Van Med J 32 (3):113 -117, 2025
DOI: 10.5505/vmj.2025.96268

Comparative Evaluation of AI-based Systems for Tinnitus

Abdulaziz Yalınkılıç[1], Mehmet Zeki Erdem[1]

[1]Van Yüzüncü Yıl University, Faculty of Medicine, Department of Otorhinolaryngology, Van, Türkiye

Abstract

Introduction: Today, with the development of technology, the variety of information sources has increased. It is now possible to access information obtained from encyclopedias in seconds with a few clicks of a button. Rapid developments in artificial intelligence (AI) and the widespread use of large language models (LLMs) such as ChatGPT, Gemini, and Perplexity have revolutionized access to medical information. However, the accuracy and readability of the answers provided by these models are critical, especially in the healthcare domain. This study evaluates the performance of ChatGPT, Gemini, and Perplexity in addressing frequently asked questions about tinnitus, a common symptom in otolaryngology practice.

Materials and Methods: Twenty frequently asked questions about tinnitus were posed to the models and their responses were evaluated by two otolaryngologists using global quality (GQS) and Likert scales for accuracy and reliability and the Gunning-Fog Index (GFI) for readability.

Results: The findings reveal no significant difference in the reliability and quality of information between the models, but it was observed that Gemini came out ahead in readability and ChatGPT in accuracy. However, Perplexity lagged in both metrics. These results highlight the varying strengths and weaknesses of LLMs, emphasizing the importance of model selection based on user needs. For example, ChatGPT is ideal for complex medical information, while Gemini is more accessible to wider audiences.

Conclusion: This study demonstrates the potential of AI-enabled systems in healthcare; however, we suggest that future improvements should increase both accuracy and accessibility.

Key words: Large language models; tinnitus; chatbots; chatGPT.

Introduction

Technology has made accessing information much easier. With the spread of the Internet and the increase in digital platforms, the process of accessing information can now be realized within seconds. Encyclopedias and printed books, which are traditional sources of information, have been replaced by online databases, social media platforms, and artificial intelligence-supported information systems. Especially in the field of health, individuals frequently use the internet to learn about diseases, treatment methods, and symptoms. However, this situation brings with it an important problem: the accuracy and reliability of online information sources. Much of the information on the Internet is unregulated and this increases the risk of misinformation, especially in a critical area such as health (1,2).

In recent years, AI-supported chatbots and LLMs have started to play an important role in this information retrieval process. Models such as ChatGPT, Gemini, and Perplexity attract attention by providing fast and comprehensive answers to users' questions. These models not only provide general information but have also been the center of attention in academic studies with their success in medical exams and responses to clinical scenarios. For example, ChatGPT's performance in the US Medical Licensing Examination (USMLE) demonstrated the potential of AI in the field of medical education and counseling. This has increased the confidence in and frequency of use of such technologies both among the public and healthcare professionals (3,4).Tinnitus is a condition seen in 10% to 15% of adults worldwide and is difficult to diagnose and treat (5,6). Tinnitus is a symptom that negatively affects the quality of life of individuals and is associated with sleep disorders, concentration problems, and psychological disorders. However, the pathogenesis of tinnitus is not fully understood and this constitutes a significant difficulty for both patients and physicians. Clinical diagnosis is usually based on the personal knowledge and experience of physicians, which makes it difficult to obtain objective diagnostic accuracy (7). The complex nature of tinnitus and uncertainties in the treatment process direct patients to alternative sources of information. In this context, AI-based information systems stand out as an important tool in the evaluation of complex symptoms such

*Corresponding Author: Abdulaziz Yalınkılıç Department of Otorhinolaryngology, Faculty of Medicine, Van Yuzuncu Yil University Van, Turkey **Email:** y_aziz21@hotmail.com **Orcid:** Abdulaziz Yalınkılıç 0000-0003-2702-5905, Mehmet Zeki Erdem 0000-0003-3263-4633

Received: 29.01.2025, Accepted: 5.05.2025

as tinnitus. AI-supported big language models are increasingly used in medical information provision and evaluation processes. Models such as ChatGPT, Gemini, and Perplexity have attracted wide attention in the literature due to their capabilities in medical subjects (8). These models not only provide information but also facilitate access to information by providing personalized answers to users' questions. However, the accuracy and readability of the answers provided by these models is an issue that needs to be carefully evaluated, especially in the field of health. The accuracy of answers guarantees that users receive reliable information, while readability makes this information comprehensible to a broad audience. In this study, the accuracy, readability, reliability, and quality of the answers given by ChatGPT, Gemini, and Perplexity to frequently asked questions about tinnitus were evaluated. The aim was to analyze the performance of these models in the field of tinnitus and to identify potential areas of use in the field of medical information provision. In addition, examining the accessibility and comprehensibility of these models for different user groups will shed light on future development processes.

Table 1: Questions are used to test large language models

1. 1- What is tinnitus?
2. 2- How often is tinnitus seen?
3. 3- What causes tinnitus?

4- Does tinnitus heal?

5- What does tinnitus sound like?

4. 6- What are the types of tinnitus?
5. 7- When does tinnitus become dangerous?

8- Which vitamin deficiency causes tinnitus in the ears?

9- In which diseases is tinnitus in the ears seen?

10- Does ringing in the ears last a lifetime?

11- What should a patient with tinnitus pay attention to?

12- What are the treatment methods for tinnitus?

13- What are the factors that worsen tinnitus?

14- Can tinnitus cause vertigo?

15- Can stress cause tinnitus in the ears?

16- Can aspirin cause tinnitus in the ears?

17- Can high blood pressure cause tinnitus in the ears?

18- Can earwax cause tinnitus in the ears?

19- Do hearing aids help with tinnitus?

20- Can tinnitus cause headaches?

Materials and Methods

Twenty of the most frequently asked questions related to tinnitus on internet search engines were selected (Table 1). These questions were asked to ChatGPT, Gemini, and Perplexity and the answers were recorded. Free versions accessible to everyone were preferred. To obtain consistent results, the first answer of each LLM to each question was accepted. Questions were entered on the same day by accessing the LLM using a single account. The answers given by two otolaryngologists with at least 10 years of experience were compared with the current literature and the quality and reliability of the information were evaluated by giving a score between 1-5 according to the global quality scale (GQS) and the accuracy was evaluated by giving a score between 1-7 according to the Likert scale. (Table 2).

Table 2: Comparison of statistical test results for ChatGPT, Gemini, and perplexity

	Group	Mean	Std. Dev.	Median	Range	*p.*
	ChatGPT	4.95	.22	5.00	1.00	
GQS	Gemini	4.90	.31	5.00	1.00	.159
	Perplexity	4.75	.44	5.00	1.00	
	ChatGPT	16.06 a	2.08	15.96	8.59	
GFI	Gemini	13.78 b	1.83	13.85	7.10	.001
	Perplexity	17.34 a	5.17	16.73	24.63	
	ChatGPT	4.90 a	.31	5.00	1.00	
Likert	Gemini	4.75 a	.44	5.00	1.00	.007
	Perplexity	4.45 b	.51	4.00	1.00	

***** Significance level between groups according to Kruskal-Wallis H test;
a,b: Shows the difference between groups according to the Bonferroni Post-Hoc test

The evaluation was determined by a joint decision. Readability was assessed with the Gunning-Fog Index (GFI), which is frequently used in the existing literature and may indicate the level of education required to understand the text at first reading (Table 3). These readability scores were calculated by automatically transferring them to the http://gunning-fog-index.com/ website. This provided an objective assessment of the accessibility of the texts for the average reader.

Ethical approval: This study does not require ethics committee approval because it involves publicly available data and does not involve human participants. The study was conducted by the guidelines and regulations on ensuring the confidentiality and integrity of data throughout the research process.

Table 3: Comparison of gunning fog index (GFI) statistical test results for chatGPT, gemini, and perplexity

		Group						*p.
		ChatGPT		Gemini		Perplexity		
		N	%	N	%	N	%	
GFI	College freshman	3	50.0%	3	50.0%	0	0.0%	.011
	College graduate	3	50.0%	1	16.7%	2	33.3%	
	College junior	6	50.0%	5	41.7%	1	8.3%	
	College senior	5	55.6%	0	0.0%	4	44.4%	
	College sophomore	0	0.0%	4	66.7%	2	33.3%	
	High school junior	0	0.0%	1	50.0%	1	50.0%	
	High school senior	1	12.5%	4	50.0%	3	37.5%	
	High school sophomore	0	0.0%	2	100.0%	0	0.0%	
	Postgraduate	2	22.2%	0	0.0%	7	77.8%	

***** Significance level according to the results of chi-square (Fisher's exact) test

Statistical analysis: In calculating the sample size of the study, the Power (Power of the Test) for each variable was determined by taking at least 80% and a Type-1 error of 5%. Shapiro-Wilk and Skewness-Kurtosis tests were used to determine whether the continuous measurements in the study were normally distributed and nonparametric tests were applied since the measurements were not normally distributed. Descriptive statistics for continuous variables in the study were expressed as mean, standard deviation (SD), median, range, number (n), and percentage (%). "Kruskal-Wallis H test" was used to compare continuous measurements according to groups. Following the Kruskal-Wallis test, "Post-Hoc Test with Bonferroni correction" was used to determine the different groups. The Chi-square (Fisher's exact) test was calculated to determine the relationships between categorical variables and groups. Statistical significance level was taken as $p<0.05$ in the calculations and SPSS (IBM SPSS for Windows, ver.26) statistical package program was used for the analyses.

Results

In this study, the performances of different artificial intelligence programs (ChatGPT, Gemini, Perplexity) were compared in terms of GKS (quality and reliability), GFI (readability), and Likert (accuracy) criteria. There was no statistically significant difference between the artificial intelligence programs in terms of GKS scores ($p=0.159$) and the median value of all groups was 5.00. However, a significant difference was observed between the groups in the GFI (readability) measurement ($p=0.001$), and Gemini's GFI score was significantly lower than the other AI programs. Similarly, a significant difference was found between the groups in Likert accuracy measurement ($p=0.007$), and Perplexity's accuracy score was significantly lower than the other programs. A statistically significant relationship was found between GFI level and artificial intelligence programs ($p=0.011$), and it was determined that the distribution of GFI level was affected by artificial intelligence groups. For example, ChatGPT showed a higher distribution at 'College freshman' and 'College junior' levels, while Gemini showed a higher distribution at 'College sophomore' and 'High school sophomore' levels. Perplexity had the highest distribution at the 'Postgraduate' level (Figure 1).

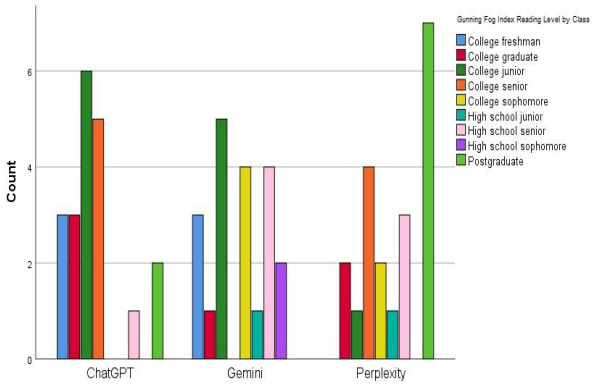

Figure 1: Gunning Fog Index (GFI) reading level by grade

Discussion

The potential of AI-assisted LLMs to provide medical information has attracted considerable interest both in academic circles and in the general public in recent years. The ability of these models to provide information and answer users' questions, especially on complex medical topics, has been addressed in many studies (9,10). The fact that LLMs facilitate access to medical information enables users to access information quickly and practically. However, the accuracy and readability of the answers given by these models is an important issue that should be carefully

evaluated, especially in the field of health (11). In this study, the accuracy and readability of the answers given by three popular LLMs such as ChatGPT, Gemini, and Perplexity to frequently asked questions about tinnitus were evaluated. The findings show that all three models perform satisfactorily in general. However, significant differences were observed between the models.

ChatGPT outperformed the other models in terms of the accuracy of the responses. In the study, it was found that ChatGPT gave correct answers to almost all questions and therefore received the highest accuracy score. This result is in line with other studies in the literature. For example, the success of ChatGPT in medical exams and clinical scenarios supports the reliability of this model in providing medical information (3,4,12). Gemini closely followed ChatGPT in terms of accuracy, but Perplexity received a lower accuracy score compared to the other two models. This reveals the limitations of Perplexity in providing medical information. However, the overall accuracy level of all three models was satisfactory, indicating the potential of these AI-supported tools to provide medical information. In the literature, there are many studies in which ChatGPT shows superior performance in terms of accuracy. For example, in a study by Johnson et al., the accuracy rate of ChatGPT's answers to medical questions was found to be over 90% (13). Similarly, in a study by Mediboina et al. (12), it was reported that ChatGPT gave more accurate answers compared to other models, especially in complex medical issues. However, some studies also reported that Gemini (formerly known as Bard) gave more accurate answers than ChatGPT in certain subjects (14). These differences can be explained by the methodologies used in the studies, the complexity of the topics evaluated, and the characteristics of the user groups. A review of 64 articles on ChatGPT and Gemini in various specialties found that ChatGPT gave more accurate answers. However, when the character length of the answers given was examined, it was found that the Gemini gave longer answers (15). This suggests that the Gemini may be better for more detailed explanations. Our study shows that the Gemini provides more readable answers. In terms of readability, Gemini performed better than the other models. Analyses using the GFI showed that Gemini's responses had a lower GFI score (mean: 13.78) and were therefore easily understandable by a wider range of users. ChatGPT (mean: 16.06) and Perplexity (mean: 17.34) had higher GFI scores and used a more complex language structure. This may limit the accessibility of Perplexity's responses in particular. In the literature, it is stated that lower GFI scores are more suitable for large audiences (16). For example, while Gemini's answers are suitable for a reading level at the first-year university level, ChatGPT's answers require a reading skill at the senior university level, and Perplexity's answers require a reading skill at the university graduate level. These differences emphasize the importance of model selection according to the educational level of the users. In the study examining the readability level of ChatGPT's health information, it was suggested that it should be adapted especially for users with low reading levels. However, in this study, the responses were made more readable by giving special commands to ChatGPT. In our study, the models were evaluated within the framework of standard user experience without any special commands. This shows that ChatGPT uses a more complex language structure with default settings and therefore readability scores are higher. Although the findings of this study are generally compatible with other studies in the literature, some differences are noteworthy. For example, some studies reported that Gemini gave more accurate answers than ChatGPT, but ChatGPT was more readable (16). These differences can be explained by the methodologies used in the studies, the complexity of the subjects evaluated, and the characteristics of the user groups. For example, in some studies, it was reported that the commands given to artificial intelligence models affected the readability and accuracy of the responses. Moreover, the education levels and information access habits of different user groups play an important role in evaluating the performance of the models.The findings of this study make an important contribution to understanding the potential and limitations of AI-supported large language models in providing medical information through the assessment of a symptom such as tinnitus. ChatGPT showed superior performance in terms of accuracy, while Gemini showed better results in terms of readability. Perplexity lagged behind the other two models in both criteria. This shows the importance of selecting models according to the needs of different user groups. For example, ChatGPT may be a more suitable option for users who need more in-depth medical information and have higher reading skills, while Gemini offers a more accessible alternative for users who want to address a wider audience.

Study limitatons: This study has some limitations. A common concern with artificial intelligence is the risk of fabrication. Not

everyone can accurately assess the accuracy of AI responses, which can lead to users trusting them too much. Therefore, medically, LLMs can be dangerous.

Conclusion

The performance of AI-supported large language models in providing medical information may vary according to the needs and expectations of users. The field of AI is undergoing rapid advancements, with frequent updates and new developments. As algorithms are increasingly adopted in healthcare settings, these updates must be made with the specific needs and requirements of the healthcare sector in mind. In the future, studies are needed to improve the readability of these models and make them accessible to a wider audience.

Ethical approval: This study does not require ethics committee approval because it involves publicly available data and does not involve human participants.

Conflict of interest: The authors have no conflict of interest regarding this study.

Financial disclosure: No financial support has been received for this study.

Author contributions: Concept (AY), Design (AY, MZE), Data Collection and Processing (AY, MZE), Analysis and Interpretation (AY, MZE)

References

1. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology 2023;307(2):e230163.
2. Mu X, Lim B, Seth I, Xie Y, Cevik J, Sofiadellis F, et al. Comparison of large language models in management advice for melanoma: Google's AI BARD, BingAI and ChatGPT. Skin Health Dis 2023;4(1):e313.
3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198.
4. Musheyev D, Pan A, Loeb S, Kabarriti AE. How Well Do Artificial Intelligence Chatbots Respond to the Top Search Queries About Urological Malignancies?. Eur Urol 2024;85(1):13-16.
5. Baguley D, McFerran D, Hall D. Tinnitus. Lancet 2013;382(9904):1600-1607.
6. Michiels S. Somatosensory Tinnitus: Recent Developments in Diagnosis and Treatment. J Assoc Res Otolaryngol 2023;24(5):465-472.
7. Piccirillo JF, Rodebaugh TL, Lenze EJ. Pharmacological Treatments for Tinnitus-Reply. JAMA 2020;324(11):1109-1110.
8. Yin Z, Kuang Z, Zhang H, Guo Y, Li T, Wu Z, et al. Explainable AI Method for Tinnitus Diagnosis via Neighbor-Augmented Knowledge Graph and Traditional Chinese Medicine: Development and Validation Study. JMIR Med Inform 2024;12:e57678.
9. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst 2023;47(1):33.
10. Sahin S, Erkmen B, Duymaz YK, Bayram F, Tekin AM, Topsakal V. Evaluating ChatGPT-4's performance as a digital health advisor for otosclerosis surgery. Front Surg 2024;11:1373843.
11. Aygul Y, Olucoglu M, Alpkocak A. Tipta uzmanlik sinavinda (tus) buyuk dil modelleri insanlardan daha mi basarili? 2024; arXiv preprint arXiv:2408.12305.
12. Mediboina A, Badam RK, Chodavarapu S. Assessing the Accuracy of Information on Medication Abortion: A Comparative Analysis of ChatGPT and Google Bard AI. Cureus 2024;16(1):e51544.
13. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Preprint Res Sq 2023;rs.3.rs-2566942.
14. Seth I, Lim B, Xie Y, Cevik J, Rozen WM, Ross RJ, et al. Comparing the Efficacy of Large Language Models ChatGPT, BARD, and Bing AI in Providing Information on Rhinoplasty: An Observational Study. Aesthet Surg J Open Forum 2023;5:ojad084.
15. Fattah FH, Salih AM, Salih AM, Asaad SK, Ghafour AK, Bapir R, et al. Comparative analysis of ChatGPT and Gemini (Bard) in medical inquiry: a scoping review. Front Digit Health 2025;7:1482712.
16. Ayoub NF, Lee YJ, Grimm D, Balakrishnan K. Comparison Between ChatGPT and Google Search as Sources of Postoperative Patient Instructions. JAMA Otolaryngol Head Neck Surg 2023;149(6):556-558.