JCTEI
JOURNAL OF CLINICAL TRIALS AND EXPERIMENTAL INVESTIGATIONS
Year: 2024 Volume: 3 Issue: 4

ORIGINAL ARTICLE

10.5281/ zenodo.14617672

Artificial intelligence in Otorhinolaryngology practice: Comparative performance of ChatGPT and Gemini AI

Ahmet Celik[1]

1. Silopi State Hospital, Department of Otorhinolaryngology, Sirnak, Turkey

Correspondence
Ahmet Celik, Silopi State Hospital, Department of Otorhinolaryngology, Sirnak, Turkey.
e-mail
ahmetcelikk1982@gmail.com

Received: 2 November 2024
Revised: 24 December 2024
Accepted: 25 December 2024
Published: 30 December 2024

Keywords
- Artificial intelligence
- ChatGPT
- Gemini
- Otorhinolaryngology

ORCID ID of the author(s):
0000-0002-6192-1546

Abstract

Objective: This study aims to evaluate the accuracy of ChatGPT and Gemini AI in the field of otorhinolaryngology.

Materials and methods: This study evaluated the performance of ChatGPT 4.0 and Gemini AI in answering 150 multiple-choice questions evenly distributed across otorhinolaryngology domains: ear, nose, and throat. Both models were tested under standardized conditions, with their responses compared to an answer key. The true and false answers were evaluated.

Results: For ear-related questions, ChatGPT correctly answered 34 (68%), while Gemini AI correctly answered 33 (66%) (p=0.832). For nose-related questions, both models achieved identical results: 34 correct answers (68%) and 16 incorrect answers (32%) (p=1.000). For throat-related questions, ChatGPT provided 34 correct answers (68%) compared to Gemini AI's 38 correct answers (76%) (p=0.373). Overall, ChatGPT achieved 102 correct answers (68%) and Gemini AI achieved 105 (70%), with no statistically significant difference between the models (p=0.708). The total correct answers across all topics were 207 (69%), and incorrect answers were 91 (31%). Binary logistic regression showed no significant differences in performance between the AI models or topics, confirming their comparable accuracy in otorhinolaryngology question sets.

Conclusion: ChatGPT 4.0 and Gemini AI demonstrated comparable performance in answering otorhinolaryngology questions, with no statistically significant differences observed across ear, nose, and throat topics. Both models achieved high accuracy rates (ChatGPT: 68%, Gemini AI: 70%), suggesting their potential applicability in clinical decision-making and supporting otorhinolaryngology-related diagnostics.

Cite as: : Celik A. Artificial intelligence in ENT practice: Comparative performance of ChatGPT and Gemini AI. J Clin Trials Exp Investig. 2024;3(4):156-162

Introduction

The integration of artificial intelligence (AI) into medicine is fundamentally transforming the healthcare landscape, offering unprecedented opportunities to enhance diagnosis, treatment, and patient care (1-4). Powered by advanced technologies such as machine learning, deep learning, and natural language processing (NLP), AI has demonstrated remarkable capacity to analyze vast and complex medical data with unparalleled speed and accuracy (5,6). This evolution positions AI as a pivotal tool in addressing critical challenges in modern healthcare, including diagnostic delays, inefficiencies in treatment, and limited accessibility in resource-constrained settings (7,8).

AI applications in healthcare have achieved significant milestones. For instance, diagnostic imaging systems powered by AI have exhibited the ability to detect conditions like cancer, cardiovascular diseases, and neurological disorders with precision that rivals and occasionally surpasses human expertise (9). Personalized medicine has also benefited greatly from AI algorithms, which enable predictive modeling of treatment responses based on individual patient data, leading to more targeted and effective therapies (10,11).

Recent advancements in NLP, particularly in models such as ChatGPT, Gemini Advanced, and Co-Pilot, have further expanded AI's impact in medicine (12). These state-of-the-art models are capable of advanced text comprehension and generation, making them invaluable for analyzing medical literature and improving communication between healthcare providers and patients. By synthesizing and contextualizing vast amounts of information, these models facilitate informed decision-making and enhance the efficiency of clinical workflows (13,14).

Guerra et al. reported an accuracy of 0.77 for GPT models in the field of neurosurgery (15), while Huang et al. found a slightly higher accuracy of 0.82 (16). Similarly, Waldock et al. assessed the performance of large language models (LLMs) on the USMLE, analyzing 14 sub-studies that included 13,535 questions, and reported an overall accuracy of 0.51 (CI 0.46–0.56) (17). Building on these findings, this study aims to evaluate the accuracy of ChatGPT and Gemini AI in the field of otorhinolaryngology.

Materials and methods

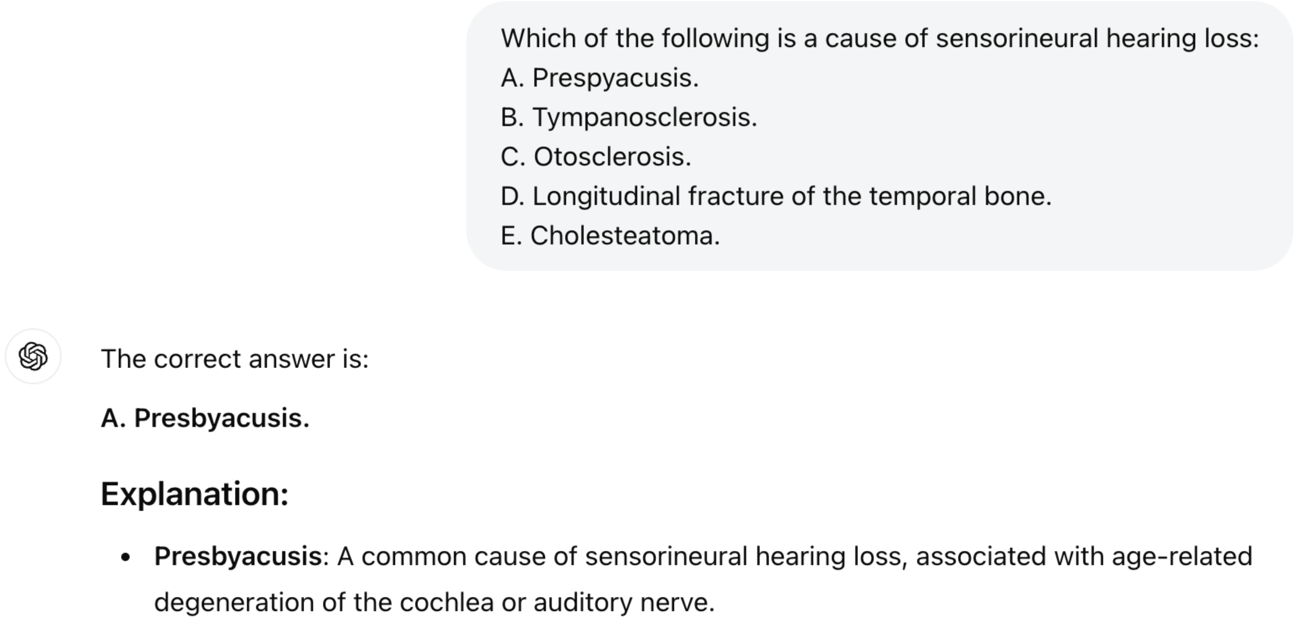
This study was conducted in September 2024 - October 2024. As it focused on AI-based analysis, ethical approval was deemed unnecessary. This study utilized two advanced artificial intelligence models, Gemini AI and ChatGPT 4.0, to evaluate their performance in answering questions the three fields of otorhinolaryngology: ear, nose, and throat. A total of 150 questions were selected, with 50 questions chosen randomly from each domain to ensure a diverse and representative dataset. All questions were multiple-choice, each having a single correct answer.

Prior to presenting the questions to the AI models, both systems were configured to operate in a "Multiple-Choice Question (MCQ)" format, ensuring they selected only one answer option per question. This setup eliminated the possibility of ambiguity in their responses and standardized the evaluation process.

The 150 questions were simultaneously asked both Gemini AI and ChatGPT 4.0. Their responses were collected and recorded, noting the number of correct and incorrect answers for each model. The correctness of their answers was determined by comparing their selections against a predetermined answer key.

Before starting the study, both ChatGPT 4.0 and Gemini AI were provided with instructions, as outlined in the text below, to ensure accurate and consistent results. “To ChatGPT and Gemini AI: I am conducting a detailed study to evaluate the accuracy and agreement between ChatGPT and Gemini AI in the field of otorhinolaryngology, covering three major areas: ear, nose, and throat. The goal of this research is to compare the performance of these AI models in answering clinically relevant questions derived from established exam preparation materials for otorhinolaryngology. For this purpose, I will present each of you (AI models) with a total of 150 questions, distributed equally across the three topics, with 50 questions for each area. The same set of questions will be provided to both models. These questions are randomly selected from the otorhinolaryngology exam preparation book to ensure they reflect real-world clinical scenarios and standard diagnostic practices. Each question is in a multiple-choice format, with only one correct answer. To maintain consistency, I request that you select only one option per question as your response. This requirement is essential for the analysis and will prevent ambiguities in the results. Please approach these questions with your best effort and answer as accurately as possible. Once all responses have been recorded, your correct and incorrect answers will be evaluated against the official answer key. The performance of ChatGPT and Gemini AI will then be compared, focusing on accuracy rates and the level of agreement between your answers."

A sample question with explanation was given in Figure 1.



Which of the following is a cause of sensorineural hearing loss:
A. Prespyacusis.
B. Tympanosclerosis.
C. Otosclerosis.
D. Longitudinal fracture of the temporal bone.
E. Cholesteatoma.

The correct answer is:

A. Presbyacusis.

Explanation:

- **Presbyacusis**: A common cause of sensorineural hearing loss, associated with age-related degeneration of the cochlea or auditory nerve.

Figure 1: An answer of ChatGPT

Statistical analysis

Data were analyzed using descriptive statistics and inferential methods. Results were expressed as number, percentages, and/or odds ratios (OR). Accuracy rates for ChatGPT 4.0 and Gemini AI were compared using chi-square tests for categorical variables. Binary logistic regression was employed to assess differences in performance between models and across topics (ear, nose, throat). Statistical significance was set at p<0.05. All analyses were conducted using Jamovi software.

Results

A total of 150 questions were included in the study, all of which were posed to both Gemini AI and ChatGPT 4.0. These questions were divided into three main categories: 33% (n=50) focused on Ear-related topics, 33% (n=50) on Nose-related topics, and 33% (n=50) on Throat-related topics (Table 1).

For Ear-related questions, ChatGPT provided 34 correct answers (68%) and 16 incorrect answers (32%), while Gemini AI provided 33 correct answers (66%) and 17 incorrect answers (34%). The total number of correct answers was 67 (67%), and incorrect answers were 33 (33%), with no significant difference between the two models (p=0.832). For Nose-related questions, both ChatGPT and Gemini AI performed identically, each providing 34 correct answers (68%) and 16 incorrect answers (32%), resulting in a total of 68 correct answers (68%) and 32 incorrect answers (32%), with no significant difference observed (p=1.000). In Throat-related questions, ChatGPT provided 34 correct answers (68%) and 16 incorrect answers (32%), whereas Gemini AI provided 38 correct answers (76%) and 12 incorrect answers (24%). The total number of

Table 1: Question distribution

Number of questions	150
Topics	
Ear	50
Nose	50
Throat	50
AI Model	
Gemini AI	150
ChatGPT 4.o	150

AI: Artificial intelligence

correct answers was 72 (72%), and incorrect answers were 28 (28%), with no statistically significant difference between the models (p=0.373). When considering all topics combined, ChatGPT provided 102 correct answers (68%) and 48 incorrect answers (32%), while Gemini AI provided 105 correct answers (70%) and 45 incorrect answers (30%). The overall total number of correct answers was 207 (69%), and incorrect answers were 91 (31%), with no significant difference between the two models (p=0.708) (Table 2 and Figure 2).

Table 2: Results of the AI answers

		ChatGPT	Gemini AI	Total	P-value
Ear	False	16 (32%)	17 (34%)	33 (33%)	0.832
	True	34 (68%)	33 (66%)	67 (67%)	
Nose	False	16 (32%)	16 (32%)	32 (32%)	1.000
	True	34 (68%)	34 (68%)	68 (68%)	
Throat	False	16 (32%)	12 (24%)	28 (28%)	0.373
	True	34 (68%)	38 (76%)	72 (72%)	
Total	False	48 (32%)	45 (30%)	91 (31%)	0.708
	True	102 (68%)	105 (70%)	207 (69%)	

AI: Artificial intelligence

The binary logistic regression analysis showed no significant differences between topics (nose, throat) or AI models (ChatGPT vs. Gemini AI) in performance (Table 3).

Table 3: Binary Logistic regression analysis results for prediction of correct answers

Predictors	Estimate	SE	Z	OR	p-value
Topic (ref: ear)					
Nose	0.045	0.302	0.151	1.053	0.880
Throat	0.236	0.308	0.767	1.271	0.443
AI model (ref:ChatGPT)					
Gemini AI	0.093	0.350	0.374	1.101	0.708

AI: Artificial intelligence, SE: Standard Error, OR: Odss Ratio

Discussion

This study compared the accuracy of ChatGPT 4.0 and Gemini AI in answering 150 otorhinolaryngology questions evenly distributed across three domains: ear, nose, and throat. Both models demonstrated similar performance, with no statistically significant differences in accuracy across topics or overall. ChatGPT achieved 68% accuracy, while Gemini AI scored 70%. Binary logistic regression confirmed comparable performance between models and across topics, highlighting their potential for clinical decision support.

Artificial intelligence and LLM models, such as ChatGPT and Gemini AI, have gained significant popularity in recent years. The introduction of ChatGPT and the advancements of GPT-based models in the medical field have enabled their use as essential reference tools in clinical practice (2,3). Huang et al. concluded that GPT-4 outperformed GPT-3.5 and Family Medicine residents on a multiple-choice medical test, demonstrating superior accuracy (82.4%), logical reasoning, and potential for enhancing medical education tools (16).

The meta-analysis of Waldock et al suggests that LLMs are able to perform with an overall medical examination accuracy of 0.61 (CI 0.58-0.64) and a USMLE accuracy of 0.51 (CI 0.46-0.56), while ChatGPT can perform with an overall medical examination accuracy of 0.64 (CI 0.6-0.67) (17). In a study by Durmaz et al. on retinopathy of prematurity, ChatGPT achieved a success rate of 98% (18). In contrast, Lee et al. reported significant differences among three LLMs in bariatric and metabolic surgery: ChatGPT-4 (85.7%), Bard (74.3%), and Copilot (25.7%) (19). In another study on LLMs assisting in glaucoma surgery cases, Carla et al. found ChatGPT to have a 58% success rate, compared to Google Gemini's 32% (p<0.001) (20). Azizoglu et al. compared the accuracy of ChatGPT-4 and ChatGPT-3.5 in the field of pediatric surgery, highlighting notable differences in performance (21). Ulus et al. reported a higher success rate for GPT-4 (75%) compared to GPT-3.5 (45%), though this difference approached

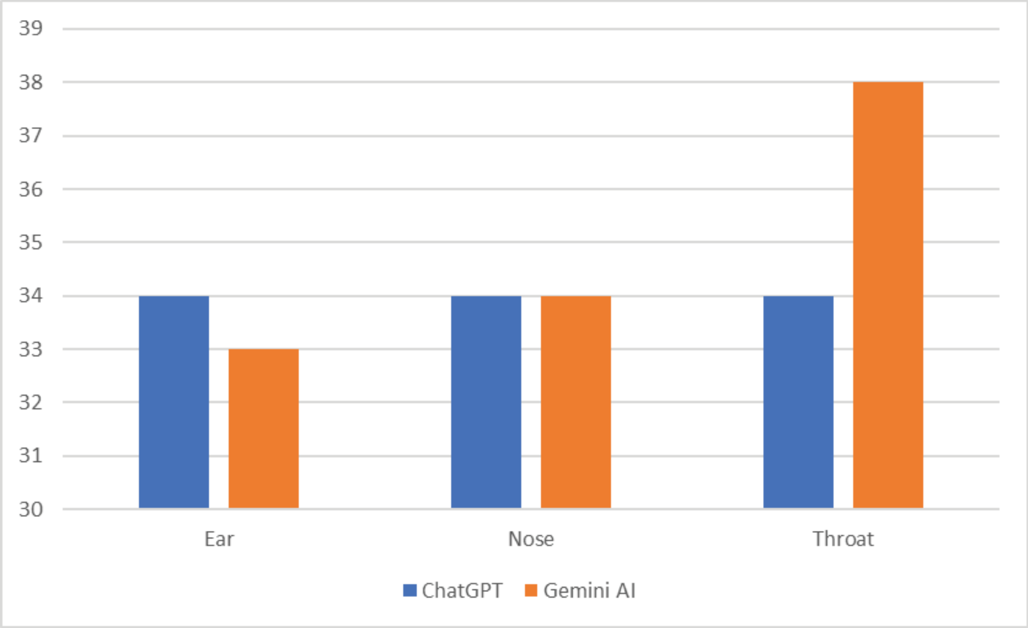


Figure 2: Correct number of questions by topics and AI model

statistical significance (p=0.053) (22). Similarly, Demir et al. demonstrated that ChatGPT-4.0 provided more detailed and accurate responses to patient inquiries about keratoconus than Google Gemini and Microsoft Copilot, with 92% of its answers rated as "agree" or "strongly agree." Numerous studies in this area indicate that various AI models (e.g., ChatGPT, Copilot, Bard, and Gemini) consistently produce meaningful results with high levels of accuracy (23). Teixeira-Marques et al concluded that while ChatGPT shows promise as a clinical decision-making tool in otorhinolaryngology, its performance lags behind specialists, requiring further development to enhance reliability, temporal stability, and accuracy (24).

All these studies demonstrate the potential applicability of artificial intelligence LLM models in the medical field. In otorhinolaryngology, several published articles consistently highlight the need for further development to establish ChatGPT as a reliable tool for clinical decision support, medical education, and patient information, even though AI models have achieved accuracy rates exceeding 70% (25-27). For instance, our study found ChatGPT to have an accuracy rate of 68% and Gemini AI 70%. While these accuracy levels are promising, they also underline certain limitations, indicating that widespread adoption in active clinical use is still premature. Nevertheless, some studies have concluded that artificial intelligence will play a pivotal role in clinical decision-making in the near future, with ChatGPT emerging as the most promising chatbot to date (24).

To the best of our knowledge, this is the first study comparing Gemini AI and ChatGPT 4.0 in the field of otorhinolaryngology. The results of our study align with existing literature, demonstrating consistency despite some variations. Both AI models achieved accuracy rates around 70% across all three subgroups (ear, nose, and throat). Notably, Gemini AI exhibited the best performance in the throat category.

This study has several limitations. First, the dataset included only 150 multiple-choice questions, which may not comprehensively represent all otorhinolaryngology topics. Second, the AI models were tested in a controlled environment, which may differ from real-world clinical scenarios. Third, while both models demonstrated high accuracy (ChatGPT: 68%, Gemini AI: 70%), limitations such as the lack of temporal stability and occasional hallucinations were evident. ChatGPT, for example, provided highly incorrect responses in some instances. Lastly, the absence of real-time medical updates in both models, limited to knowledge up to September 2021, hinders their reliability for current medical decision-making.

Conclusions

ChatGPT 4.0 and Gemini AI demonstrated comparable accuracy in answering otorhinolaryngology questions across ear, nose, and throat topics. With accuracy rates of 68% and 70%, respectively, both models show promise as clinical decision-support tools. However, their limitations, including occasional hallucinations and outdated medical knowledge, highlight the need for further development.

Conflict of interest: The authors report no conflict of interest.

Funding source: No funding was required.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent: This article does not contain any studies with human participants performed by any of the authors.

Acknowledgments: None

Peer-review: Externally. Evaluated by independent reviewers working in at least two different institutions appointed by the field editor.

Data availability: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Contributions

Research concept and design: AC

Data analysis and interpretation: AC

Collection and/or assembly of data: AC

Writing the article: AC

Critical revision of the article: AC

Final approval of the article: AC

All authors read and approved the final version of the manuscript..

References

1. Xie Q, Chen Q, Chen A, Peng C, Hu Y, Lin F, et al. Medical Foundation Large Language Models for Comprehensive Text Analysis and Beyond. Res Sq [Preprint]. 2024:rs.3.rs-5456223.
2. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.
3. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res. 2023;25:e48568.
4. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus. 2023;15(2):e35179.
5. Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK. ChatGPT in Dentistry: A Comprehensive Review. Cureus. 2023;15(4):e38317.
6. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature. 2023;614(7947):224-6.
7. Mokmin NAM, Ibrahim NA. The evaluation of chatbot as a tool for health literacy education among undergraduate students. Educ Inf Technol (Dordr). 2021;26(5):6033-49.
8. Kitamura FC. ChatGPT Is Shaping the Future of Medical Writing But Still Requires Human Judgment. Radiology. 2023;307(2):e230171.
9. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review. J Med Internet Res. 2020;22(10):e20346.
10. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. doi: 10.48550/arXiv.2005.14165.
11. Bhattacharya K, Bhattacharya A, Bhattacharya N, Yagnik Vd, Garg P, Kumar S. ChatGPT in surgical practice—a new kid on the block. Indian J Surg. 2023;22:1–4.
12. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol. 2023;228(6):696-705.
13. Ray PP. Broadening the horizon: a call for extensive exploration of ChatGPT's potential in obstetrics and gynecology. Am J Obstet Gynecol. 2023;229(6):706.
14. Ray PP. Bridging the gap: integrating ChatGPT into obstetrics and gynecology research-a call to action. Arch Gynecol Obstet. 2024;309(3):1111-3.
15. Guerra GA, Hofmann H, Sobhani S, Hofmann G, Gomez D, Soroudi D, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg. 2023;179:e160–e165.
16. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung FH. Assessment of Resident and AI Chatbot Performance on the University of Toronto Family Medicine Residency Progress Test: Comparative Study. JMIR Med Educ. 2023;9:e50514.
17. Waldock WJ, Zhang J, Guni A, Nabeel A, Darzi A, Ashrafian H. The Accuracy and Capability of Artificial Intelligence Solutions in Health Care Examinations and Certificates: Systematic Review and Meta-Analysis. J Med Internet Res. 2024;26:e56532.
18. Durmaz Engin C, Karatas E, Ozturk T. Exploring the Role of ChatGPT-4, BingAI, and Gemini as Virtual Consultants to Educate Families about Retinopathy of Prematurity. Children (Basel). 2024;11(6):750.
19. Lee Y, Shin T, Tessier L, Javidan A, Jung J, Hong D, et al. Harnessing artificial intelligence in bariatric surgery: comparative analysis of ChatGPT-4, Bing, and Bard in generating clinician-level bariatric surgery recommendations. Surg Obes Relat Dis. 2024;20(7):603-8.
20. Carlà MM, Gambini G, Baldascino A, Boselli F, Giannuzzi F, Margollicci F, et al. Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. Graefes Arch Clin Exp Ophthalmol.

2024;262(9):2945-59.

21. Azizoglu M, Aydogdu B. How does ChatGPT perform on the European Board of Pediatric Surgery examination? A randomized comparative study. Acad J Health Sci. 2024;39(1):23-6.

22. Ulus SA. How does ChatGPT perform on the European Board of Orthopedics and Traumatology examination? A comparative study. Acad J Health Sci. 2023;38(6):43-6.

23. Demir S. Evaluation of Responses to Questions About Keratoconus Using ChatGPT-4.0, Google Gemini and Microsoft Copilot: A Comparative Study of Large Language Models on Keratoconus. Eye Contact Lens. 2024 Dec 4. doi: 10.1097/ICL.0000000000001158. Epub ahead of print.

24. Teixeira-Marques F, Medeiros N, Nazaré F, Alves S, Lima N, Ribeiro L, et al. Exploring the role of ChatGPT in clinical decision-making in otorhinolaryngology: a ChatGPT designed study. Eur Arch Otorhinolaryngol. 2024;281(4):2023-30.

25. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol. 2023;280(9):4271-8.

26. Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA, et al. Accuracy of ChatGPT-Generated Information on Head and Neck and Oromaxillofacial Surgery: A Multicenter Collaborative Analysis. Otolaryngol Head Neck Surg. 2024;170(6):1492-503.

27. Qu RW, Qureshi U, Petersen G, Lee SC. Diagnostic and Management Applications of ChatGPT in Structured Otolaryngology Clinical Scenarios. OTO Open. 202322;7(3):e67.

Publisher's Note: Unico's Medicine remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.