AMERICAN ACADEMY OF OTOLARYNGOLOGY–HEAD AND NECK SURGERY FOUNDATION

Original Research—Pediatric Otolaryngology

Otolaryngology–Head and Neck Surgery
2025, Vol. 173(1) 228–236

DOI: 10.1002/ohn.1225
http://otojournal.org
WILEY

Comparative Analysis of Information Quality in Pediatric Otorhinolaryngology: Clinicians, Residents, and Large Language Models

Eleonora M. C. Trecca, MD, PhD[1,2], Vito Carlo Alberto Caponio, DDM, PhD[3], Mario Turri-Zanoni, MD, PhD[1,4,5], Antonella Miriam di Lullo, MD, PhD[1,6,7], Michele Gaffuri, MD[1,8,9], Jérôme R. Lechien, MD, PhD, FACS[10,11,12], Antonino Maniaci, MD, PhD[1,13], Giuseppe Maruccio, MD[14,15], Marella Reale, MD[1,16], Irene Claudia Visconti, MD[1,17], and Virginia Dallari, MD[1,18]

Abstract

Objective. Pediatric otorhinolaryngology (ORL) addresses complex conditions in children, requiring a tailored approach for patients and families. With artificial intelligence (AI) gaining traction in medical applications, this study evaluates the quality of information provided by large language models (LLMs) in comparison to clinicians, identifying strengths and limitations in the field of pediatric ORL.

Study Design. Comparative blinded study.

Setting. Controlled research environment using LLMs.

Methods. Fifty-four items of increasing difficulty, namely 18 theoretical questions, 18 clinical scenarios, and 18 patient questions, were posed to ChatGPT-3.5, -4.0, -4o, Claude-3, Gemini, Perplexity, Copilot, a second-year resident, and an expert in the field of pediatric ORL. The Quality Analysis of Medical Artificial Intelligence (QAMAI) tool was used for blinded evaluation of the quality of medical information by a panel of expert members from the Young Otolaryngologists Group of the Italian Society of ORL and the International Federation of ORL Societies.

Results. LLMs performed comparably to specialist in theoretical and standardized clinical scenarios, with Bing Copilot achieving the highest QAMAI scores. However, AI responses lacked transparency in citing reliable sources and were less effective in addressing patient-centered questions. Poor

[1]GOS, Young Otolaryngologists Group of the Italian Society of Otorhinolaryngology–Head and Neck Surgery, Rome, Italy
[2]Department of Otorhinolaryngology and Maxillofacial Surgery, IRCCS Research Hospital Casa Sollievo della Sofferenza, San Giovanni Rotondo, Foggia, Italy
[3]Department of Clinical and Experimental Medicine, University of Foggia, Foggia, Italy
[4]Department of Otolaryngology–Head and Neck Surgery, ASST Lariana, Ospedale Sant'Anna, University of Insubria, Como, Italy
[5]Department of Biotechnology and LifeSciences, University of Insubria, Varese, Italy
[6]Otorhinolaryngology–Head and Neck Surgery Unit, AORN "San Pio", Benevento, Italy
[7]CEINGE-Advanced Biotechnology, Naples, Italy
[8]Department of Otolaryngology and Head and Neck Surgery, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico of Milan, Milan, Italy
[9]Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy
[10]Department of Otolaryngology–Head and Neck Surgery, Division of Laryngology and Bronchoesophagology, Baudour, Belgium
[11]Department of Surgery, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium
[12]Department of Otolaryngology–Head Neck Surgery, Foch Hospital, School of Medicine, UFR Simone Veil, Université Versailles Saint-Quentin-en-Yvelines (Paris Saclay University), Paris, France
[13]Department of Otolaryngology, Kore University, Enna, Italy
[14]Department of Otolaryngology–Head and Neck Surgery, IRCCS Azienda Ospedaliero-Universitaria of Bologna, Bologna, Italy
[15]Alma Mater Studiorum - University of Bologna, Bologna, Italy
[16]Otorhinolaryngology Unity, Head and Neck Department, Meyer Children's Hospital IRCCS, Florence, Italy
[17]ENT Unit, Morgagni Pierantoni Hospital, Azienda USL della Romagna, Forlì, Forlì-Cesena, Italy
[18]Department of Otolaryngology Head and Neck Surgery, Santa Maria delle Croci Hospital, AUSL della Romagna, Ravenna, Italy

Corresponding Author:
Eleonora M. C. Trecca, MD, PhD, Department of Otorhinolaryngology and Maxillofacial Surgery, IRCCS Research Hospital Casa Sollievo della Sofferenza, Viale Cappuccini, 1, 71013 San Giovanni Rotondo, Foggia, Italy.
Email: eleonoramc.trecca@gmail.com; e.trecca@operapadrepio.it
EleonoraTrecca@Twitter.com

interrater agreement among reviewers highlighted challenges in distinguishing human-generated from AI-generated responses. Rhinology topics received the highest scores, whereas laryngology and patient-centered questions showed lower agreement and performance.

Conclusion. LLMs show promise as supportive resources in pediatric ORL, particularly in theoretical learning and standardized cases. However, significant limitations remain, including source transparency and contextual communication in patient interactions. Human oversight is essential to mitigate risks. Future developments should focus on refining AI capabilities for evidence-based and empathetic communication to support both clinicians and families.

Keywords

artificial intelligence, ChatGPT, digital health, eHealth, large language models, otolaryngology, pediatric otorhinolaryngology, pediatrics

Received December 18, 2024; accepted February 27, 2025.

Pediatric otorhinolaryngology (ORL) focuses on diagnosing and treating a wide range of diseases in pediatric patients. This field requires a tailored approach that addresses both the unique physiological characteristics of children and the concerns of their families. Unlike adult ORL, pediatric ORL often involves treating patients from birth through adolescence. This continuum of care presents challenges, as many conditions are congenital, anatomical spaces are small and variable, and the quality of life must be carefully considered in treatment decisions. In this context, parents play a crucial role as primary decision-makers, often having to choose the best treatment path for their children, which may include complex surgical interventions.[1]

In recent years, artificial intelligence (AI) has emerged as a transformative tool in the medical field, with applications ranging from clinical decision support to patient self-management, real-world drug research, and assistance in research studies.[2] In ORL, AI has shown promise for automating classification tasks, analyzing clinical data, and simulating preoperative outcomes, which can aid physicians in providing precise, personalized care.[3] Among AI advancements, language-based models like ChatGPT (OpenAI, Microsoft) have gained popularity due to their ability to generate human-like responses. These models can make medical information more accessible to patients and clinicians; however, concerns persist regarding the accuracy and reliability of the medical advice they provide, as errors or misunderstandings in responses could pose risks to patient safety.[4-6]

Given these opportunities and challenges, this study seeks to evaluate the potential use of AI in pediatric ORL, specifically in comparison to medical practitioners, to assess the strengths and limitations of AI as a supportive tool in this complex subspecialty.

The aim of this study is to assess the quality of medical information provided by large language models (LLMs) in pediatric ORL. We intend to evaluate the accuracy and reliability of AI responses across three key areas: theoretical knowledge, clinical decision-making, and patient-centered advice. Comparisons will be made with responses from human practitioners of varying expertize, including residents and experienced clinicians. This study also aims to identify both the advantages and pitfalls of using AI in pediatric ORL clinical practice, with particular emphasis on aspects that impact physicians and families.

Materials and Methods

This study evaluated responses to a set of 54 questions, designed to represent a range of complexities in pediatric ORL (Supplemental Appendices 1-3, available online). These questions were divided into three categories: 18 theoretical questions, 18 clinical scenarios, and 18 patient-centered questions. The three types of questions were grouped and pertained to the fields of pediatric otology, rhinology, and laryngology for a total of 162 items. For this study, three experts—a pediatric otologist (M.R.), a rhinologist (A.M.d.L.), and a laryngologist (I.C.V.)—generated fictitious results from hypothetical patients inspired by their clinical activities. For these reasons, patient safety was guaranteed, and formal ethical approval was waived. The items were designed to range from easy to more complex and were graded on a Likert scale (1-5) as easy (1-2), medium (3-4), and difficult (5). Agreement on the difficulty levels of the items was reached by the first and last authors (E.M.C.T., V.D.).

Responses were obtained from a range of AI models, including ChatGPT (OpenAI, Microsoft, versions 3.5, 4.0, and 4o), Claude-3 (Anthropic), Gemini (Google), Perplexity AI (Inc.), and Copilot (Microsoft). For comparison, responses were also collected from a second-year resident (G.M.) and an experienced pediatric ORL specialist (M.T.-Z.). All answers were extracted from the AI tool during the week of September 23 to 29, 2024, whereas resident and specialist were given 1 month to answer the questions (September 2024).

A panel of experts (M.G., A.M., and J.R.L.) from the Young Otolaryngologists Group of the Italian Society of ORL and the Young Otolaryngologists of IFOS (International Federation of ORL Societies) evaluated each response, scoring them on the accuracy and relevance of the medical information provided. The panel, namely reviewer 1 (Rev1), reviewer 2 (Rev2), and reviewer 3 (Rev3), was selected based on their proven record of a high number of publications in the field of pediatric ORL.

The Quality Analysis of Medical AI (QAMAI)[7] tool was used for a blinded assessment of response quality; the

scale analyzes six parameters of medical information: accuracy, clarity, relevance, completeness, sources, and usefulness. Each parameter is evaluated using a Likert scale from 1 (strongly agree) to 5 (strongly disagree). This methodology allowed for an objective comparison between AI-generated responses and those provided by human practitioners at different levels of expertize, enabling us to assess the feasibility of using AI as a supportive tool in the clinical practice of pediatric ORL.

We also asked the panel (Rev1, Rev2, and Rev3) to indicate for each item whether the author of the response was a human or if the answer was AI-generated. This step was included to ensure a blinded evaluation and to assess the ability of the AI tool to appear human-like in its responses. Furthermore, this approach was implemented to mitigate any risk of bias, as the QAMAI is a score created to evaluate the quality of medical information generated by AI. In this article, however, it was also used to assess the quality of medical information provided by humans.

Statistical Analysis

The statistical analysis aimed to evaluate the overall quality of responses provided by AI tools compared to clinicians in the field of pediatric ORL. QAMAI scores were summarized as total and mean scores, based also on subscores (ie, accuracy, clarity, relevance, etc.). Results were presented using descriptive statistics across categories (ie, responder type, question difficulty, and topics). Scores were averaged and tested for normality, using the Shapiro-Wilk test to guide the choice of parametric or nonparametric tests. Analysis of variance (ANOVA) or Kruskal-Wallis tests were employed to find differences in QAMAI scores among the question variables (ie, topic, difficulty, etc.) and responder type (ie, resident, Claude-3, others AI, etc.). Post hoc pairwise Wilcoxon test with Bonferroni correction was applied for multiple pairwise comparisons. The adjusted *P*-value is considered statistically significant if it is below .05. SPSS® Statistics applies the correction by maintaining the threshold at .05 and multiplying the retrieved *P*-value by the number of comparisons.

Interrater reliability was measured using Fleiss' kappa for multirater agreement across reviewers.

The IBM® SPSS® Statistics version 25 and R statistics were used to perform statistical analysis.

Results

Rater Agreement Among Reviewers

Multirater agreement among reviewers, evaluated using Fleiss' kappa, resulted in .05, indicating poor agreement. The three reviewers provided the same answers in only 35.8% of the cases (n = 174). Rev3 showed a stronger tendency to classify responses as AI-generated, whereas Rev1 and Rev2 exhibited a more balanced classification pattern. Specifically, Rev1 and Rev2 classified, respectively, 292 (60.1%) and 290 (59.7%) answers as AI, against 376 (77.4%) of Rev3. Agreement varied according to the difficulty of the questions, decreasing from 40.1% for easy questions to 36.4% for medium-difficulty questions and 30.9% for hard questions ($P = .003$). In topic-based analysis, the highest agreement was recorded in rhinology (50%), followed by otology (33.3%), and laryngeal topics (24.1%) ($P = .001$). When classifying patient-generated questions as either human- or AI-generated, agreement was the lowest (29.6%), whereas similar values were observed for clinical scenario and theoretical questions (37% and 40.7%, respectively) ($P = .005$). Data with specific subgroup topic, category, and difficulty are in **Table 1** with complete agreement.

When assessing the accuracy of the reviewers in correctly identifying AI- versus human-generated responses, Rev3 achieved the highest accuracy, correctly classifying 84% of responses. This result is likely due to the previously described tendency to assign responses as AI-generated. Rev1 and Rev2 had similar classification accuracy rates, correctly identifying 58.4% and 59.3% of responses, respectively ($P < .001$).

Overall Quality of Medical Information Provided by LLMs

Looking at the quality of the answers provided, the average QAMAI total score was 22.9 (SD = 3.1), with a median of 23 (range: 13.7-30). The average score across single subfields was consistent, with a mean of 4 (SD = 0.6) for usefulness, 4 (SD = 0.5) for completeness, 4.1 (SD = 0.6) for relevance and accuracy, and 4.2 (SD = 0.5) for clarity. The lowest score was recorded for the provided sources, with a mean value of 2.5 (SD = 1).

Comparisons Based on Topic, Category, and Difficulty

Average total QAMAI score, SD, median, minimum, and maximum values combined for topic, category, and difficulty are collected in **Table 2**. The Kruskal-Wallis test found statistically significant differences among topics and categories ($P = .007$ and $P = .02$; difficulty resulted in a nonstatistically significant difference, $P = .381$). Based on actual responders, grouped as AI, specialist, and resident, no statistically significant difference was found from the Kruskal-Wallis test ($P = .155$) ($\text{mean}_{AI} = 23$, SD = 2.9; $\text{mean}_{resident} = 22.1$, SD = 4; $\text{mean}_{specialist} = 23.6$, SD = 3.4). Two-way ANOVA models were built to investigate differences in mean total QAMAI scores between responder groups (AI, resident, and specialist) and, respectively, the topic, category, and difficulty. A single main effects analysis showed that there was a statistically significant higher score for the specialist only when compared to the resident (adj. $P = .03$), whereas no differences were found when comparing their answers to those provided by the AI. A single main effects analysis found similar results when considering the topic alone,

Table 1. Combined Group Based on Topic, Category, and Difficulty With Respective Full Agreement Among Reviewers

Question variables	Cases of full agreement	Total	Cases of full agreement, %
Larynx: clinical scenario, easy	6	18	33.3
Larynx: clinical scenario, hard	4	18	22.2
Larynx: clinical scenario, medium	4	18	22.2
Larynx: patient's question, easy	5	18	27.8
Larynx: patient's question, hard	0	18	0.0
Larynx: patient's question, medium	4	18	22.2
Larynx: theoretical, easy	4	18	22.2
Larynx: theoretical, hard	7	18	38.9
Larynx: theoretical, medium	5	18	27.8
Oto: clinical scenario, easy	6	18	33.3
Oto: clinical scenario, hard	6	18	33.3
Oto: clinical scenario, medium	5	18	27.8
Oto: patient's question, easy	5	18	27.8
Oto: patient's question, hard	7	18	38.9
Oto: patient's question, medium	5	18	27.8
Oto: theoretical, easy	8	18	44.4
Oto: theoretical, hard	5	18	27.8
Oto: theoretical, medium	7	18	38.9
Rhino: clinical scenario, easy	9	18	50.0
Rhino: clinical scenario, hard	9	18	50.0
Rhino: clinical scenario, medium	11	18	61.1
Rhino: patient's question, easy	9	18	50.0
Rhino: patient's question, hard	3	18	16.7
Rhino: patient's question, medium	10	18	55.6
Rhino: theoretical, easy	13	18	72.2
Rhino: theoretical, hard	9	18	50.0
Rhino: theoretical, medium	8	18	44.4

with higher scores achieved in the rhinology group only when compared to otology (adj. $P = .001$). The interaction between responder type and topic was not statistically significant. Two-way ANOVA was also conducted to examine the effect of responders and category with respect to the total mean QAMAI score. There was a statistically significant interaction between responders and question category ($P = .02$). The specialist achieved higher scores in theoretical scenarios compared to the resident answering clinical and theoretical questions (respectively, adj. $P = .02$, and $=.01$) and AI answering patients' questions (adj. $P = .04$). When difficulty was considered in the model, its simple main effects and interaction with the responder group were not statistically significant different ($P = .48$ and $=.99$, respectively).

LLMs Comparisons

When comparing different AI models, a statistically significant difference was found ($P = .023$), with Bing Copilot reporting the highest mean score (24.1, SD = 3). This score was significantly higher than that of Google Gemini (adj. $P = .032$) and ChatGPT-4o (adj. $P = .020$). Notably, Bing Copilot also achieved a significantly higher mean total QAMAI score compared to residents (adj. $P = .026$), whereas no significant difference was found between specialists and any of the AI models (**Table 3**).

Analysis of QAMAI Subfields

In the last instance, single QAMAI subfields were analyzed to determine which variables had the greatest impact on the overall total score. Means and SDs are depicted in **Table 4**. Overall, most subfields received high score, while the "resources provided" subfield had the lowest scores, particularly in the rhinology topic. The Kruskal-Wallis test showed no statistically significant differences in the mean scores of QAMAI subfields based on the difficulty of the questions or category, except for the "source provided" subfield ($P = .005$). However, all subfields showed statistically significant differences based on topics ($P < .001$). The highest scores were awarded in the rhinology topic, whereas otology and larynx topics had similar mean scores. When analyzing the responders' QAMAI subfield scores, aspects such as accuracy, clarity, relevance, and usefulness did not show significant differences. However, variable scores were observed for

Table 2. Average Total Quality Analysis of Medical Artificial Intelligence (QAMAI) Score, SD, Median, Minimum (min), and Maximum (max) Recorded Values for Topic, Category, and Difficulty

Topic	Category	Difficulty	Mean for total QAMAI score	SD	Median	Min	Max
Larynx	Clinical scenario	Easy	23.3	3.24	22.3	18.7	29.7
Larynx	Clinical scenario	Hard	22.9	2.73	23.3	19.3	28
Larynx	Clinical scenario	Medium	23.5	2.86	23.5	18.3	28.3
Larynx	Patient's question	Easy	22.3	3.41	22.3	15.7	27.3
Larynx	Patient's question	Hard	20.7	2.35	20	17.7	25.3
Larynx	Patient's question	Medium	20.4	3.13	19.7	15.7	26
Larynx	Theoretical	Easy	24.2	2.58	24.2	20.7	29
Larynx	Theoretical	Hard	23.9	2.95	23	17.7	28.3
Larynx	Theoretical	Medium	24.2	2.52	23.8	20	30
Oto	Clinical scenario	Easy	22.6	3.63	22.3	13.7	27.3
Oto	Clinical scenario	Hard	22.4	3.6	22.3	13.7	27
Oto	Clinical scenario	Medium	21.8	3.85	22	15	28
Oto	Patient's question	Easy	21.9	3.88	22.2	13.7	28
Oto	Patient's question	Hard	22.3	3.45	22.8	16	27
Oto	Patient's question	Medium	22.5	3.7	22.7	13.7	27.3
Oto	Theoretical	Easy	22.6	4.19	22.8	13.7	29.7
Oto	Theoretical	Hard	22.8	3.63	23.3	14.3	27
Oto	Theoretical	Medium	22.7	3.87	23	15	28.3
Rhino	Clinical scenario	Easy	22.7	1.76	22.8	17.7	25.3
Rhino	Clinical scenario	Hard	23.2	1.56	23.5	19.7	25.3
Rhino	Clinical scenario	Medium	22.6	2.21	22.7	16.7	25.7
Rhino	Patient's question	Easy	24.5	2.47	24.8	18.7	27.3
Rhino	Patient's question	Hard	23.7	2.17	23.5	20.3	28
Rhino	Patient's question	Medium	24.8	1.95	24.7	21	27.7
Rhino	Theoretical	Easy	23.9	3.07	24	17	30
Rhino	Theoretical	Hard	22.4	2.07	22.7	18.3	25.3
Rhino	Theoretical	Medium	24.6	1.88	24.2	21.7	28

Table 3. Total Mean Quality Analysis of Medical Artificial Intelligence Scores Specific to Resident, Specialist, and Single Artificial Intelligence (AI) Models

Real responder	Mean	Std. deviation	N
Resident	22.1	4.0	54
Specialist	23.6	3.4	54
ChatGPT-3.5	23.2	2.6	54
ChatGPT-4.0	23.1	2.7	54
Perplexity AI	22.9	2.9	54
Google Gemini	22.3	2.7	54
Bing Copilot	24.1	3.0	54
Claude-3	22.9	2.7	54
ChatGPT-4o	22.3	3.5	54
Total	22.9	3.1	486

the "sources provided" and completeness ($P < .001$ and $P = .048$, respectively). Specifically, the expert received higher scores compared to the resident, whereas no statistically significant differences were found among the other comparisons.

Discussion

To the best of our knowledge, this is the first study to evaluate the quality of information provided by seven LLMs in the field of pediatric ORL using blinded evaluation with human expertize. By evaluating responses from various LLMs, alongside those of a resident and an experienced specialist, we were able to assess the potential of AI as a supportive tool in clinical practice.

Our analysis revealed that AI models can achieve comparable quality scores to human respondents, particularly specialists, in theoretical and clinical scenarios. This indicates that AI tools have reached a level of sophistication where they can contribute meaningfully to decision-making in pediatric ORL. Among the AI models, Bing Copilot emerged as the top performer, underscoring the variability in performance across different platforms. This is consistent with the current literature with Bing Copilot demonstrating superior performance across several key metrics, including readability and comprehensiveness,[8] and superior accuracy compared to other LLMs in various areas of research.[9] In a related study, interestingly, Copilot demonstrated remarkable performance, achieving the second-highest

Table 4. Mean Quality Analysis of Medical Artificial Intelligence (QAMAI) Subfield Scores by Group, Topic, and Category

Topic	Category	Responder	QAMAI total (mean ± SD)	Accuracy (mean ± SD)	Clarity (mean ± SD)	Relevance (mean ± SD)	Completeness (mean ± SD)	Source (mean ± SD)	Usefulness (mean ± SD)
Larynx	Clinical scenario	AI	23.26 ± 2.77	3.97 ± 0.51	4.17 ± 0.45	3.98 ± 0.49	3.99 ± 0.45	3.18 ± 0.62	3.96 ± 0.50
Larynx	Clinical scenario	Resident	22.33 ± 2.80	3.89 ± 0.45	4.11 ± 0.40	3.72 ± 0.64	3.72 ± 0.39	3.05 ± 0.61	3.83 ± 0.46
Larynx	Clinical scenario	Specialist	24.05 ± 4.14	4.17 ± 0.70	4.17 ± 0.55	4.17 ± 0.7	4.17 ± 0.69	3.22 ± 0.93	4.17 ± 0.69
Larynx	Patient's question	AI	20.96 ± 2.68	3.59 ± 0.45	3.87 ± 0.40	3.61 ± 0.44	3.59 ± 0.42	2.71 ± 0.66	3.59 ± 0.44
Larynx	Patient's question	Resident	21.83 ± 4.42	3.89 ± 0.69	3.89 ± 0.72	3.89 ± 0.69	3.61 ± 0.77	2.72 ± 0.83	3.83 ± 0.78
Larynx	Patient's question	Specialist	21.39 ± 4.47	3.72 ± 0.71	3.72 ± 0.74	3.78 ± 0.65	3.72 ± 0.71	2.72 ± 1.00	3.72 ± 0.71
Larynx	Theoretical	AI	24.20 ± 2.55	4.17 ± 0.44	4.28 ± 0.43	4.23 ± 0.45	4.17 ± 0.49	3.23 ± 0.64	4.11 ± 0.46
Larynx	Theoretical	Resident	22.44 ± 2.75	3.83 ± 0.40	4.22 ± 0.50	3.94 ± 0.38	3.77 ± 0.45	2.88 ± 0.62	3.78 ± 0.45
Larynx	Theoretical	Specialist	25.17 ± 2.89	4.39 ± 0.49	4.33 ± 0.52	4.38 ± 0.49	4.28 ± 0.44	3.61 ± 0.83	4.17 ± 0.55
Oto	Clinical scenario	AI	22.53 ± 3.38	3.96 ± 0.57	4.09 ± 0.53	3.94 ± 0.57	3.95 ± 0.52	2.62 ± 0.84	3.97 ± 0.58
Oto	Clinical scenario	Resident	20.67 ± 5.78	3.55 ± 1.00	3.94 ± 0.77	3.50 ± 0.98	3.66 ± 0.84	2.50 ± 1.22	3.50 ± 1.00
Oto	Clinical scenario	Specialist	21.89 ± 3.09	3.89 ± 0.54	4.05 ± 0.39	3.89 ± 0.54	3.94 ± 0.53	2.22 ± 0.62	3.89 ± 0.54
Oto	Patient's question	AI	22.15 ± 3.52	3.86 ± 0.61	4.10 ± 0.53	3.87 ± 0.62	3.88 ± 0.54	2.55 ± 0.82	3.88 ± 0.62
Oto	Patient's question	Resident	23.67 ± 5.08	4.05 ± 0.88	4.28 ± 0.64	4.00 ± 0.90	4.17 ± 0.75	3.00 ± 1.05	4.17 ± 0.91
Oto	Patient's question	Specialist	21.17 ± 2.70	3.72 ± 0.44	4.00 ± 0.47	3.72 ± 0.44	3.78 ± 0.50	2.17 ± 0.41	3.78 ± 0.50
Oto	Theoretical	AI	22.57 ± 3.43	3.91 ± 0.57	4.11 ± 0.52	3.92 ± 0.58	3.96 ± 0.54	2.70 ± 0.82	3.96 ± 0.57
Oto	Theoretical	Resident	21.89 ± 5.80	3.89 ± 1.00	4.05 ± 0.74	3.89 ± 1.00	3.67 ± 0.81	2.61 ± 1.27	3.78 ± 1.02
Oto	Theoretical	Specialist	24.44 ± 4.55	4.28 ± 0.77	4.39 ± 0.53	4.28 ± 0.77	4.28 ± 0.68	2.94 ± 1.18	4.28 ± 0.77
Rhino	Clinical scenario	AI	22.81 ± 1.72	4.20 ± 0.35	4.32 ± 0.37	4.48 ± 0.30	4.17 ± 0.40	1.52 ± 0.64	4.13 ± 0.37
Rhino	Clinical scenario	Resident	21.39 ± 2.06	4.11 ± 0.58	4.22 ± 0.40	4.17 ± 0.28	3.83 ± 0.55	1.11 ± 0.27	3.95 ± 0.39
Rhino	Clinical scenario	Specialist	24.39 ± 1.60	4.61 ± 0.49	4.67 ± 0.21	4.72 ± 0.25	4.50 ± 0.28	1.33 ± 0.52	4.55 ± 0.34
Rhino	Patient's question	AI	24.27 ± 2.14	4.52 ± 0.40	4.50 ± 0.35	4.67 ± 0.34	4.40 ± 0.40	1.73 ± 0.90	4.43 ± 0.40
Rhino	Patient's question	Resident	24.67 ± 3.04	4.55 ± 0.40	4.67 ± 0.42	4.72 ± 0.33	4.50 ± 0.59	1.78 ± 0.96	4.44 ± 0.54
Rhino	Patient's question	Specialist	24.67 ± 2.18	4.72 ± 0.32	4.61 ± 0.53	4.83 ± 0.28	4.50 ± 0.50	1.50 ± 0.66	4.50 ± 0.50
Rhino	Theoretical	AI	23.91 ± 2.04	4.38 ± 0.35	4.45 ± 0.29	4.40 ± 0.29	4.21 ± 0.38	2.17 ± 0.97	4.30 ± 0.33
Rhino	Theoretical	Resident	19.83 ± 2.68	3.83 ± 0.55	3.83 ± 0.62	3.89 ± 0.45	3.39 ± 0.49	1.39 ± 0.13	3.50 ± 0.62
Rhino	Theoretical	Specialist	25.28 ± 2.27	4.67 ± 0.42	4.72 ± 0.32	4.67 ± 0.30	4.50 ± 0.41	2.11 ± 0.93	4.61 ± 0.39

Abbreviation: AI, artificial intelligence.

score among 108 otolaryngologists who participated in a public health care system ORL job competition examination.[10]

However, AI models were less effective when addressing patient-centered questions, suggesting limitations in contextual understanding and empathy. In fact, although LLMs are capable of identifying and addressing emotions, they lack the ability to perform deep reflective analysis of emotional experiences and the motivational aspects of emotions. This limitation is particularly evident in the use of generic empathic phrases that lack genuine contextual adaptation and a tendency toward overly verbose or unnecessarily elaborate replies. These findings underscore the need for significant improvements in AI performance to generate nuanced and contextually appropriate empathic communication.[11,12] Furthermore, they highlight the necessity for more robust and standardized evaluation strategies to effectively measure such soft skills in complex clinical contexts, such as those involving pediatric patients and their families, who often pose a wide range of emotionally charged and intricate questions.

Our results evidenced that AI models demonstrated strong performance in clarity, relevance, and usefulness of their responses. However, they consistently underperformed in providing transparent and reliable sources, which is a key limitation for integrating AI into evidence-based clinical practice. As highlighted by other authors,[13,14] ChatGPT has a tendency to generate references that may be accurate, erroneous, or entirely inexistent, depending on their alignment with existing records in scientific databases. This shortcoming highlights the need for improving AI models' ability to cite reliable medical literature and enhance the trustworthiness of their outputs to avoid risks of fabrication and plagiarism.[15]

Reviewer agreement decreased as question difficulty increased, with laryngology and patient-centered questions showing the lowest agreement rates. Interestingly, AI tools and specialist achieved higher scores in rhinology-related questions, suggesting that this topic might be more suited for AI-driven analysis due to a relatively higher degree of standardization in the field.[16,17]

The potential for AI in pediatric ORL is significant, especially for supporting clinicians in theoretical learning and standardized clinical scenarios. However, its current limitations—particularly in providing source transparency and effectively addressing nuanced patient-centered queries—must be addressed before widespread adoption. Additionally, the integration of AI into clinical workflows should be approached cautiously, with human oversight to mitigate risks associated with erroneous or incomplete AI-generated advice.

A major strength of this study is that, to the best of our knowledge, it evaluates the highest number of LLMs to date in this context. This broad inclusion allows for a comprehensive understanding of AI performance across different models, providing valuable insights into their potential role in pediatric ORL practice. Although the results provide meaningful insights, several limitations of the study should be acknowledged. First, the clinical scenarios were generated based on fictional patient cases, which, although inspired by real practice, may not fully capture the complexity and variability of actual clinical encounters. Also, responses were evaluated in a controlled setting rather than in real-world clinical environments where the dynamic nature of interactions with patients, caregivers, and health care providers could influence outcomes. Second, the study focused on a specific set of AI tools available at a particular time. Current advancements in AI may yield different results, limiting the generalizability of our findings to future models. Finally, although reviewers were blinded, inherent biases may have influenced their judgments. Notably, the QAMAI is a validated scoring system specifically designed to evaluate the quality of medical information generated by AI. Extending its use to assess responses provided by specialist and resident may introduce potential biases. However, reviewers often encountered objective difficulties in distinguishing whether responses were human- or AI-generated, a challenge particularly evident in Rev3's tendency to classify responses as AI-generated. Therefore, these observations support the suitability of QAMAI for research purposes, even when applied to the evaluation of human responses.

Our study underscores the importance of continued development and training of AI systems. Enhancing AI models' capabilities in sourcing reliable evidence and improving their understanding of patient-centered contexts could bridge existing gaps. Future research should focus on the longitudinal assessment of AI implementation in clinical practice to evaluate its impact on patient outcomes, clinician efficiency, and educational training.[18] In fact, especially following the COVID-19 pandemic, ORL resident education has increasingly incorporated asynchronous learning methods, including web-based platforms.[19] Although traditional textbook- and didactic-based education remains the gold standard, there is a growing need to consider a paradigm shift within academic ORL education. This shift should include the integration of high-quality electronic resources, where, after thorough evaluation and refinement, LLMs could play a central role.[20]

Conclusions

This study demonstrates that LLMs can perform comparably to human respondents in many areas of pediatric ORL, with specialists only marginally outperforming the highest-performing AI tools. However, LLMs still faces significant challenges, particularly in patient communication and source transparency. These findings highlight the dual nature of AI in clinical practice: as a promising tool with considerable potential and as one requiring cautious integration and refinement. AI can serve as a valuable complementary resource in clinical practice and medical education, particularly for residents and early-career clinicians. However, its limitations underscore the need

for continued development and human oversight to ensure patient safety and optimal care outcomes. With iterative advancements and careful integration, AI could become a transformative tool in pediatric ORL, supporting clinicians and enhancing the quality of care for young patients and their families.

Author Contributions

All authors (Eleonora M. C. Trecca, Vito Carlo Alberto Caponio, Mario Turri-Zanoni, Antonella Miriam di Lullo, Michele Gaffuri, Jérôme R. Lechien, Antonino Maniaci, Giuseppe Maruccio, Marella Reale, Irene Claudia Visconti, and Virginia Dallari) meet the ICMJE criteria: (1) Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work. (2) Drafting the work or revising it critically for important intellectual content. (3) Final approval of the version to be published. (4) Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Disclosures

Competing interests: The authors declare that there is no conflicts of interest.

Funding source: None.

Data Availability Statement

Supplementary material is available upon request from the corresponding author. For further details, Appendices 1-3 are available.

Supplemental Material

Additional supporting information is available in the online version of the article.

ORCID iD

Eleonora M. C. Trecca http://orcid.org/0000-0001-6490-1746
Vito Carlo Alberto Caponio http://orcid.org/0000-0001-5080-5921
Mario Turri-Zanoni http://orcid.org/0000-0002-3678-9088
Antonella Miriam di Lullo https://orcid.org/0000-0002-4482-6782
Michele Gaffuri http://orcid.org/0000-0002-5435-685X
Jérôme R. Lechien http://orcid.org/0000-0002-0845-0845
Antonino Maniaci http://orcid.org/0000-0002-1251-0185
Giuseppe Maruccio https://orcid.org/0009-0008-6474-2122
Marella Reale http://orcid.org/0000-0002-9716-8605
Irene Claudia Visconti http://orcid.org/0000-0002-5831-1042
Virginia Dallari http://orcid.org/0000-0001-9260-1003

X (formerly known as Twitter)

Eleonora M. C. Trecca 𝕏@EleonoraTrecca

References

1. Trecca EMC, Gaffuri M, Molinari G, et al. Impact of the COVID-19 pandemic on paediatric otolaryngology: a nationwide study. *Acta Otorhinolaryngol Ital*. 2023;43(5):352-359. doi:10.14639/0392-100X-N2452
2. Ariyaratne S, Iyengar KP, Nischal N, Chitti Babu N, Botchu R. A comparison of ChatGPT-generated articles with human-written articles. *Skeletal Radiol*. 2023;52(9):1755-1758. doi:10.1007/s00256-023-04340-5
3. Seifen C, Huppertz T, Gouveris H, et al. Chasing sleep physicians: ChatGPT-4o on the interpretation of polysomnographic results. *Eur Arch Otrhinolaryngol*. Published online October 20, 2024. doi:10.1007/s00405-024-08985-3
4. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *J Med Internet Res*. 2023;25:e46924. doi:10.2196/46924
5. Vaishya R, Misra A, Vaish A. ChatGPT: is this version good for healthcare and research? *Diabetes Metab Syndr Clin Res Rev*. 2023;17(4):102744. doi:10.1016/j.dsx.2023.102744
6. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*. Published online February 19, 2023. doi:10.7759/cureus.35179
7. Vaira LA, Lechien JR, Abbate V, et al. Validation of the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool: a new tool to assess the quality of health information provided by AI platforms. *Eur Arch Otrhinolaryngol*. 2024;281(11):6123-6131. doi:10.1007/s00405-024-08710-0
8. Lim B, Lirios G, Sakalkale A, Satheakeerthy S, Hayes D, Yeung JMC. Assessing the efficacy of artificial intelligence to provide peri-operative information for patients with a stoma. *ANZ J Surg*. Published online December 2, 2024. doi:10.1111/ans.19337
9. Kaftan AN, Hussain MK, Naser FH. Response accuracy of ChatGPT 3.5 Copilot and Gemini in interpreting biochemical laboratory data a pilot study. *Sci Rep*. 2024;14(1):8233. doi:10.1038/s41598-024-58964-1
10. Mayo-Yáñez M, Lechien JR, Maria-Saibene A, Vaira LA, Maniaci A, Chiesa-Estomba CM. Examining the performance of ChatGPT 3.5 and Microsoft Copilot in otolaryngology: a comparative study with otolaryngologists' evaluation. *Indian J Otolaryngol Head Neck Surg*. 2024;76(4):3465-3469. doi:10.1007/s12070-024-04729-1
11. Vzorin GD, Bukinich AM, Sedykh AV, Vetrova II, Sergienko EA. The emotional intelligence of the GPT-4 large language model. *Psychol Russ State Art*. 2024;17(2):85-99. doi:10.11621/pir.2024.0206
12. Sorin V, Brin D, Barash Y, et al. Large language models and empathy: systematic review. *J Med Internet Res*. 2024;26:e52597. doi:10.2196/52597
13. Lechien JR, Briganti G, Vaira LA. Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology–head and neck surgery. *Eur Arch Otrhinolaryngol*. 2024;281(4):2159-2165. doi:10.1007/s00405-023-08441-8
14. Frosolini A, Franz L, Benedetti S, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. *Eur Arch Otrhinolaryngol*. 2023;280(11):5129-5133. doi:10.1007/s00405-023-08205-4
15. Elali FR, Rachid LN. AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns*. 2023;4(3):100706. doi:10.1016/j.patter.2023.100706

16. Ghosh Moulic A, Gaurkar SS, Deshmukh PT. Artificial intelligence in otology, rhinology, and laryngology: a narrative review of its current and evolving picture. *Cureus*. Published online August 2, 2024. doi:10.7759/cureus.66036
17. Radulesco T, Saibene AM, Michel J, Vaira LA, Lechien JR. ChatGPT-4 performance in rhinology: a clinical case series. *Int Forum Allergy Rhinol*. 2024;14(6):1123-1130. doi:10.1002/alr.23323
18. Dallari V, Sacchetto A, Saetti R, Calabrese L, Vittadello F, Gazzini L. Is artificial intelligence ready to replace specialist doctors entirely? ENT specialists vs ChatGPT: 1-0, ball at the center. *Eur Arch Otrhinolaryngol*. 2024;281(2):995-1023. doi:10.1007/s00405-023-08321-1
19. Malka RE, Marinelli JP, Newberry TR, Carlson ML, Bowe SN. Asynchronous learning among otolaryngology residents in the United States. *Am J Otolaryngol*. 2022;43(5):103575. doi:10.1016/j.amjoto.2022.103575
20. Dallari V, Liberale C, De Cecco F, et al. The role of artificial intelligence in training ENT residents: a survey on ChatGPT, a new method of investigation. *Acta Otorhinolaryngol Ital*. 2024;44(3):161-168. doi:10.14639/0392-100X-N2806