Acta Medica Alanya

RESEARCH ARTICLE

ARAŞTIRMA MAKALESİ

Acta Medica Alanya

2025;9(3): 228-235

DOI: 10.30565/medalanya.1807012

How Well Do AI Models Inform Otitis Media Patients? A Comparative Study of Large Language Models

Yapay Zeka Modelleri Otitis Media Hastalarını Ne Kadar İyi Bilgilendiriyor? Büyük Dil Modellerinin Karşılaştırmalı Çalışması

Yeşim Yüksel[1]*, Rezarta Taga Senirli[1]

1.Department of Otorhinolaryngology, University of Health Sciences, Antalya Training and Research Hospital, Antalya, Türkiye

ABSTRACT

Aim: The aim of this study was to evaluate the accuracy, comprehensiveness, and readability of information provided by large language models (LLMs) supported by natural language processing (NLP) technologies regarding otitis media (OM) based on responses to questions asked by patients and their relatives.

Method: In this descriptive, cross-sectional evaluation study, 60 frequently asked questions by patients and their relatives regarding OM were classified under four subheadings (general information, diagnosis, follow-up and therapy, and surgery and complications) and answered by three different LLMs (Google Gemini 2.5 Flash, Microsoft Copilot, ChatGPT-4o). The answers were evaluated for accuracy by two experienced otorhinolaryngology specialists using a 5-point Likert scale. The readability of the responses was analyzed using the Coleman-Liau Index (CLI) and Simple Measure of Gobbledygook (SMOG) index to determine readability levels corresponding to academic education levels and to compare the models.

Results: The artificial intelligence (AI) models received similarly high scores for accuracy in their responses to patient questions related to OM. In the readability analysis, Gemini responses were found to be statistically significantly more readable than those of the other models, according to the SMOG and CLI indices. The ChatGPT responses required a higher level of education; in particular, the readability of the answers under the "diagnosis" subheading was found to have the highest rate of graduate-level education requirement.

Conclusion: Although the three commonly used AI models provided similarly accurate responses to OM-related questions, differences in readability were observed among the LLMs. For AI to effectively support patient education and promote treatment adherence, both the accuracy and readability of the content are essential.

Keywords: otorhinolaryngologic disease, otitis media, artificial intelligence, large language models.

ÖZ

Amaç: Bu çalışmada, doğal dil işleme (NLP) teknolojileri ile desteklenen büyük dil modellerinin (LLM) hastalar ve hasta yakınları tarafından yöneltilen sorulara verdikleri yanıtlar üzerinden, otitis media (OM) hakkında sundukları bilgilerin doğruluğu, kapsamlılığı ve okunabilirliği değerlendirildi.

Yöntem: Bu tanımlayıcı kesitsel değerlendirme çalışmasında, hastalar ve hasta yakınları tarafından OM ile ilgili sıkça sorulan 60 soru dört alt başlık (genel bilgi, tanı, takip ve tedavi ve cerrahi ve komplikasyonlar) altında gruplandırıldı ve üç farklı büyük dil modeli (Google Gemini 2.5 Flash, Microsoft Copilot, ChatGPT-4o) tarafından yanıtlandı. Yanıtlar, iki deneyimli Kulak Burun Boğaz (KBB) uzmanı tarafından 5 puanlık Likert ölçeği kullanılarak doğruluk açısından değerlendirildi. Yanıtların okunabilirliği Coleman-Liau İndeksi (CLI) ve Gobbledygook'un Basit Ölçüm (SMOG) indeksi kullanılarak analiz edildi ve bu analiz ile okunabilirlik düzeylerine karşılık gelen akademik eğitim düzeyleri belirlendi ve modeller karşılaştırıldı.

Bulgular: Yapay zeka modellerinin OM ile ilgili hasta sorularına verdiği yanıtlar, doğruluk açısından benzer şekilde yüksek puanlar aldı. Okunabilirlik analizinde, SMOG ve CLI indekslerine göre Gemini'nin yanıtlarının diğer modellere kıyasla istatistiksel olarak anlamlı düzeyde daha okunabilir olduğu bulundu. ChatGPT'nin yanıtları daha yüksek bir eğitim seviyesi gerektirmekle birlikte, özellikle "tanı" başlığı altındaki yanıtların okunabilirliği için lisansüstü eğitim gerekliliği en yüksek oranda bulundu.

Sonuç: Yaygın olarak kullanılan üç yapay zeka modeli, OM ile ilgili sorulara benzer düzeyde doğru yanıtlar vermiş olsa da, LLM'ler arasında okunabilirlik açısından farklılıklar gözlendi. Yapay zekânın hasta eğitimini etkin bir şekilde desteklemesi ve tedaviye uyumu artırması için hem içeriğin doğruluğu hem de okunabilirliği önemlidir.

Anahtar Kelimeler: otorinolarengolojik hastalık, otitis media, yapay zeka, büyük dil modelleri.

Recieved Date: 20.10.2025 / Accepted Date: 06.12.2025 / Published (Online) Date:31.12.2025

*Corresponding Author: Yeşim Yüksel. Department of Otorhinolaryngology, University of Health Sciences, Antalya Training and Research Hospital. Antalya, Türkiye, Phone: +90 532 303 76 19 mail: yesimgedikli@gmail.com

ORCID: 0000-0003-2280-3843

To cited: Yuksel Y, Senirli RT. How Well Do AI Models Inform Otitis Media Patients? A Comparative Study of Large Language Models. Acta Med. Alanya 2025;9(3)::228-235 DOI: 10.30565/medalanya.1807012

Introduction

Otitis media (OM) is among the most common reasons for presentation to ear, nose, and throat (ENT) clinics. Although it occurs most frequently in childhood, it can be seen across all age groups [1-3]. There are different subtypes, including acute OM (AOM), OM with effusion (OME), and chronic suppurative OM (CSOM). Although less common in adults, the possibility of an underlying disease in adult OM patients should not be overlooked. Serious diseases which may accompany OM, such as immunodeficiencies and malignancies of the head and neck region that predispose to infection, must be carefully evaluated.

Acute OM has a rapid onset and is typically characterized by pronounced symptoms such as otalgia and high fever, whereas OME usually presents with subtler symptoms that may progress gradually [1,2]. All forms of OM carry the risk of serious complications, including hearing loss, tympanic membrane perforation, and, in severe cases, intracranial sequelae such as meningitis and brain abscess. The diagnosis is primarily made through clinical evaluation and otoscopic examination; in some cases, it may be supported by audiometry and tympanometry [1-3]. Management of AOM involves antibiotic use in addition to symptomatic treatment in severe cases, while treatment of OME usually requires supportive therapies such as antihistamines, decongestants, and nasal steroids in the presence of a concomitant upper respiratory tract infection. Antibiotic therapy may be initiated when a bacterial infection is suspected or confirmed. During follow-up, ventilation tubes may be inserted, when necessary, to drain fluid accumulated in the middle ear, equalize ear pressure, and provide ventilation [1,2]. Consequently, OM is regarded by clinicians as a challenging condition to manage and represents one of the otorhinolaryngological disorders about which patients and their families most frequently seek information. This is attributable to its high prevalence, broad range of etiological factors, prolonged treatment and follow-up requirements, impact on daily functioning, and potential for serious complications.

In recent years, artificial intelligence (AI) technologies in healthcare have advanced substantially, particularly in remote disease diagnosis, radiological interpretation, clinical decision-making, and diagnosis, prognosis, and treatment planning. These developments have contributed to the increasingly widespread adoption of AI tools by clinicians [3-7]. Using large datasets collected over time on OM patients, AI systems can support clinicians by facilitating differential diagnosis, predicting disease risk, personalizing treatment strategies, and optimizing overall disease management [3,4]. In the study by Dink et al. [3], the contributions of AI in OM management by clinicians and the effects of integrating AI into OM care on patients, physicians, surgeons, and the entire multidisciplinary care team were discussed. Natural language processing (NLP) is a specialized branch of AI that focuses on understanding, interpreting, and generating human language in an intelligent and meaningful way. Large language models (LLMs) are AI systems developed to understand, produce, and interpret human language. LLMs have recently been used to power NLP applications, leading to significant improvements. Different LLMs optimize NLP for different tasks. The release of Chat Generated Pre-Trained Transformer (ChatGPT) (OpenAI, November 2022) marked a significant NLP milestone, achieving the fastest growth in consumer app history by reaching 100 million users within two months [5]. It is still only emerging in healthcare and promises a wide range of future applications. Interest in application NLP to otolaryngology has surged since the introduction of LLMs [8]. The NLP technologies are becoming increasingly valuable for clinicians, including otorhinolaryngology specialists, by improving workflow efficiency and supporting the comprehensive range of their clinical responsibilities. The literature contains numerous studies published in recent years on various topics related to AI in otorhinolaryngology [3-8]. Some of these studies have addressed the timely and accurate completion of patient applications, ensuring the patient's active participation in the treatment and follow-up process, and increasing patient awareness regarding the development of complications through the use of AI models. In this context, the necessity of using AI in developing and improving patient education has

been evaluated [5,9-11].

For OM and similar otorhinolaryngological conditions where patients and their families often have a substantial need for reliable information, it is essential to evaluate the accuracy and adequacy of content generated by AI models that offer enhanced accessibility and convenience. Additionally, assessing whether different AI models provide variable responses to identical questions and examining the readability of these responses are important considerations. In the present study, we, therefore, aimed to evaluate and compare the accuracy, comprehensiveness, and readability of content generated by different AI models which are frequently used, easily accessible, and open access, in response to questions asked by OM patients and their relatives on general information, diagnosis, follow-up, therapy, surgery and complications.

Methods

This descriptive cross-sectional evaluation study did not require Ethics Committee approval, as all data sources were publicly accessible and contained no identifiable or sensitive information. The authors declare no affiliations with, or involvement in, any of the AI research and development companies referenced in this study.

This study was conducted using a dataset of questions related to “otitis media” asked by patients and their relatives. To increase inclusivity and patient representation, questions were selected from patient support groups, social media posts, medical websites related to ENT diseases, educational health platforms for patients, and online question-and-answer platforms such as Quora. All questions, including those containing technical medical terms that could describe situations patients might encounter during their medical processes, were carefully screened by the participating authors and selected by mutual agreement to enable the AI model to assess its ability to provide information to patient questions. Each question was carefully evaluated for its suitability for inclusion in the study. Questions with similar meanings or similar answers, questions containing ambiguous expressions, and non-medical questions related to the disease were excluded from the evaluation. Finally, a total of 60 questions related to the topic of otitis media were selected and divided into groups of 15 questions each: (1) general information, (2) diagnosis, (3) follow-up and therapy, and (4) surgery and complications.

Google Gemini 2.5 Flash (Google DeepMind, London, UK), Microsoft Copilot (Microsoft-11) (Microsoft, Redmond, WA, USA), and ChatGPT-4o (May 2024, OpenAI, San Francisco, California, USA) LLMs were included in this study. All questions posed to the LLMs were asked in English, with each question entered independently and separately, and the responses were recorded. Two experienced otorhinolaryngologists, actively engaged in academic practice, independently reviewed and evaluated the responses for accuracy, resolving any discrepancies through consensus. They conducted these evaluations using a 5-point Likert scale which is a standard measurement tool frequently used in medical and academic evaluations. The LLM responses were analyzed and rated as follows: 1- Strongly disagree / 2- Disagree / 3- Partially agree and partially disagree / 4- Agree / 5- Strongly agree.

The LLM outputs were evaluated in terms of readability. Before evaluation, elements which could influence the scores, such as author information, legal notices, sources, and publication dates, were removed from the text. To assess the readability of the texts, the Coleman-Liau Index (CLI) and Simple Measure of Gobbledygook (SMOG) Index criteria were used [11,12]. The responses given by each AI model to the questions were scored separately in the readability indices. Along with the scores obtained in both readability criteria, the academic education levels corresponding to the readability level according to the SMOG index score were also recorded. The SMOG index score was evaluated as high school for scores between 9 and 12.9, undergraduate for scores between 13 and 16.9, and graduate for scores of 17 or higher. Higher scores correspond to greater reading difficulty. Higher academic education levels indicate more difficult-to-read texts. For all questions and subcategory groups related to OM, the AI models were compared using their Likert scores, CLI scores, SMOG index scores, and distributions by academic grade level.

Statistical Analysis

Statistical analysis was performed using the SPSS for Windows version 11.5 software (SPSS Inc., Chicago, IL, USA). Continuous data were presented in mean ± standard deviation (SD) or median (min-max), while categorical data were presented in number and frequency. Comparisons between the three groups were performed using one-way analysis of variance (ANOVA) and the Bonferroni test for quantitative variables and the chi-square test for qualitative variables. A p value of <0.05 was considered statistically significant.

Results

The accuracy of the responses provided by AI models, based on a Likert scale, showed that all responses scored 4 or 5 points. The percentage of responses scoring 5 points on the Likert scale was determined to be 53.3% for ChatGPT, 61.7% for Copilot, and 68.3% for Gemini. The mean scores given by the authors for the AI models and question groups, along with their comparison results, are presented in Table 1. Accordingly, no statistically significant difference was observed in the evaluations within the four different question groups, each consisting of 15 questions covering the topics of “general information”, “diagnosis”, “follow-up and therapy” and “surgery and complications”, and in the comparison of the Likert score averages of the three AI models for a total of 60 questions (Table 1).

The descriptive values and comparison results of readability scores calculated using the SMOG index according to AI models and question groups are presented in Table 2. Accordingly, the SMOG index readability scores for the answers to the questions in the general information and follow-up-therapy subheadings, each consisting of 15 questions, did not show any statistically significant difference among the AI models. However, the SMOG index obtained from Gemini for diagnosis and surgical-complication question answers was found to be statistically significantly lower than ChatGPT (p=0.008 and p=0.006, respectively) and Copilot (p=0.043 and p=0.020, respectively). For all questions, statistically significant differences were observed among the ChatGPT, Gemini, and Copilot AI models (p<0.001) (Table 2).

In the assessment of academic education levels corresponding to readability levels based on the SMOG index score, proportional differences were observed both between AI models and within the subheadings of question groups for each AI model. Considering the responses to all questions, the readability education level required according to the SMOG index was most frequently found to be at the undergraduate level for all AI models. The undergraduate level education requirement rates were determined as 51.7% for ChatGPT, 81.7% for Gemini, and 63.3% for Copilot. Comparing the readability education level requirements according to the SMOG index among the three different AI models revealed that the undergraduate education requirement rate for the readability of Gemini responses was significantly higher than that of the ChatGPT (p=0.001) and Copilot (p=0.018) AI models. While 41.7% of ChatGPT responses to all questions required graduate-level education for readability, this rate was lower for Gemini (8.3%) and Copilot (28.3%). Particularly, in ChatGPT question-and-answer section under the subheading “diagnosis”, the readability requirement at the graduate level was found to be the highest at 60%. Unlike other AI models, only Gemini had a higher readability rate at the high school level (10%) than at the graduate level (8.3%) in all question-answer pairs (Table 2).

The descriptive values and comparison results of readability scores calculated using the CLI according to AI models and question groups are presented in Table 3. Accordingly, the comparison of the three AI models revealed that the index scores calculated from the responses obtained in the Gemini AI model were statistically significantly lower than those in ChatGPT in the subheadings of diagnosis, follow-up-therapy, and surgical-complications, as well as in all questions. Similarly, in the Copilot AI model, the mean CLI score was found to be significantly lower than ChatGPT in the general information and diagnosis subheadings and in all questions. In the comparison between Gemini and Copilot, the CLI scores calculated from the answers obtained in the follow-up-therapy and surgical-complications subheadings and all questions were statistically significantly lower than those of the Copilot AI model (Table 3).

Table 1. Likert scale scores given by reviewers for the responses of AI models to questions, along with the comparison results.

Question groups	Likert Scale			LLMs Comparison Results		
	ChatGPT	Gemini	Copilot	ChatGPT-Gemini	ChatGPT-Copilot	Gemini-Copilot
	Mean±SD	Mean±SD	Mean±SD			
General information	4,60±0,51	4,80±0,41	4,60±0,51	0,189	1,000	0,189
Diagnosis	4,47±0,52	4,67±0,49	4,60±0,51	0,082	0,433	0,670
Follow-up and Therapy	4,53±0,52	4,60±0,51	4,53±0,52	0,751	1,000	0,719
Surgical Treatment and Complication	4,53±0,52	4,67±0,49	4,73±0,46	0,499	0,384	0,670
Overall	4,53±0,50	4,68±0,47	4,62±0,49	0,072	0,358	0,398

Table 2. Comparative results based on the SMOG index scores and education level distributions of AI models.

Question groups	SMOG Index						LLMs Comparison Results		
	ChatGPT		Gemini		Copilot		ChatGPT- Gemini	ChatGPT- Copilot	Gemini -Copilot
	Mean±SD		Mean±SD		Mean±SD		p	p	p
General information	15,69±3,13		15,12±2,38		14,17±1,55		0,599	0,110	0,201
Diagnosis	18,36±3,81		14,62±2,38		16,48±2,13		**0,008**	0,092	**0,043**
Follow-up and Therapy	17,00±3,24		15,75±1,93		16,28±1,82		0,063	0,419	0,099
Surgical Treatment and Complication	16,86±2,26		15,24±1,13		16,37±1,78		**0,006**	0,455	**0,020**
Overall	16,97±3,22		15,12±1,83		15,83±2,03		**0,001**	**0,010**	**0,039**
Education Levels	n	%	n	%	n	%	p	p	p
High school	4	6,7	6	10,0	5	8,3	**0,001**	0,310	**0,018**
Undergraduate	31	51,7	49	81,7	38	63,3			
Graduate	25	41,7	5	8,3	17	28,3			

Table 3. Coleman-Liau index scores and comparison results based on AI models and query groups

Question groups	Coleman- Liau Index			LLMs Comparison Results		
	ChatGPT	Gemini	Copilot	ChatGPT-Gemini	ChatGPT-Copilot	Gemini-Copilot
	Mean±SD	Mean±SD	Mean±SD			
General information	14,07±1,96	13,52±1,70	12,74±1,71	0,147	**0,002**	0,099
Diagnosis	15,59±2,89	13,83±1,52	14,47±1,78	**0,009**	**0,033**	0,055
Follow-up and Therapy	15,57±2,90	14,14±1,66	15,75±1,93	**0,024**	0,816	**0,013**
Surgical Treatment and Complication	16,13±1,75	13,66±0,99	16,02±1,58	**0,001**	0,812	**0,001**
Overall	15,34±2,49	13,79±1,48	14,75±2,15	**0,001**	**0,029**	**0,001**

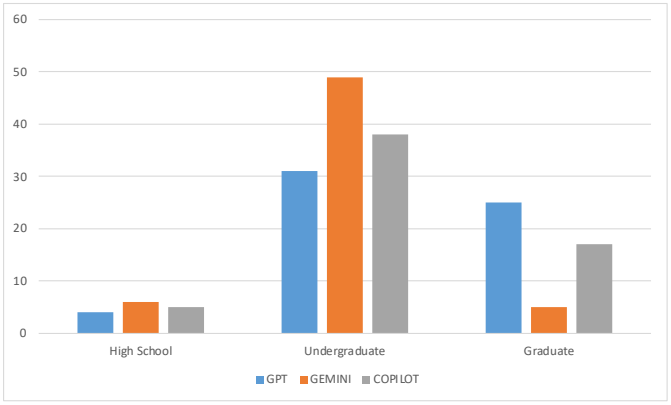

Figure 1. Education level distributions of AI models based on the SMOG readability index

The correlation between SMOG and CLI readability indices was evaluated for the AI-generated answers to all questions. Statistically significant positive correlations were observed between SMOG and CLI scores for all three models: ChatGPT (r = 0.479, p = 0.001), Gemini (r = 0.655, p = 0.001), and Copilot (r = 0.518, p = 0.001).

Discussion

In the fields of otorhinolaryngology and related clinical science, the use of online AI models which facilitate access to information for clinicians, patients, and their relatives and offer open access has increased significantly in recent years [5]. Among the primary LLM-based AI systems developed by various technology companies, ChatGPT (OpenAI), Gemini (Google), and Copilot

(Microsoft) are the most frequently preferred applications. The potential offered by these NLP in terms of accessibility and ease of use is also recognized in various studies in the literature [5,10,13,14]. There are studies evaluating the use of AI models as a source of information to increase patient education and participation in treatment [8,10,13-21]. Patient information content generated by AI may replace professional patient information brochures and patient/caregiver education materials in the coming years due to its similar information quality. However, studies have shown that the readability of patient information content generated by AI is still difficult [14-16, 18-20]. To ensure optimal benefit from AI-generated patient education materials, the content must be medically accurate and presented at an appropriate level of readability and comprehensibility [11,16,18,20]. In our study, we evaluated the content accuracy and comprehensiveness of the responses provided by different AI models to accurately inform patients and their relatives about OM disease, as well as the readability levels of the texts. The main finding of this study was that although all three AI models provided similarly accurate information, their readability levels differed significantly, with Gemini generating the most readable responses.

In the current study, the content of responses provided by ChatGPT (OpenAI), Gemini (Google), and Copilot (Microsoft) AI models to questions about OM posed by patients and their relatives was evaluated using a Likert scale. Similarly, in a study by Bellinger et al. [14], which evaluated three similar LLMs on common rhinology pathologies, no significant difference in accuracy was found between the ChatGPT and Gemini AI models, and the accuracy of both AI models was found to be higher than that of Copilot. When the responses provided by each AI model were evaluated by otorhinolaryngology specialists using the Likert scale, the quality, accuracy, and clarity of the responses were found to be satisfactory and similarly high. Therefore, in our study, the readability index scores became decisive in the comparative evaluation of the responses provided by the AI models to all questions under the subheadings of general information, diagnosis, follow-up-therapy, and surgery-complications related to OM.

The SMOG and CLI readability indices are used in the literature to assess the readability of texts, particularly in cases where comprehensibility is critical, such as healthcare information and educational materials. These formulas estimate the level of education needed to understand a text, an important step in ensuring that the evaluated material is accessible to its intended audience. The SMOG formula is preferred more often owing to its consistency and simplicity in medical topics. The CLI index offers an advantageous assessment option, as it provides fast and automatic analysis in digital texts and contains fewer errors than character counting, compared to syllable counting [11,12,16,20]. In our study, we found a positive correlation between the SMOG and CLI index readability assessments of the responses given by all three AI models to questions related to OM.

In the present study, when evaluating the readability of responses provided by ChatGPT, Gemini, and Copilot AI models to questions related to OM, significant differences were observed among the LLMs. In general, ChatGPT responses demonstrated higher readability scores, indicating greater reading difficulty. Compared to Gemini in particular, this increased difficulty was evident across all question subheadings except for general information. For all AI models, when both readability indices were considered together, responses to questions in the general information category demonstrated lower readability scores than those in the other question groups, with the exception of the diagnosis subheading in Gemini. Accordingly, the most easily readable responses were those provided for questions in the general information subcategory. For both readability indices, the lowest scores, and therefore the highest readability, were observed in Copilot responses within this category. Consistent with the high readability scores observed across all question groups for ChatGPT, 41.7% of its responses corresponded to a graduate-level academic reading requirement. The requirement for a graduate education level to understand the text was the highest in ChatGPT for responses to questions related to OM diagnosis. This can be attributed to the fact that AI models, particularly ChatGPT, have been trained on more technically dense medical corpora such as medical diagnosis, thereby leading to more detailed, but also more

complex and technical responses. Of note, as demonstrated in this study, a model's greater amount of information does not necessarily translate into more readable responses. On the contrary, information density can sometimes result in the excessive use of technical terms rather than simplification. There are studies in the literature reporting similar results and comments regarding ChatGPT [14-16,18,19,22]. For diseases such as OM, answers to questions related to diagnosis may usually need to be explained using medical terms. If the model is not sufficiently optimized to simplify these terms, its responses may become more difficult to understand. Simplification, on the other hand, can result in shorter responses lacking crucial details [15,22]. Currently, AI models should aim not only to provide information, but also to ensure that their outputs are understandable, user-centered, and responsive to patients' needs. In the literature, three studies comparing LLMs on otorhinolaryngologic diseases have shown that ChatGPT outperformed Gemini (formerly Bard) and Copilot (Bing) in terms of response accuracy and quality of responses, while ChatGPT responses are more difficult to read than those of Gemini (formerly Bard) and Copilot (Bing) [10,13,14].

Nonetheless, the present study has certain limitations. First, only three AI tools were evaluated, and the inclusion of additional models might have yielded different or more favorable results. Readability indices may not fully capture the true readability of the responses, as these metrics rely on variables such as syllable or character counts; the presence of medical terms with fewer syllables or characters may inaccurately lower readability scores. Additionally, the use of English-only questions limits the generalizability of the findings to patient groups with diverse linguistic and cultural backgrounds. The absence of patient feedback regarding comprehensibility may also restrict the ability of the results to fully reflect patient perspectives. Thus, the lack of direct patient evaluation represents an important limitation. Finally, as LLMs are continuously updated, their responses may vary over time. This dynamic nature necessitates ongoing monitoring and verification to ensure the consistency, scope, and accuracy of the information provided.

Conclusion

In conclusion, NLP currently demonstrates high accuracy and efficiency in extracting information from diverse sources and supporting qualitative analysis. However, continuous updates and validation of AI-generated information are essential to ensure alignment with current medical guidelines. Since patients may seek information from AI models before initiating treatment, the accuracy and reliability of this information are critical to prevent misunderstanding or misinformation. It is imperative that treatment decisions are not delayed or disrupted due to misleading or inaccurate AI-derived content. Similarly, during the treatment and follow-up process, it is expected that the responses obtained from LLM will ensure the patient's proper participation in treatment and increase their compliance with treatment. Our study results suggest that the accuracy and comprehensiveness of the responses of ChatGPT, Gemini, and Copilot are similar across models, while Gemini responses seem to be more readable compared to ChatGPT and Copilot. Enhancing health literacy in the era of AI is essential to ensure that individuals can effectively engage with and interpret AI-generated medical information. Further well-designed studies are warranted to gain a deeper understanding of how AI-based tools influence patient comprehension, decision-making, and clinical outcomes.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: This study received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Ethics Committee Approval: This descriptive cross-sectional evaluation study did not require Research Ethics Board approval, as the data sources are publicly available and do not contain identifiable or sensitive information. The authors declare no affiliation with or involvement in any of the AI research and development companies included in this study

ORCID and Author Contributions: Y.Y. (0000-0003-2280-3843), R.T.S. (0000-0002-3866-2152). All authors contributed to all stages of the study. All authors read and approved the final

manuscript.

Peer-review: Externally peer reviewed.

REFERENCES

1. Schilder AG, Chonmaitree T, Cripps AW, Rosenfeld RM, Casselbrant ML, Haggard MP, et al. Otitis media. Nat Rev Dis Primers. 2016;2(1):16063. doi: 10.1038/nrdp.2016.63.
2. Jamal A, Alsabea A, Tarakmeh M, Safar A. Etiology, Diagnosis, Complications, and Management of Acute Otitis Media in Children. Cureus. 2022;14(8):e28019. doi: 10.7759/cureus.28019.
3. Ding X, Huang Y, Tian X, Zhao Y, Feng G, Gao Z. Diagnosis, Treatment, and Management of Otitis Media with Artificial Intelligence. Diagnostics (Basel). 2023;13(13):2309. doi: 10.3390/diagnostics13132309.
4. Huang AE, Valdez TA. Artificial Intelligence and Pediatric Otolaryngology. Otolaryngol Clin North Am. 2024;57(5):853-62. doi: 10.1016/j.otc.2024.04.011.
5. Banyi N, Ma B, Amanian A, Bur A, Abdalkhani A. Applications of Natural Language Processing in Otolaryngology: A Scoping Review. Laryngoscope. 2025;135(9):3049-63. doi: 10.1002/lary.32198.
6. Amanian A, Heffernan A, Ishii M, Creighton FX, Thamboo A. The Evolution and Application of Artificial Intelligence in Rhinology: A State of the Art Review. Otolaryngol Head Neck Surg. 2023;169(1):21-30. doi: 10.1177/01945998221110076.
7. Wilson BS, Tucci DL, Moses DA, Chang EF, Young NM, Zeng FG, et al. Harnessing the Power of Artificial Intelligence in Otolaryngology and the Communication Sciences. J Assoc Res Otolaryngol. 2022;23(3):319-49. doi: 10.1007/s10162-022-00846-2.
8. Lechien JR, Rameau A. Applications of ChatGPT in Otolaryngology-Head Neck Surgery: A State of the Art Review. Otolaryngol Head Neck Surg. 2024;171(3):667-77. doi: 10.1002/ohn.807.
9. Aliyeva A, Sari E, Alaskarov E, Nasirov R. Enhancing Postoperative Cochlear Implant Care With ChatGPT-4: A Study on Artificial Intelligence (AI)-Assisted Patient Education and Support. Cureus. 2024;16(2):e53897. doi: 10.7759/cureus.53897.
10. Ostrowska M, Kacała P, Onolememen D, Vaughan-Lane K, Sisily Joseph A, Ostrowski A, et al. To trust or not to trust: evaluating the reliability and safety of AI responses to laryngeal cancer queries. Eur Arch Otorhinolaryngol. 2024;281(11):6069-81. doi: 10.1007/s00405-024-08643-8.
11. Alamleh S, Mavedatnia D, Francis G, Le T, Davies J, Lin V, et al. Readability, Reliability, and Quality Analysis of Internet-Based Patient Education Materials and Large Language Models on Meniere's Disease. J Otolaryngol Head Neck Surg. 2025;54:1-10. doi: 10.1177/19160216251360651.
12. Grose EM, Holmes CP, Aravinthan KA, Wu V, Lee JM. Readability and quality assessment of internet-based patient education materials related to nasal septoplasty. J Otolaryngol Head Neck Surg. 2021;50(1):16. doi: 10.1186/s40463-021-00507-z.
13. de Souza LL, Santos-Silva AR, Hagag A, Alzahem A, Vargas PA, Lopes MA. Evaluating AI models in head and neck cancer research: the use of NCI data by ChatGPT 3.5, ChatGPT 4.0, Google Bard, and Bing Chat. Oral Surg Oral Med Oral Pathol Oral Radiol. 2024;138(3):453-7. doi: 10.1016/j.oooo.2024.05.012.
14. Bellinger JR, Kwak MW, Ramos GA, Mella JS, Mattos JL. Quantitative Comparison of Chatbots on Common Rhinology Pathologies. Laryngoscope. 2024;134(10):4225-31. doi: 10.1002/lary.31470.
15. Abou-Abdallah M, Dar T, Mahmudzade Y, Michaels J, Talwar R, Tornari C. The quality and readability of patient information provided by ChatGPT: can AI reliably explain common ENT operations? Eur Arch Otorhinolaryngol. 2024;281(11):6147-53. doi: 10.1007/s00405-024-08598-w.
16. Ajit-Roger E, Moise A, Peralta C, Orishchak O, Daniel SJ. Enhancing Multilingual Patient Education: ChatGPT's Accuracy and Readability for SSNHL Queries in English and Spanish. OTO Open. 2024;8(4):e70048. doi: 10.1002/oto2.70048.
17. Moise A, Centomo-Bozzo A, Orishchak O, Alnoury MK, Daniel SJ. Can ChatGPT Guide Parents on Tympanostomy Tube Insertion? Children (Basel). 2023;10(10):1634. doi: 10.3390/children10101634.
18. Shamil E, Ko TK, Fan KS, Schuster-Bruce J, Jaafar M, Khwaja S, et al. Assessing the Quality and Readability of Online Patient Information: ENT UK Patient Information e-Leaflets versus Responses by a Generative Artificial Intelligence. Facial Plast Surg. 2025;41(4):472-81. doi: 10.1055/a-2413-3675.
19. Campbell DJ, Estephan LE, Sina EM, Mastrolonardo EV, Alapati R, Amin DR, et al. Evaluating ChatGPT Responses on Thyroid Nodules for Patient Education. Thyroid. 2024;34(3):371-77. doi: 10.1089/thy.2023.0491.
20. Carnino JM, Rampam S, Puyo EM, Kennedy DG, Levi JR. Readability of Pediatric Otolaryngology Information: Comparing AI-Generated Content With Google Search Results. Otolaryngol Head Neck Surg. 2025;173(6):1478-184. doi: 10.1002/ohn.70011.
21. Kucu O, Pamuk AE, Sütay Süslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? Front Oncol. 2023;13:1-7. doi: 10.3389/fonc.2023.1256459.
22. Shen SA, Perez-Heydrich CA, Xie DX, Nellis JC. ChatGPT vs. web search for patient questions: what does ChatGPT do better? Eur Arch Otorhinolaryngol. 2024;281(6):3219-25. doi: 10.1007/s00405-024-08524-0.