European Archives of Oto-Rhino-Laryngology (2025) 282:4325–4334
https://doi.org/10.1007/s00405-025-09504-8

MISCELLANEOUS

Clinical decision support using large language models in otolaryngology: a systematic review

Rania Filali Ansary[1] · **Jerome R. Lechien**[1,2,3,4]

Received: 3 May 2025 / Accepted: 27 May 2025 / Published online: 6 June 2025

Abstract

Objective This systematic review evaluated the diagnostic accuracy of large language models (LLMs) in otolaryngology-head and neck surgery clinical decision-making.

Data sources PubMed/MEDLINE, Cochrane Library, and Embase databases were searched for studies investigating clinical decision support accuracy of LLMs in otolaryngology.

Review methods Three investigators searched the literature for peer-reviewed studies investigating the application of LLMs as clinical decision support for real clinical cases according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The following outcomes were considered: diagnostic accuracy, additional examination and treatment recommendations. Study quality was assessed using the modified Methodological Index for Non-Randomized Studies (MINORS).

Results Of the 285 eligible publications, 17 met the inclusion criteria, accounting for 734 patients across various otolaryngology subspecialties. ChatGPT-4 was the most evaluated LLM (n = 14/17), followed by Claude-3/3.5 (n = 2/17), and Gemini (n = 2/17). Primary diagnostic accuracy ranged from 45.7 to 80.2% across different LLMs, with Claude often outperforming ChatGPT. LLMs demonstrated lower accuracy in recommending appropriate additional examinations (10–29%) and treatments (16.7–60%), with substantial subspecialty variability. Treatment recommendation accuracy was highest in head and neck oncology (55–60%) and lowest in rhinology (16.7%). There was substantial heterogeneity across studies for the inclusion criteria, information entered in the application programming interface, and the methods of accuracy assessment.

Conclusions LLMs demonstrate promising moderate diagnostic accuracy in otolaryngology clinical decision support, with higher performance in providing diagnoses than in suggesting appropriate additional examinations and treatments. Emerging findings support that Claude often outperforms ChatGPT. Methodological standardization is needed for future research.

Level of evidence NA.

Keywords Artificial intelligence · Large language model · Otolaryngology · Otorhinolaryngology · Generative artificial intelligence

Introduction

Large language models (LLMs) are increasingly used in medicine and surgery as adjunctive tools for clinical and basic science research, scientific paper grammar and spelling improvement, referencing, patient information, real cases and clinical vignette management [1–4]. Some practitioners report using LLMs in clinical practice [5, 6], making the assessment of diagnostic accuracy an important topic of research. To date, Chatbot Generative Pre-trained Transformer (ChatGPT) is considered as the most popular and primary LLM used in Medicine with variable degrees of diagnostic and treatment accuracy on fictive clinical

✉ Jerome R. Lechien
Jerome.Lechien@umons.ac.be

1 Department of Surgery, Faculty of Medicine, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), University of Mons, 6, Mons B7000, Belgium

2 Department of Otolaryngology-Head and Neck Surgery, School of Medicine, Foch Hospital, University Paris Saclay, Paris, France

3 Department of Otolaryngology-Head and Neck Surgery, CHU Saint-Pierre, Brussels, Belgium

4 Department of Otolaryngology, Elsan Hospital, Paris, France

vignettes, real clinical cases, or very rare conditions [3, 4, 7]. In a 2024 state-of-the-art review [3], preliminary data suggested that ChatGPT reported 47–79% accuracy in providing a plausible primary diagnosis, and low scores for selecting the most adequate additional examinations. Since the publication of this state-of-the-art review, the number of studies dedicated to the assessment of LLM accuracy in the management of real clinical cases has substantially increased, particularly regarding other LLMs, such as Claude Sonnet, Gemini, Bard, and Large Language Model Meta AI (LLaMA). Moreover, updated versions of some LLMs can analyze clinical images, radiological studies, pathological findings, and video content, which may theoretically improve their accuracy [8, 9].

This systematic review evaluated the diagnostic accuracy of large language models (LLMs) in otolaryngology-head and neck surgery clinical decision-making.

Materials and methods

The criteria for study inclusion and exclusion were based on the population, intervention, comparison, outcome, timing, and setting (PICOTS) framework [10]. The data collection was performed by three independent authors (RFA, KC and JRL) according to the PRISMA checklist for systematic reviews [11].

Types of studies

Retrospective and prospective case series, along with controlled and uncontrolled prospective studies published between January 2020 and February 2024 were included if they investigated LLM accuracy or concordance with practitioner decisions in managing real clinical cases from otolaryngology settings. The studies were published in English, Spanish, or French peer-reviewed journals. The authors considered pre-print papers. Case reports were excluded, and only studies reporting data for ≥ 5 cases were considered.

Populations, inclusion, and exclusion criteria

Studies evaluating LLM performance in real clinical cases were included if they reported clear methodology, inclusion and exclusion criteria. Potential overlap between clinical studies published by the same research teams was assessed, with smaller cohorts excluded. However, when overlapping studies presented substantially different outcomes, all relevant studies were included in the analysis.

Outcomes

The following outcomes were evaluated: study design, number of cases, subspecialty distribution, patient demographics, LLM configurations, data input methodology, evaluation metrics (including evaluations of diagnostic accuracy, primary diagnosis establishment, additional examination and treatment recommendations), and model performance comparisons (comparative studies).

Intervention and comparison

The methodology for evaluating LLM outputs and their comparative assessment *versus* practitioners, expert panels, or other LLMs was systematically analyzed.

Timing and setting

There were no criteria for specific timing in the evaluation process of the LLMs.

Search strategy

The literature research was conducted by two investigators (RFA, JRL, KC) through PubMed/MEDLINE, Cochrane Library and Embase databases for relevant peer-reviewed publications related to the LLM accuracy in clinical decision-making. The following keywords were used: ‘artificial intelligence’, ‘large language model’, ‘machine learning’, ‘ChatGPT’, ‘GPT-4’, ‘Claude’, ‘Gemini’, ‘LLaMA’, ‘Bard’, ‘clinical decision’, ‘diagnosis’, ‘treatment’, ‘management’, ‘otorhinolaryngology’, ‘otolaryngology’, ‘accuracy’, ‘performance’, and ‘evaluation’ to identify clinical studies, reviews, and meta-analyses. The authors considered studies with and without database abstracts. The papers had available full texts or titles with the search terms. The findings were reviewed for relevance, and the reference lists of these articles were examined for additional pertinent studies. Potential discrepancies in the literature search results were resolved by an external senior otolaryngologist.

Bias analysis

The authors carried out a bias analysis with the Methodological Index for Non-Randomized Studies (MINORS) tool [12], which is a validated instrument designed for assessing the quality of non-randomized studies. The MINORS tool includes 12 items related to the analysis of methodological points of studies. The items were scored 0 if absent; 1 when reported but inadequate; and 2 when reported and adequate.

The inclusion of cases was evaluated in terms of consecutive inclusion (0 or 2), while the data collection was rated as prospective (2), retrospective analysis of prospectively collected data (1), or absent (0). The quality of endpoints was judged as high (2) when authors evaluated the LLM accuracy in primary and differential diagnosis, additional examinations, and treatment recommendations; incomplete (1) when assessing only diagnostic accuracy; and low (0) when evaluations focused on non-diagnosis accuracy or if there were lacking clear evaluation criteria. The evaluations of outcomes by independent judges was ideal (2), while the evaluation by a single judge or two unblinded judges was evaluated as inadequate. About the follow-up period, the assessment of the stability of LLM's outputs over regenerated prompts was evaluated as present/adequate (2) or lacking (0). The study size calculation needed to be carried out (2), mentioned as unnecessary (1), or absent (0). The ideal MINORS score was 16 for non-comparative studies and 24 for comparative studies.

Results

Of the 1,132 identified studies, 17 publications met our inclusion criteria (Fig. 1) [4, 7, 8, 9, 13–25]. All studies were cross-sectional with either prospective or retrospective case inclusion. Studies represented various otolaryngology subspecialties: head and neck oncology (n=6) [9, 15, 16, 18, 22, 23], general otolaryngology (n=4) [4, 13, 21], laryngology (n=3) [8, 17, 19], oral and maxillofacial surgery (n=1) [14], rhinology (n=1) [20], pediatric otolaryngology (n=1) [24], otology (n=1) [25], and very rare otolaryngological disorders (n=1) (Table 1).[7] Excluding a potential overlap study [4], the present review included findings of 734 patients. Among studies reporting data of clinical/histopathological images, [8, 9, 22, 25] it was not possible to determine the patient count in one study [25]. Studies used the following LLMs for their analyses: ChatGPT-4 (n=14/17), [4, 8, 9, 14–18, 20–25] ChatGPT-3.5 (n=3/17) [13, 14, 19], ChatGPT-4o n=1/17) [7], Claude-3.0 Opus (n=17) [15], Claude-3.5 Sonnet (n=1/17) [7], Google Bard (n=1/17) [13], Gemini-1.5-Pro (n=1/17) [7], Gemini Advanced (n=1/17) [16], Llama-2.0 (n=1/17) [17], and Bing-GPT4 (n=1/17) [13].

Evaluation methodologies

Expert panel was used in the majority of studies (n=15/17), ranging from 2 to 57 independent experts (Table 1) [4, 7, 8, 13–22, 24, 25]. Alami et al. compared the accuracy of LLM in oncological case decisions against both NCCN guidelines and multidisciplinary oncological board determinations [23]. In the study of Schmidl et al., LLM was evaluated for diagnostic accuracy in clinical oncological cases (carcinomas) and leukoplakia without expert assessment of the LLM responses [9]. The profile of experts consisted of otolaryngologists-head and neck surgeons (n=10/15) [4, 7, 8, 13, 15, 18–21, 24], dental or maxillofacial practitioners (n=1/15) [14], or association of otolaryngologists with the following specialists: maxillofacial surgeons (n=1/15) [16], speech-therapist (n=1/15) [17], pathologist (n=1/15) [22], and pediatricians [25].

The accuracy evaluation of LLM was carried out with the following tools: validated artificial intelligence performance instrument (AIPI) (n=12/17), [4, 7–9, 14–18, 20, 21, 24] a binary evaluation (yes/no) (n=2/17) [13, 22], variable Likert-scale (n=3/17) [8, 9, 14], total disagreement score (n=1/17) [16], and a modified version of the Ottawa clinical assessment tool (n=1/17) [19].

Large language model accuracy for diagnoses

ChatGPT-3.5/4/4o, Claude-3/3.5, Bard, Gemini, and Bing-GPT4 were assessed for primary diagnostic accuracy in at least one study (Table 2), whereas only ChatGPT-3.5/4/4o, Claude-3/3.5, and Gemini were evaluated across multiple investigations. There were substantial consistent ranges of primary diagnosis accuracy values across studies for the several versions of ChatGPT and Claude Sonnet (Table 2). Tomo et al. compared the accuracy of ChatGPT-3.5 versus ChatGPT-4 for oral and maxillofacial conditions associated with typical pictures [14]. While they reported similar stability of both LLM versions through regenerated prompts, ChatGPT-4 surpassed ChatGPT-3.5 for primary diagnosis accuracy with 80.2% *versus* 61.8% of correct diagnoses. ChatGPT-4 was compared to Claude-3 and 3.5 in two studies [7, 15]. Schmidl et al. showed that both LLMs were similar in terms of oncological treatment recommendations and explanations, but Claude-3 demonstrated higher accuracy for diagnostic work-up than ChatGPT-4 [15]. ChatGPT-4o and Claude-3.5 were challenged for the diagnosis of very rare conditions in otolaryngology (case reports) [7]. In this study, Claude-3.5 reported a significantly higher rate of accurate primary diagnosis than ChatGPT-4 (54.3% *versus* 45.7%); the advantage of Claude-3.5 over ChatGPT-4o being stable when considering common otolaryngological consultation situations [7].

Accuracy for differential diagnoses was evaluated in 4 studies [7, 8, 19, 21]. The differential diagnosis accuracy ranges from 28.3 to 90% for ChatGPT. The lowest accuracy (28.3%) was found when considering differential diagnoses of laryngology images [8], whereas ChatGPT-4 performances were substantially higher for otological [25], and histopathological [22] images. For studies evaluating the differential diagnosis accuracy in clinical cases without image interpretation, the differential diagnosis accuracy

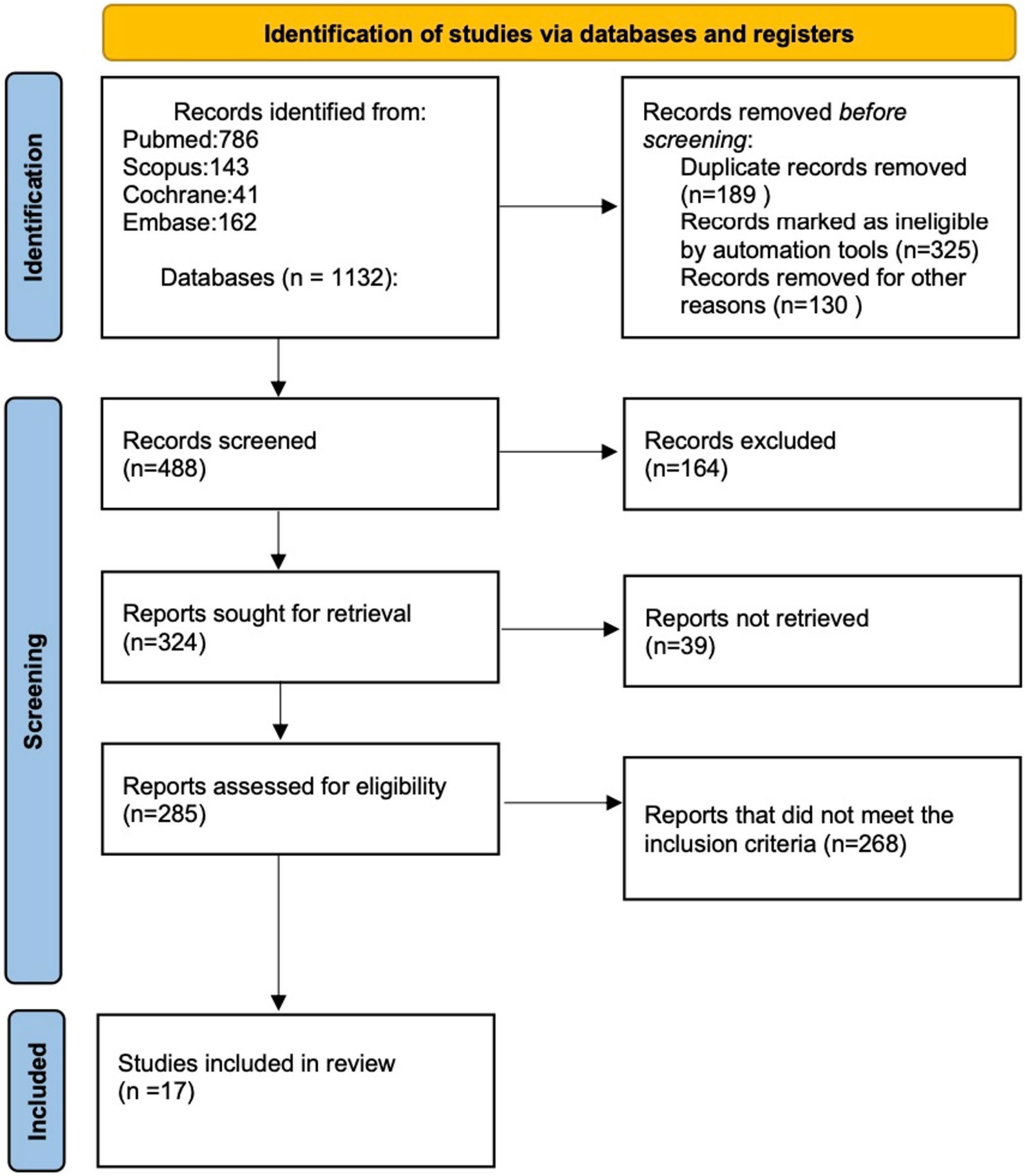

Fig. 1 Chart flow. Three independent investigators conducted the literature search

Table 1 Studies assessing the large Language model accuracy in clinical decision support

References	Subspecialty & cases	Expert/panel	LLMs	Tools	Outcomes	Results
Warrier [13]	General OTO (n=100)	18 OTO	ChatGPT-3.5	Correct	PD accuracy (GPT3.5-Bard-Bing)	89%-82%-74%; GPT3.5 > Bard > Bing
			Google Bard	PD		
			Bing-GPT4			
Tomo [14]	Oral-Maxillofacial (n=37)	18 OPT	ChatGPT-3.5	AIPI		*GPT3.5/4/OPT/GDS/DS*
		23 GDS	ChatGPT-4	4-point	Mean correct diagnoses	64.9-80.2-86.6%-24.3%-16.7%
		16 DS		LS	Mean correct primary diagnosis	46.0-61.8-82.3%-22.7%-15.8%
					Consistency (test-retest - GPT3.5, 4.0)	k = 0.532, 0.533
Schmidl [15]	Head and Neck (n=50)	2 independent OTO	Claude-3	AIPI	Diagnostic work-up	Claude 3 > ChatGPT4.0.
			ChatGPT-4		Treatment recommendations	Claude 3 = ChatGPT4.0
					Explanation/summarization	Claude 3 = ChatGPT4.0
Lorenzi [16]	Head and Neck (n=5)	2 blinded panels	ChatGPT-4	TDS	Treatment recommendations	ChatGPT4.0 > Gemini
		Each = 1 OTO-1 MF	Gemini	AIPI	Adherence to guidelines	ChatGPT4.0 > Gemini
		NCCN guidelines	Advanced	QAMAI	AIPI	ChatGPT4.0 > Gemini
Lechien [4]	Laryngology (n=34)	2 independent OTO	ChatGPT-4	AIPI	Additional examination (mean/patient N)	3.3 vs. 2.1; ChatGPT-4 > OTO
	Head and Neck (n=29)	per subspecialty			Work-up, PD, TT accuracy	18-62%-25%
	Rhinology (n=26)					
	Otology (n=11)					
Dronkers [17]	Laryngology (n=20)	Local laryngology	ChatGPT-4	AIPI	Treatment consistency	ChatGPT-4 > Llama-2
	BVFP	team: 2 LA, 1 SPL	Llama-2.0		ChatGPT-4 accuracy	50%
					Llama-2.0 accuracy	15%
Lechien [18]	Head and Neck (n=20)	2 independent OTO	ChatGPT-4	AIPI	Additional examination (mean/patient N)	3.6 vs. 5.2; ChatGPT-4 > OTO
					Add. Examination, TT, cTNM accuracy	25 − 55 − 95%
Lechien [19]	Laryngology and	3 independent OTO	ChatGPT-3.5	OCAT	Additional examination (mean/patient N)	2.8 vs. 1.8; ChatGPT-4 > OTO
	Head and Neck (n=40)				Add. Examination, TT, PD, DD accuracy	10%, 60%, 65%, 90%
Radulesco [20]	Rhinology (n=40)	3 independent OTO	ChatGPT-4	AIPI	Additional examination (mean/patient N)	3.3 vs. 1.7; ChatGPT-4 > OTO
					Add. Examination, TT, PD accuracy	15.8-16.7-50.8%
Lechien [21]	General OTO (n=45)	2 independent OTO	ChatGPT-4	AIPI	Most pertinent additional examination (%)	ChatGPT-4 = 27.5%
					Add. Examination, TT, PD, DD accuracy	29%-22%-63.5%-63.5%
Sievert [22]	Patho-oncological images	2 independent OTO	ChatGPT-4	2-item	Diagnosis accuracy	71.2% (GPT-4) 88.5% (experts)
	(n=139 from 5 patients)	1 pathologist		score	Consistency (test-retest - GPT-4)	k = 0.773
Schmidl [9]	Clinical oncological	None	ChatGPT-4	AIPI	Cancer - leukoplakia detection based on image	26 − 86.7%
	image (n=45)			4-point LS	Cancer-leukoplakia detection based on image/MH	73.3 − 93.3%
					Cancer-leukoplakia detection based on MH	73.3 − 80%
Alami [23]	Head and Neck (n=263)	NCCN guidelines	ChatGPT-4	MOM	MOM-ChatGPT-4 TT agreement, SE, SP	k = 0.35, 67 − 89%
		MOM decisions		decision	NCCN-ChatGPT-4 TT agreement, SE, SP	k = 0.72, 86 − 96%
Maniaci [8]	Laryngology (n=40)	3 independent	ChatGPT-4	AIPI	Additional examination (N tot)	153 − 93; ChatGPT-4 > OTO
	Image (n=20/40)	Laryngologists		5-point LS	Add. Examination, TT, PD, DD accuracy	12.5 − 25% − 22.5 − 28.3%
Maniaci [24]	Pediatric OTO (n=49)	2 independent OTO	ChatGPT-4	ACCS	Additional examination (N tot)	111 − 60; ChatGPT-4 > OTO
				AIPI	PD, management plan accuracy	78.6% − 28.6%
Noda [25]	Otology (n=305 images)	8 PED, 8 OTO,	ChatGPT-4	NP	Image, PD AOM, PD OME accuracy	82.1%, 89.2%, 85.7%

Table 1 (continued)

References	Subspecialty & cases	Expert/panel	LLMs	Tools	Outcomes	Results
		8 OTOR, 6 EOTO			ChatGPT-4 versus human (performance)	ChatGPT-4 > PED
Lechien [7]	Rare General OTO (n = 35)	2 independent OTO	ChatGPT-4o	AIPI	PD accuracy (Claude-ChatGPT-Gemini) RD	54.3% −45.7% −28.6%
			Claude-3.5		DD accuracy (Claude-ChatGPT-Gemini) RD	Claude > ChatGPT > Gemini
			Gemini-1.5		PD accuracy (Claude-ChatGPT-Gemini) CD	88.6% −77.1% −71.4%

Abbreviations: Abbreviations: ACCS = Amsterdam Clinical Challenge Scale; AOM = acute otitis media; BVFP = bilateral vocal fold paralysis; CD = common disease; DD = differential diagnosis; DS = dental students; EOTO = experienced OTO; GDS = general dental surgeons; HN = head and neck; LA = laryngologist; LS = Likert scale; MF = maxillofacial surgeon; MH = medical history; MOM = multidisciplinary oncology meeting; NCCN = National Comprehensive Cancer Network; NP = not provided; OCAT = Ottawa clinical assessment tool; OME = otitis media with effusion; OPT = oral pathology training dental practitioner; OTOR = resident in otolaryngology; PD = primary diagnosis; PED = pediatricians; QAMAI = Quality Assessment of Medical Artificial Intelligence; RD = rare disease; SE = sensitivity; SLT = second line treatment; SP = specificity; SPL = speech language pathologist; TDS = Total disagreement score; TT = treatment

ranged from 63.5 to 90% [19, 21], with Claude-3.5 demonstrating superior performance compared to ChatGPT-4o and Gemini-1.5 [7].

Large Language model accuracy for additional examinations and treatments

The performance of LLMs, particularly ChatGPT-3.5 and 4, in recommending the most appropriate additional examinations was investigated in seven studies [4, 8, 18–21, 24]. ChatGPT-3.5/4 recommended adequate additional examinations in 10 to 29% of cases (Table 2), which represents a lower mean accuracy rate compared to those related to primary and differential diagnoses. The low accuracy of ChatGPT in proposing adequate additional examinations is related to its inability to select the most appropriate examinations. This finding was illustrated in 5 studies with a significantly higher number of recommended additional examinations per patient from ChatGPT *versus* otolaryngologists [8, 18–20, 24]. Schmidl et al. compared diagnostic work-up performance of Claude-3 *versus* ChatGPT-4, demonstrating that Claude-3 more effectively selected appropriate additional examinations than ChatGPT-4 [15]. Apart from the study by Schmidl et al., the capability of LLMs to recommend adequate additional examinations was not extensively evaluated for the other LLMs.

ChatGPT-3.5 and particularly ChatGPT-4 were the most extensively evaluated LLMs for treatment recommendation accuracy across otolaryngological subspecialties (Table 2). The accuracy of ChatGPT in proposing appropriate therapeutic options demonstrated marked subspecialty variability, ranging from 16.7 to 60%. Oncological applications yielded the highest accuracy rates (55–60%) [18, 19], while significantly lower performance was observed in laryngology (25%) and rhinology (16.7%) [20]. Treatment recommendations were compared across LLMs in 3 studies [15–17]. Lorenzi et al. suggested better therapeutic oncological recommendations from ChatGPT-4 over Gemini [16], which was attributed to ChatGPT-4's superior adherence to clinical guidelines compared to Gemini. Alami et al. [23]. corroborated this finding, demonstrating that ChatGPT-4 exhibited high rates of adherence to clinical oncological guidelines when proposing primary or alternative therapeutic recommendations. The potential superiority of ChatGPT-4 over other LLMs in oncology was not supported by the findings of Schmidl et al., who did not find significant differences between Claude-3 and ChatGPT-4 [15]. In laryngology, Dronkers et al. compared the therapeutic recommendations of Llama-2 *versus* ChatGPT-4 for bilateral vocal cord paralysis. In this specific topic, ChatGPT-4 surpassed Llama-2 with 50% versus 15% of correct management [17].

Table 2 Summary of accuracy findings of large Language models

LLM	Accuracy Findings				
	Primary Diagnosis	Differential Diagnosis	Additional Examinations	Treatment Recommendations	Image-Based Diagnosis
ChatGPT-3.5	46–89%	90%	10%	60%	N/A
ChatGPT-4/4o	28.5–82.3%	28.3–90%	12.5–29%	16.7–60%	22.5–89.2%
Claude-3/3.5	54.3–88.6%	Higher than ChatGPT/Gemini	N/A	N/A	Higher than ChatGPT/Gemini
Google Bard	82%	N/A	N/A	N/A	N/A
Google Gemini	28.6–71.4%	Lower than Claude/ChatGPT	N/A	Lower than Claude/ChatGPT	N/A
Llama-2	N/A	N/A	N/A	15% (vs. 50% for GPT-4)	N/A
Bing-GPT4	74%	N/A	N/A	N/A	N/A

Abbreviations: LLM = large language model; NA = not available

Table 3 Bias analysis

Studies	Clearly Stated Aim	Inclusion of consecutive patients	Prospective data collection	Endpoints appropriate quality	Unbiased endpoint assessment	Study size prospective calculation	< 5% of lost of follow-up	Follow-up Output stability	Total MINOR (/16)
Warrier [13]	2	0	1	1	1	0	-	0	5
Tomo [14]	2	0	1	1	2	0	-	2	8
Schmidl [15]	2	2	2	2	2	0	-	0	10
Lorenzi [16]	2	0	1	2	2	0	-	0	7
Lechien [4]	2	1	2	2	2	0	-	0	9
Dronkers [17]	2	0	1	1	1	0	-	0	5
Lechien [18]	2	2	2	2	2	0	-	2	12
Lechien [19]	2	1	2	2	2	0	-	0	9
Radulesco [20]	2	2	2	2	2	0	-	2	12
Lechien [21]	2	2	2	2	2	0	-	2	12
Sievert [22]	2	0	2	1	2	0	-	2	9
Schmidl [9]	2	0	2	1	2	0	-	0	7
Alami [23]	2	1	0	1	0	0	-	0	4
Maniaci [8]	2	2	2	2	2	0	-	2	12
Maniaci [24]	2	0	1	2	2	0	-	2	9
Noda [25]	2	2	1	1	2	0	-	0	8
Lechien [7]	2	0	2	2	2	0	-	0	8

Table 4 Summary of large Language model performance

Clinical function	Performance summary	Implications for practice
Primary diagnosis (text-based)	Moderate to high accuracy (45–80%) across ChatGPT and Claude [4, 7, 14, 15]	May be used as a second opinion generator in common and rare conditions.
Differential diagnosis (text-based)	High accuracy in head and neck oncology and general ORL (up to 90%) [7, 15, 19, 21]	Valuable adjunct for refining differential diagnoses in complex or overlapping presentations.
Image interpretation (multimodal LLMs)	Highly variable. High with context (up to 93.3%) [9, 22, 25], poor without context [8]	Effective only when paired with structured clinical data; not standalone.
Recommendation of additional investigations	Low reliability (10–29%) in all subspecialties [4, 8, 18–21, 24].	LLMs tend to over-request. Clinical oversight is essential.
Therapeutic recommendation	Variable. Stronger in oncology (55–60%) [15, 18], weak in laryngology and rhinology [17, 20]	May assist in guideline adherence in oncology; less reliable in functional or surgical cases.
Rare disease identification	Claude-3.5 > ChatGPT-4o (54% vs. 45.7%) [7]	Promising tool in identifying overlooked conditions; requires expert verification.

Bias analysis

The mean MINORS was 8.6 ± 2.5 (Table 3). Heterogeneity among included articles in LLM prompts, inclusion/exclusion clinical case criteria, and accuracy outcomes precluded statistically pooling the data into a formal meta-analysis, thereby limiting the analysis to a qualitative rather than quantitative summary of the available information. There was substantial heterogeneity across studies for the inclusion of consecutive clinical cases from the consultation, and the prospective collected data inclusion in the LLM's interface. The LLM's responses were appropriately analyzed in all studies with at least two independent judges (Table 3). There were no investigations with sample size prospective calculation. The LLM's responses were regenerated at least one time in 8 studies, all of them reporting moderate-to-high stability of outputs [8, 14, 18, 20–22, 24].

Discussion

The increasing clinical integration of LLMs underscores the need for rigorous performance assessment in otolaryngology [26]. A comprehensive state-of-the-art review published in September 2024 examining the use of ChatGPT in otolaryngology identified only five studies that objectively evaluated the diagnostic accuracy in real otolaryngologic case series [6]. As of March 2025, the present systematic review reports findings from 17 publications, representing a substantial increase in peer-reviewed literature on this subject within a six-month period.

The present discussion primarily addresses ChatGPT performance metrics, as this platform represents the most extensively validated LLM in the contemporary medical literature. Although it remains the most widely evaluated and validated LLM in the otolaryngologic literature, recent investigations have revealed that alternative models such as Claude-3.0 and 3.5 may exhibit superior diagnostic reasoning in specific clinical contexts. For instance, Claude-3.5 outperformed ChatGPT-4 in the accurate identification of rare otolaryngological conditions and demonstrated higher consistency in formulating guideline-concordant diagnostic workups in head and neck oncology [7,15.] Thus, these results may reflect a more refined handling of clinical subtleties and a better integration of contextual cues in Claude's architecture. Conversely, Gemini's performance was more variable and often lagged both Claude and ChatGPT in diagnostic accuracy and therapeutic relevance [7,16.] However, its potential for rapid evolution may position it as a competitive model in future iterations. While the current evidence base for Claude and Gemini remains limited relative to that of ChatGPT, these early comparative findings underscore the importance of broadening LLM evaluation.

The findings of this systematic review support that ChatGPT exhibits varying degrees of accuracy in diagnosis-making, additional examination selection, and treatment recommendations. ChatGPT exhibited highest accuracy in proposing correct primary or differential diagnoses, followed by indicating the most appropriate treatments and additional examinations. While most studies reported moderate-to-high stability of regenerated outputs [14, 20, 22], the accuracy of ChatGPT's responses appeared to be influenced by several interconnected factors, including the subspecialty of clinical cases, disease rarity (case reports *versus* common conditions), the inclusion of images, and the type of images (pathology, imaging, clinical) [8, 14, 18, 20, 21, 25]. Radulesco et al. evaluated the performance of ChatGPT-4 in the clinical management of 40 rhinological cases [20]. In this study, ChatGPT-4 proposed a correct diagnosis in 50.8% of cases, which was lower than the primary diagnosis accuracy rates of studies using standardized protocols in general otolaryngology [21], otology [25], head and neck [18], and maxillofacial surgery [14]. In laryngology, Maniaci et al. similarly observed low diagnostic accuracy [8], which was particularly related to the ChatGPT's capability to recognize laryngeal lesions in the uploaded images. Specifically, their findings revealed that ChatGPT-4 exhibited limited capability in analyzing clinical laryngeal images [8]. The observation of ChatGPT-4's lowest accuracy when interpreting laryngological clinical images was not corroborated in maxillofacial and oncological subspecialties, where the model achieved superior diagnostic concordance rates: 80.2% for maxillofacial clinical images (surpassing ChatGPT-3.5's 64.9%) [14], and 71.2% when analyzing histopathological oncological specimens [22]. Importantly, Schmidl et al. demonstrated that the accuracy of ChatGPT-4 in analyzing clinical images was significantly influenced by the inclusion of medical information, with highest accuracy achieved when clinical images are inputted with patient's clinical information (73.3–93.3%) compared to when they are presented without context (26–86.7%) [9].

The integration of image-based reasoning into LLMs, as exemplified by GPT-4 Vision, represents a significant evolution in AI-assisted clinical decision-making. While multimodal configurations theoretically enhance diagnostic capability by enabling visual pattern recognition, their actual performance remains highly context dependent. Schmidl et al. [9]. demonstrated that diagnostic accuracy for oncological lesions markedly improved when clinical images were paired with structured patient data, rising from 26% with isolated image interpretation to 93.3% when contextual clinical information was included. Similar findings were reported in otologic disease classification using GPT-4 Vision, where high performance was contingent on multimodal input structures [25]. Conversely, when deprived of clinical context, LLMs frequently failed to extract relevant features or prioritize differentials appropriately, even more so in laryngological applications [8].

The present systematic review suggests similar variability in ChatGPT's accuracy for proposing appropriate treatments in the management of real clinical cases. In the therapeutic domain, ChatGPT-4 exhibited superior treatment recommendation concordance for oncological cases, demonstrating statistically significant adherence to both multidisciplinary tumor board consensus protocols and international evidence-based guidelines for head and neck cancer management [15, 18, 23]. Although the majority of studies investigated only ChatGPT, emerging evidence suggests Claude's superior performance in specific clinical applications, including the recommendations of guideline-concordant oncological diagnostic workups [15], and the establishment of accurate primary and differential diagnoses across both common and very rare otolaryngological disorders [7].

The lack of comparison of ChatGPT performance with other LLMs is the primary limitation of the current research [3]. In the otolaryngological literature, comparative analyses between ChatGPT and alternative large language models (e.g., Google Bard, Claude, Gemini) remain limited [3], with a few existing studies evaluating performance across multiple clinical domains including patient information, clinical decision support, general knowledge related to diseases.

The rapid advancement in LLM capabilities and emerging comparative efficacy data support the need to conduct prospective comparative studies evaluating LLM diagnostic concordance, therapeutic recommendation accuracy, and clinical workflow integration using standardized clinical vignettes from diverse subspecialties. Finally, the absence of a quantitative meta-analysis was related to the important heterogeneity across studies for input/prompt forms, subspecialties, evaluation metrics, and diagnostic reference standards making statistical aggregation inappropriate and potentially misleading. In accordance with best practices in diagnostic accuracy research and international guidance such as PRISMA, a structured qualitative synthesis was deliberately chosen to preserve analytical integrity and ensure a contextually valid interpretation of findings. Future studies with standardized protocols could address the heterogeneities highlighted in the present review, which constitute another primary limitation, restricting the ability to draw valid conclusions.

To facilitate clinical translation, we summarized the specific domains where current LLMs demonstrate robust performance versus where their application remains limited.

Conclusion

Large language models, especially ChatGPT, demonstrate promising moderate diagnostic accuracy in otolaryngology clinical decision support, with higher performance in providing diagnoses than in suggesting appropriate additional examinations and treatments. The inclusion of clinical images, the rarity of cases, and the subspecialty could influence the performance of LLMs. Although emerging findings support that Claude often outperforms ChatGPT, the lack of comparison between ChatGPT performance and other LLMs is the main limitation of the current research. Methodological standardization is needed for future research.

Acknowledgements KC, Librarian, for the literature review conduction.

Author contributions Filali Ansary R.: data interpretation, revising the manuscript for important intellectual content; final approval of the version to be published, final approval, and accountability for the work; 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. Lechien JR: data interpretation, revising the manuscript for important intellectual content; final approval of the version to be published, final approval, and accountability for the work; 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.

Funding None.

Data availability Data can be available on request.

Declarations

Ethic committee Not required.

Informed consent Not required.

Competing interests The author Jerome R. Lechien was not involved with the peer review process of this article.

References

1. Lechien JR, Gorton A, Robertson J, Vaira LA (2023) Is ChatGPT-4 accurate in proofread a manuscript in Otolaryngology-Head and neck surgery? Otolaryngol Head Neck Surg. https://doi.org/10.1002/ohn.526
2. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, Cotofana S, Alfertshofer M (2023) ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification Preparation questions. Eur Arch Otorhinolaryngol 280(9):4271–4278. https://doi.org/10.1007/s00405-023-08051-4
3. Lechien JR, Rameau A (2024) Applications of ChatGPT in Otolaryngology-Head neck surgery: A state of the Art review. Otolaryngol Head Neck Surg 171(3):667–677. https://doi.org/10.1002/ohn.807
4. Lechien JR, Naunheim MR, Maniaci A, Radulesco T, Saibene AM, Chiesa-Estomba CM, Vaira LA (2024) Performance and consistency of ChatGPT-4 versus otolaryngologists: A clinical case series. Otolaryngol Head Neck Surg 170(6):1519–1526. https://doi.org/10.1002/ohn.759
5. Naddaf M (2025) How are researchers using AI? Survey reveals pros and cons for science. Nat Feb 4, https://doi.org/10.1038/d41586-025-00343-5
6. Lechien JR et al (2025) Artificial Intelligence-Assisted diagnosis of an unusual cause of periodic epistaxis: A case report. Ear Nose Throat J. https://doi.org/10.1177/01455613251335385

7. Lechien JR, Maniaci A (2025) October,. Large Language Models as Adjunctive Tools for Diagnosing Rare Diseases in Otolaryngology: A Controlled Study. Accepted for oral communication, American Academy of Otolaryngology Head Neck Surgery Annual Meeting, Indianapolis
8. Maniaci A, Chiesa-Estomba CM, Lechien JR (2024) ChatGPT-4 consistency in interpreting laryngeal clinical images of common lesions and disorders. Otolaryngol Head Neck Surg 171(4):1106–1113. https://doi.org/10.1002/ohn.897
9. Schmidl B, Hütten T, Pigorsch S, Stögbauer F, Hoch CC, Hussain T, Wollenberg B, Wirth M (2025) Artificial intelligence for image recognition in diagnosing oral and oropharyngeal cancer and leukoplakia. Sci Rep 15(1):3625. https://doi.org/10.1038/s41598-025-85920-4
10. Thompson M, Tiwari A, Fu R, Moe E, Buckley DI (2012) A Frame- work to facilitate the use of systematic reviews and Meta-Analyses in the design of primary research studies. Agency for Healthcare Research and Quality
11. McInnes MDF, Moher D, Thombs BD et al (2018) Preferred reporting items for a systematic review and Meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 319(4):388–396
12. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J (2003) Methodological index for non-randomized studies (minors): development and validation of a new instrument. ANZ J Surg 73(9):712–716. https://doi.org/10.1046/j.1445-2197.2003.02748.x
13. Warrier A, Singh R, Haleem A, Zaki H, Eloy JA (2024) The comparative diagnostic capability of large Language models in otolaryngology. Laryngoscope 134(9):3997–4002. https://doi.org/10.1002/lary.31434
14. Tomo S, Lechien JR, Bueno HS, Cantieri-Debortoli DF, Simonato LE (2024) Accuracy and consistency of ChatGPT-3.5 and– 4 in providing differential diagnoses in oral and maxillofacial diseases: a comparative diagnostic performance analysis. Clin Oral Investig 28(10):544. https://doi.org/10.1007/s00784-024-05939-1
15. Schmidl B, Hütten T, Pigorsch S, Stögbauer F, Hoch CC, Hussain T, Wollenberg B, Wirth M (2024) Assessing the use of the novel tool Claude 3 in comparison to ChatGPT 4.0 as an artificial intelligence tool in the diagnosis and therapy of primary head and neck cancer cases. Eur Arch Otorhinolaryngol 281(11):6099–6109. https://doi.org/10.1007/s00405-024-08828-1
16. Lorenzi A, Pugliese G, Maniaci A, Lechien JR, Allevi F, Boscolo-Rizzo P, Vaira LA, Saibene AM (2024) Reliability of large Language models for advanced head and neck malignancies management: a comparison between ChatGPT 4 and gemini advanced. Eur Arch Otorhinolaryngol 281(9):5001–5006. https://doi.org/10.1007/s00405-024-08746-2
17. Dronkers EAC, Geneid A, Al Yaghchi C, Lechien JR (2024) Evaluating the Potential of AI Chatbots in Treatment Decision-making for Acquired Bilateral Vocal Fold Paralysis in Adults. J Voice. Apr 6:S0892-1997(24)00059– 6. https://doi.org/10.1016/j.jvoice.2024.02.020
18. Lechien JR, Chiesa-Estomba CM, Baudouin R, Hans S (2024) Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. Eur Arch Otorhinolaryngol 281(4):2105–2114. https://doi.org/10.1007/s00405-023-08326-w
19. Lechien JR, Georgescu BM, Hans S, Chiesa-Estomba CM (2024) ChatGPT performance in laryngology and head and neck surgery: a clinical case-series. Eur Arch Otorhinolaryngol 281(1):319–333. https://doi.org/10.1007/s00405-023-08282-5
20. Radulesco T, Saibene AM, Michel J, Vaira LA, Lechien JR (2024) ChatGPT-4 performance in rhinology: A clinical case series. Int Forum Allergy Rhinol 14(6):1123–1130. https://doi.org/10.1002/alr.23323
21. Lechien JR, Maniaci A, Gengler I, Hans S, Chiesa-Estomba CM, Vaira LA (2024) Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the artificial intelligence performance instrument (AIPI). Eur Arch Otorhinolaryngol 281(4):2063–2079. https://doi.org/10.1007/s00405-023-08219-y
22. Sievert M, Aubreville M, Mueller SK, Eckstein M, Breininger K, Iro H, Goncalves M (2024) Diagnosis of malignancy in oropharyngeal confocal laser endomicroscopy using GPT 4.0 with vision. Eur Arch Otorhinolaryngol 281(4):2115–2122. https://doi.org/10.1007/s00405-024-08476-5
23. Alami K, Willemse E, Quiriny M, Lipski S, Laurent C, Donquier V, Digonnet A (2024) Evaluation of ChatGPT-4's performance in therapeutic Decision-Making during multidisciplinary oncology meetings for head and neck squamous cell carcinoma. Cureus 16(9):e68808. https://doi.org/10.7759/cureus.68808
24. Maniaci A, Lazzeroni M, Cozzi A, Fraccaroli F, Gaffuri M, Chiesa-Estomba C, Capaccio P (2024) Can chatbots enhance the management of pediatric sialadenitis in clinicalpractice? Eur Arch Otorhinolaryngol 281(11):6133–6140. https://doi.org/10.1007/s00405-024-08798-4
25. Noda M, Yoshimura H, Okubo T, Koshu R, Uchiyama Y, Nomura A, Ito M, Takumi Y (2024) Feasibility of multimodal artificial intelligence using GPT-4 vision for the classification of middle ear disease: qualitative study and validation. JMIR AI 3:e58342. https://doi.org/10.2196/58342
26. Demir E, Uğurlu BN, Uğurlu GA, Aydoğdu G (2025 Feb) Artificial intelligence in otorhinolaryngology: current trends and application areas. Eur Arch Otorhinolaryngol 17. https://doi.org/10.1007/s00405-025-09272-5