGPT, exhibit comparatively lower processing power. Unlike ChatGPT and most of its Western counterparts, DeepSeek's LLMs are open-source, allowing users to access, modify, and customize the source code to enhance functionality and adaptability.[4]

Physicians recognize their limited training in oral health and their difficulties in recognizing oral diseases.[5] This situation can be particularly relevant when it comes to identifying oral manifestations of systemic diseases and preventing diagnostic delays in oral cancer.[6] As a result, LLMs could be applied in a medical field with great training deficits as oral medicine. In a previous study, we have demonstrated that ChatGPT-4V can be useful for facilitating the diagnosis of oral diseases and conditions from the image challenges published in the New England Journal of Medicine.[7]The aim of this study was to assess the performance of the new DeepSeek for the diagnosis of oral diseases and conditions.

Materials and methods

Using the same methodology of the previously cited work,[7] we re-evaluated the "Image Challenges" corresponding to oral diseases and conditions published in the New England Journal of Medicine (NEJM) (https://www.nejm.org/image-challenge). As DeepSeek currently lacks the capability to directly analyze or interpret images, the performance of both LLMs, ChatGPT-4o- (OpenAI, San Francisco, CA, USA) and DeepSeek-R1 (DeepSeek Inc., Hangzhou, China), was evaluated using only written information with each case being presented with a written description and a multiple-choice question offering five possible diagnoses as proposed by the journal. The percentage of correct responses was then compared to those provided by the journal's readers, recording both the number and distribution of votes for each case.

Of the 52 cases initially reviewed, 36 were selected based on the inclusion of an image accompanied by explanatory text. Each selected case comprised a brief clinical description, relevant images, five diagnostic options, and the question: "What is the diagnosis?" To minimize the risk of memorization, each query was conducted in a new session.

Results

Of the 36 cases ultimately selected, ChatGPT-4o and DeepSeek-R1 achieved correct diagnoses in 88.9 % and 91.6 % of the cases, respectively. In comparison, the journal's regular readers (averaging 97,944 participants per case; range, 33,280 to 176,981 readers) had an average accuracy rate of 47.8 %, with correct responses ranging from 34 % to 83 % per case (Table 1).

Table 1 Comparison of diagnostic accuracy among NEJM readers, ChatGPT-4o, and DeepSeek-R1 in 36 selected NEJM image challenge cases involving oral diseases and conditions.

Case identification/Correct answer	NEJM readers % correct answers/ (Total responses)	ChatGPT 4o textCorrect/ Incorrect	DeepSeek text Correct/Incorrect
March 09, 2023/**Malignant acanthosis nigricans**	54% (44548)	Correct	Correct
February 02, 2023/**Odontogenic cutaneous fistula**	83% (33280)	Correct	Correct
November 17, 2022/**Pyogenic granuloma**	37% (69216)	Correct	Correct
May 19, 2022/**Keratoderma blenorrhagicum**	58% (50178)	Correct	Correct
January 06, 2022/**Vitamin B12 deficiency**	57% (60791)	Correct	Correct
December 30, 2021/**Leukemic infiltration of the gingiva**	57% (101058)	Correct	Correct
December 09, 2021/**Giant-cell arteritis**	47% (72929)	Correct	Correct
November 18, 2021/**Pyostomatitis vegetans**	43% (87698)	Correct	Correct
November 11, 2021/**Gingival melanoma**	56% (71115)	Correct	Correct
June 24, 2021/**Thromboembolism**	39% (126692)	Incorrect	Incorrect
May 13, 2021/**Leukoplakia**	66% (59817)	Correct	Correct
March 18, 2021/**Streptococcal pharyngitis**	51% (97905)	Correct	Correct
August 16, 2020/**Metastasis of colorectal cancer**	42% (91079)	Correct	Correct
July 23, 2020/**Extranodal natural killer T-cell lymphoma**	36% (143612)	Correct	Correct
June 25, 2020 **Cytomegalovirus infection**	40% (93763)	Correct	Correct
June 18, 2020 **Uremic stomatitis**	34% (137667)	Incorrect	Incorrect
June 11, 2020 **Tissue plasminogen activator (tPA)-associated angioedema**	47% (89215)	Correct	Correct
December 19, 2019 **Hereditary hemorrhagic telangiectasia**	41% (105242)	Correct	Correct

(*continued on next page*)

Table 1 (*continued*)

Case identification/Correct answer	NEJM readers % correct answers/ (Total responses)	ChatGPT 4o textCorrect/ Incorrect	DeepSeek text Correct/Incorrect
October 19, 2019	57% (83275)	Correct	Correct
Pernicious anemia			
July 18, 2019	46% (89696)	Correct	Correct
Tonsillar cancer			
July 11, 2019	34% (137752)	Correct	Correct
Multiple endocrine neoplasia type 2B (MEN 2B).			
July 4, 2019	54% (79586)	Correct	Correct
Ludwig's angina			
May 30, 2019	49% (81373)	Correct	Correct
Disseminated coccidioidomycosis			
February 28, 2019	43% (142708)	Correct	Correct
Group A streptococcus			
February 21, 2019	52% (114708)	Correct	Correct
Spindle-cell sarcoma			
January 24, 2019	60% (77356)	Correct	Correct
Peutz—Jeghers syndrome			
November 1, 2018	53% (116835)	Correct	Correct
Lead poisoning			
September 20, 2018	38% (176981)	Correct	Correct
***Mycoplasma pneumoniae*-associated mucositis**			
October 5, 2017	46% (116480)	Correct	Correct
Cowden syndrome			
July 21, 2016	76 % (102283)	Correct	Correct
Varicella-zoster virus			
May 12, 2016	45% (93268)	Incorrect	Correct
Langerhans'-cell histiocytosis			
February 11, 2016	54% (97514)	Correct	Correct
Geographic tongue			
January 21, 2016	58% (85606)	Correct	Correct
***Treponema pallidum* infection**			
January 1, 2015	40% (121117)	Incorrect	Incorrect
Exanthematous pustulosis			
July 24, 2014 —	41% (98333)	Correct	Correct
Amyotrophic lateral sclerosis			
June 26, 2014-**Zinc**	39% (88153)	Correct	Correct
TOTAL (36 cases)	**Correct 47.8% (range: 34—83%); mean: 97.944 (range: 32,280 —176,981)**	**Correct (32/36; 88.9%)**	**Correct (33/36; 91.6%)**

Discussion

In this study, ChatGPT-4o and DeepSeek-R1demonstrated high diagnostic accuracy.

Currently, DeepSeek is unable to directly analyze or interpret images. However, it may provide relevant information if users describe visual characteristics of a condition, such as its appearance, location, color, associated symptoms, and duration, in textual form (DeepSeek, 2025. https://www.deepseek.com). Although ChatGPT-4o offers this capability, a previous study found that adding images to case descriptions did not improve its diagnostic accuracy.[7]

Image-based diagnosis alone showed limited sensitivity (and specificity for detecting oral cancer; however, diagnostic accuracy improved significantly when clinical history was included in the prompt.[8] Pradhan[9] evaluated the diagnostic accuracy of ChatGPT 3.5, 4.0, 4o, and Gemini in identifying oral potentially malignant lesions using both text-based case reports and image recognition. Based on 42 case reports and publicly available images, it was found that GPT-4o achieved the highest accuracy (65 %), followed by GPT-4.0 (47 %), GPT-3.5 (42 %), and Gemini (35 %). Moreover, for image-based diagnosis, GPT-4o correctly identified 66 % of cases, outperforming Gemini (45 %).[9] ChatGPT-4 has also been evaluated as a diagnostic tool by testing its performance on 50 complex cases from the Journal of the American Dental Association's Diagnostic Challenge section (2017—2024); using the text-based interpretations alone,

ChatGPT-4 achieved a 62 % accuracy rate; however, when provided with five possible differential diagnoses from the literature, its accuracy rate improved to 80 %.[10]

The potential benefits and limitations of ChatGPT are currently under scientific investigation, with no established consensus on its various applications. Our data indicate that the current image recognition capabilities of LLMs are insufficient for interpreting images in specialized medical fields. Another potential application could be to develop and support differential diagnosis, saving time and reducing the need for additional diagnostic tests.[10]

Considering the inherent limitations of this pilot study—such as the sample size, the provision of possible answers, and the absence of a well-defined comparator group, given the diverse expertise within the journal's readership—it appears that DeepSeek is a promising AI tool for facilitating the diagnosis of oral diseases and conditions. It demonstrates diagnostic proficiency that surpasses that of the medical community, with significant potential for improvement upon incorporating image interpretation capabilities.

Declaration of competing interest

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

Acknowledgments

The authors received no financial support for the research, authorship, and/or publication of this article.

References

1. Eriksen AV, Möller S, Ryg J. Use of GPT4 to diagnose complex clinical cases. *NEJM AI* 2024;1. AIp.2300031.
2. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. *Commun Med* 2025;5:26.
3. Gibney E. Scientists flock to DeepSeek: how they're using the blockbuster AI model. *Nature*. Published online January 29, 2025.
4. Normile D. Chinese firm's large language model makes a splash. *Science* 2025;387:238.
5. Stephens MB, Wiedemer JP, Kushner GM. Dental problems in primary care. *Am Fam Physician* 2018;98:654—60.
6. Shimpi N, Schroeder D, Kilsdonk J, et al. Medical providers' oral health knowledgeability, attitudes, and practice behaviors: an opportunity for interprofessional collaboration. *J Evid Base Dent Pract* 2016;16:19—29.
7. Diniz-Freitas M, Lago-Méndez L, Limeres-Posse J, Diz-Dios P. Challenging ChatGPT-4V for the Diagnosis of Oral Diseases and Conditions. *Oral Dis*. Published online October 25, 2024.
8. Schmidl B, Hütten T, Pigorsch S, et al. Artificial intelligence for image recognition in diagnosing oral and oropharyngeal cancer and leukoplakia. *Sci Rep* 2025;15:3625.
9. Pradhan P. Accuracy of ChatGPT 3.5, 4.0, 4o and Gemini in diagnosing oral potentially malignant lesions based on clinical case reports and image recognition. *Med Oral Patol Oral Cir Bucal* 2025;30(2):e224—31.
10. Danesh A, Danesh A, Danesh F. Innovating dental diagnostics: ChatGPT's accuracy on diagnostic challenges. *Oral Dis*. Published online July 22, 2024.