American Journal of Otolaryngology and Head and Neck Surgery

Research Article
Published: 06 Nov, 2025

Evaluating Five AI-Powered Language Models as Otolaryngology Clinical Support Tools in Rural Kenya

Lise Sogalow[1], Louis Victoor[1], Mohamad Khalife[2] and Jérôme R. Lechien[1-5]*

[1]Department of Surgery, UMONS Research Institute for Health Sciences and Technology, Mons, Belgium

[2]Department of Otolaryngology-Head and Neck Surgery, EpiCURA Hospital, Baudour, Belgium

[3]Department of Otorhinolaryngology and Head and Neck Surgery, Foch Hospital, Paris Saclay University, Paris, France

[4]Department of Otorhinolaryngology and Head and Neck Surgery, CHU Saint-Pierre, Brussels, Belgium

[5]Department of Otorhinolaryngology and Head and Neck Surgery, Elsan Poitiers Hospital, Poitiers, France

Abstract

Objectives: The accuracy of Large Language Models (LLMs) as adjunctive clinical tools has been investigated in numerous studies including real otolaryngology cases recruited in Western country populations. The aim of this study was to evaluate the accuracy of five AI-powered LLMs available on smartphones in the management of real clinical cases in Sub-Saharan resource-limited settings.

Methods: Demographic and clinical data of patients consulting in Iten County Hospital for otolaryngological conditions were prospectively recruited from December 2024 to January 2025. Anonymized data and primary clinical examination findings were entered into the APIs of ChatGPT-4o, Gemini-2.0-Flash, Claude-Sonnet-3.7, DeepSeek-R1, and Mistral-Large2 for clinical patient management. Two practitioners independently assessed LLM recommendations using the Artificial Intelligence Performance Instrument (AIPI). The Intraclass Correlation Coefficient (ICC) was used to measure the interrater agreement.

Results: Sixty-three patients were included (41.3% female; mean age: 24.2 ± 25.2 years). ChatGPT-4o achieved the highest total AIPI score (12.1 ± 2.3; p=0.036), outperforming other LLMs across differential diagnosis (2.3 ± 0.5; p=0.004), management plan (p=0.003), and diagnosis (5.4 ± 1.1; p=0.037). ChatGPT-4o and Claude-Sonnet-3.7 (2.0 ± 0.7 and 2.0 ± 0.5; p=0.053) reported the higher scores for treatment plan compared to other LLMs. The performance of LLMs for differential diagnoses was low-to-moderate with ChatGPT-4o having the highest correct rate of differential diagnoses (25.4%; p=0.008). Management responses were most accurate with Gemini-2.0-Flash (49.2%) and Claude-Sonnet-3.7 (39.7%; p=0.002). Interrater reliability was excellent with an ICC>0.8 for each LLM.

Conclusion: Large Language Models demonstrate promising clinical performance in Sub-Saharan otolaryngology settings affected by a shortage of otolaryngologists. ChatGPT-4o consistently outperformed other models across key diagnostic and management tasks.

Level of Evidence: IV.

Keywords: Large language models; ChatGPT; Claude; Sub-Saharan; Africa; Artificial intelligence; Otolaryngology; Head neck surgery; Accuracy

OPEN ACCESS

Correspondence:
Lise Sogalow, Department of Surgery, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Avenue du Champ de Mars, 6, B7000 Mons, Belgium, Tel: +32 65 37 22 53;
E-mail: lise.sogalow@umons.ac.be

Received Date: *11 Oct 2025*
Accepted Date: *03 Nov 2025*
Published Date: *06 Nov 2025*

Citation:
Sogalow L, Victoor L, Khalife M, Lechien JR. Evaluating Five AI-Powered Language Models as Otolaryngology Clinical Support Tools in Rural Kenya. Am J Otolaryngol Head Neck Surg. 2025; 8(1): 1266.

Introduction

The use of Artificial Intelligence-Powered Large Language Models (LLMs) is emerging in medicine, including otolaryngology, with widespread accessibility to patient and practitioner populations with internet access [1] The accessibility and popularity of LLMs have encouraged conducting studies, assessing their accuracy, stability, and consistency with practitioner decisions in the management of common or complex clinical cases in otolaryngology-head and neck surgery [2-5]. To date, studies evaluating LLM accuracy as an adjunctive clinical tool focused on real clinical cases recruited in otolaryngology departments of Western country setting with full access to technology-based additional examinations [4,5]. A recent preliminary study suggested that AI-powered LLMs may serve as valuable adjunctive clinical tools in Kenya, and more broadly low- and middle-income countries [6] where otolaryngologist shortages and limited healthcare resources

are key barriers to delivering healthcare to rural populations [7,8]. Despite limited access to healthcare resources, the population, including practitioners, has access to substantial telecommunications infrastructure with up to 80% of Sub-Saharan inhabitants having smartphones and access to the Internet [8].

This study evaluated the accuracy of five widely used and smartphone available LLMs as adjunctive tools for diagnosis, management plan, and treatment recommendations in a resource-limited rural Kenyan settings without permanent otolaryngologist practitioners.

Materials and Methods

Patients and Settings

Demographic and clinical data of 63 consecutive patients recruited between December 14, 2024, to January 3, 2025 were prospectively collected during a humanitarian outreach program in Kenya (Iten Referral County Hospital, Iten). Data were collected during consultations performed by a senior board-certified otolaryngologist (JRL). The Iten Referral County Hospital provides primary care services for patients with otolaryngological conditions, with limited materials and infrastructure, lacking basic equipment such as nasofibroscopy, ear microscopes, tympanometry device, or otoacoustic devices. For each patient, the following information was collected: medical and surgical history; symptoms; clinical examination; treatment; and access to medication. Patients included were those who came for ear, nose, and throat consultation and agreed to participate to the study. For children, parental consent was obtained. The study adhered to the STROBE guidelines for observational studies [9]. The overall study design is illustrated in (Figure 1).

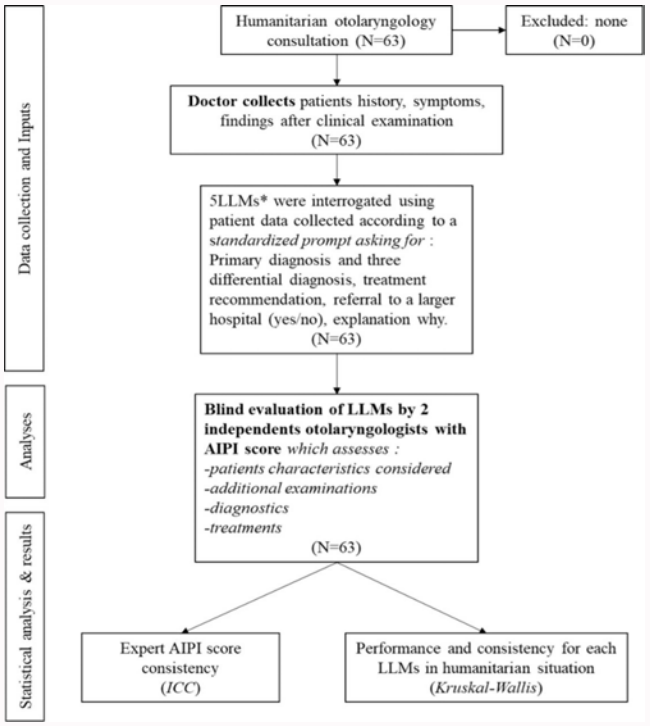

Figure 1: Chart Flow.
*ChatGPT-4o, Gemini-2.0-flash, Claude-Sonnet-3.7, Deep Seek-R and Mistral-Large-2.
Abbreviations: AIPI: Artificial Intelligence Performance Instrument; ICC: Intraclass Correlation Coefficient; LLM: Large Language Models.

The study was approved by the board of doctors of the Iten hospital (Dr. Mugala). However, because this hospital has no official IRB, the study was approved by the institutional review board of CHU Saint-Pierre to have an official IRB approval for including anonymized patient information into LLMs (CHUSP, n°BE0762023230708). The informed consent was obtained thanks to a local doctor who was in consultation with the primary practitioner. The humanitarian mission was supported by the Elsan Hospital Group (Elsan, Paris, France). The data of this study were used in a preliminary paper using Poe-based Q-Exp-Claude-3.5-Sonnet algorithm for supporting care delivery by primary care Kenyan practitioners [6].

Data Collection and Anonymization

The clinical information was entered through standardized sentences into the API of the five following LLMs: ChatGPT-4o (open AI, San Francisco, USA), Claude-Sonnet-3.7 (Anthropic, San Francisco, USA), Gemini-2.0-Flash (Google, New York, USA), DeepSeek-R1 (Together AI, Hangzhou, China) and Mistral-Large-2 (Mistral, France). Deep seek is the new Chinese LLM, while Mistral is the new French LLM, both being recent and their accuracies have never been tested in clinical setting. The data were anonymized, and the prompt included only the symptoms and the clinical findings, while specifying the local shortage of additional examinations and medications, reflecting the real humanitarian conditions. The following sentences were used: I am a doctor in a Humanitarian mission in an ENT department. I have a patient with the following symptoms. My examination without fiberoptic reported. I have limited medications, only the primary and most common medication in Rural Africa. What is your primary and three most prevalent differential diagnoses? What are your treatment recommendations? Do you believe that the patient needs to be addressed to a university hospital (yes/no?) and why.

The output information was collected in a structured database by an independent researcher (LV). LLM responses were independently assessed by two otolaryngologist practitioners using the Artificial Intelligence Performance Instrument (AIPI) [10]. The AIPI is a valid and reliable multi-criterion scoring tool for assessing the performance of LLMs in the management of real clinical cases. The characteristics evaluated are the patient's feature score (/6), the diagnosis score (/7), the additional examination score (/5), and the treatment score (/3). Each LLM's response received a total AIPI score (/20) ranging from 0 (inadequate management) to 20 (excellent management).

Statistical Analysis

Statistical analyses were performed using the Statistical Package for the Social Sciences for Windows (SPSS version 29.0; IBM Corp, Armonk, NY, USA). The AIPI scores were compared across LLMs using a Kruskal-Wallis test. An intraclass correlation coefficient was calculated between raters for assessing the interrater reliability. A 95% confidence interval was used and statistical significance was defined as $p<0.05$.

Results

Of the 63 recruited patients, there were 26 (41.3%) females and 37 (58.7%) males with a mean age of 24.2 ± 25.2 years (Table 1). The patient sample included 26 adults and 37 children. The distribution of otolaryngologist primary diagnoses is described in Table 1. The most prevalent diagnoses included 9 (14.3%) suspected allergic or nonallergic rhinitis, 8 (12.7%) adenoid/tonsil hypertrophy with potential sleep apnea syndrome, and 5 (7.9%) suspected

Table 1: Characteristics of Patients.

Demographics	Adults (N=26)	Children (N=37)	Total (N, %)
Age (mean, SD)	49.1 ± 18.7	5.1 ± 3.1	24.2 ± 25.2
Gender (N, %)			
Females	14 (53.8)	12 (32.4)	26 (41.3)
Males	12 (46.2)	25 (67.6)	37 (58.7)
Primary otolaryngologist's diagnoses			
Suspected allergic or nonallergic rhinitis	4 (15.4)	5 (13.5)	9 (14.3)
Adenoid/tonsil hypertrophy & apnea	0 (0)	8 (21.6)	8 (12.7)
Suspected LPRD (RSS>13)	4 (15.4)	1 (2.7)	5 (7.9)
Suspected GERD (Lyon consensus) with LPRD (RSS>13)	4 (15.4)	1 (2.7)	5 (7.9)
Chronic otitis media with cholesteatoma	3 (11.5)	2 (5.4)	5 (7.9)
Acute pharyngitis (virus, foreign body irritation)	3 (11.5)	1 (2.7)	4 (6.3)
Recurrent tonsillitis	0 (0)	4 (10.8)	4 (6.3)
Chronic otitis media with suppuration (no complication)	0 (0)	4 (10.8)	4 (6.3)
Adenoid hypertrophy with/without adenoiditis	0 (0)	3 (8.1)	3 (4.8)
Snoring without apnea related to nasal deviation	2 (7.7)	0 (0)	2 (3.2)
Epistaxis	1 (3.8)	1 (2.7)	2 (3.2)
Chronic adhesive otitis media	1 (3.8)	1 (2.7)	2 (3.2)
Thyroid goiter	1 (3.8)	0 (0)	1 (1.6)
Idiopathic dysphonia (suspicion of vocal cord cancer)	1 (3.8)	0 (0)	1 (1.6)
Laryngotracheitis (acute)	1 (3.8)	0 (0)	1 (1.6)
Eustachian tube dysfunction	1 (3.8)	0 (0)	1 (1.6)
Diabetes mellitus	1 (3.8)	0 (0)	1 (1.6)
Nasal foreign body	0 (0)	1 (2.7)	1 (1.6)
Complicated maxillary rhinosinusitis (acute)	0 (0)	1 (2.7)	1 (1.6)
Ear wax	0 (0)	1 (2.7)	1 (1.6)
External otitis (acute)	0 (0)	1 (2.7)	1 (1.6)
External ear duct eczema	0 (0)	1 (2.7)	1 (1.6)
Short tongue-ties	0 (0)	1 (2.7)	1 (1.6)

Abbreviations: AIPI: Artificial Intelligence Performance Instrument; GERD: Gastroesophageal Reflux Disease; LPRD: Laryngopharyngeal Reflux Disease; N: Number; RSS: Reflux Symptom Score; SD: Standard Deviation

Table 2: Clinical Performances of Large Language Models.

AIPI management outcomes	ChatGPT-4o	Gemini-2.0-Flash	Claude-3.7	DeepSeek-R1	Mistral	p-value
1. Consideration of medical history (/2)	1.81 ± 0.30	1.74 ± 0.37	1.82 ± 0.30	1.74 ± 0.35	1.73 ± 0.31	NS
2. Consideration of symptoms (/2)	1.80 ± 0.28	1.73 ± 0.31	1.79 ± 0.28	1.75 ± 0.32	1.74 ± 0.30	NS
3. Consideration of physical examination findings (/2)	1.09 ± 0.55	0.86 ± 0.61	1.00 ± 0.55	0.95 ± 0.55	0.95 ± 0.63	NS
Patient feature score (/6)	4.70 ± 0.90	4.33 ± 1.03	4.60 ± 0.82	4.44 ± 0.93	4.42 ± 0.96	NS
4. Differential diagnosis (/3)	2.32 ± 0.54	2.08 ± 0.48	2.09 ± 0.47	2.05 ± 0.46	2.04 ± 0.49	0.004
5. Primary diagnosis (/3)	2.40 ± 0.63	2.32 ± 0.56	2.44 ± 0.56	2.42 ± 0.51	2.22 ± 0.61	NS
6. Management plan (/1)	0.66 ± 0.23	0.75 ± 0.25	0.67 ± 0.24	0.59 ± 0.20	0.61 ± 0.23	0.003
Diagnosis score (/7)	5.38 ± 1.09	5.14 ± 1.03	5.20 ± 0.90	5.06 ± 0.78	4.87 ± 0.97	0.037
7. Additional examinations (/3)	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	NS
8. The most relevant additional examination (/1)	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	NS
Additional examination score (/4)	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	NS
9. Treatment (/3)	2.05 ± 0.72	1.82 ± 0.63	2.01 ± 0.49	1.96 ± 0.59	1.80 ± 0.68	0.053*
AIPI total score (/20)	12.13 ± 2.30	11.29 ± 2.32	11.81 ± 1.86	11.46 ± 1.90	11.10 ± 2.81	0.036

Abbreviations: AIPI: Artificial Intelligence Performance Instrument; NS: Non-Significant

Table 3: Proportions of Correct or Adequate Clinical Responses.

AIPI Outcomes	ChatGPT-4o N=63	Gemini-2.0 Flash N=63	Claude-3.7 N=63	DeepSeek_R1 N=6	Mistral N=63	p-value
Primary Diagnosis (N (%))						
Correct	27 (42.9)	20 (31.7)	26 (41.3)	22 (34.9)	14 (22.2)	NS
Plausible	13 (20.6)	12 (19.0)	15 (23.8)	17 (27.0)	21 (33.3)	
Not plausible	23 (36.5)	30 (47.6)	22 (34.9)	24 (38.1)	28 (44.4)	
Absent	0 (0)	1 (1.6)	0 (0)	0 (0)	0 (0)	
Differential diagnosis						
Correct	16 (25.4)	4 (6.3)	2 (3.2)	6 (9.5)	3 (4.8)	0.008
Plausible	38 (60.3)	41 (65.1)	49 (77.8)	45 (71.4)	45 (71.4)	
Not plausible	9 (14.3)	18 (28.6)	12 (19.0)	12 (19.0)	15 (23.8)	
Absent	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	
Relevant additional examination						
Pertinent and necessary	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	NS
Pertinent and not necessary	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	
Pertinent, necessary and inadequate	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	
Only inadequate examinations	63 (100)	63 (100)	63 (100)	63 (100)	63 (100)	
Treatment						
Pertinent and necessary	12 (19.0)	8 (12.7)	6 (9.5)	7 (11.1)	5 (7.9)	NS
Pertinent and incomplete	20 (31.7)	13 (20.6)	27 (42.9)	21 (33.3)	18 (28.6)	
Association of pertinent/necessary and inadequate	29 (46.0)	41 (65.1)	30 (47.6)	35 (55.6)	39 (61.9)	
No adequate strategy	2 (3.2)	1 (1.6)	0 (0)	0 (0)	1 (1.6)	
Management plant (0-1)						
Pertinent	20 (31.7)	31 (49.2)	25 (39.7)	12 (19.0)	15 (23.8)	0.002
Not pertinent	43 (68.3)	32 (50.8)	38 (60.3)	51 (81.0)	48 (76.2)	

Abbreviations: NS: Non-Significant.

laryngopharyngeal reflux disease (reflux symptom score>13).

Clinical Performances of Large Language Models

The clinical performances of LLMs showed that ChatGPT-4o had significantly higher AIPI results compared to other LLMs in various key characteristics (Table 2). The mean AIPI score of ChatGPT-4o (12.13 ± 2.30) was significantly higher than those of Gemini-2.0-Flash, Claude-Sonnet-3.7, DeepSeek-R1, and Mistral, ranging from 11.10 to 11.81, respectively. Gemini-2.0-Flash and Mistral reported the lowest AIPI scores.

Among the AIPI sub-categories, ChatGPT-4o outperformed the other LLMs for differential diagnoses (2.32 ± 0.54), with DeepSeek-R1 and Mistral reporting the lowest performances (2.05 and 2.04, respectively). The primary diagnosis score analysis showed that ChatGPT-4o and Claude-Sonnet-3.7 reported significantly higher AIPI sub scores compared to other LLMs (p=0.037), with the lowest score being reported for Mistral. The LLM management plan scores, assessed in the context of additional examination and medication shortages, revealed that Gemini-2.0-Flash outperformed other LLMs (Table 2). The five LLMs reported similar therapeutic option scores (p=0.053).

Proportion Analyses

The analysis of adequate response proportion across LLMs is available in Table 3. All LLMs did not consider the limitation of additional examinations in the humanitarian outreach context, proposing some unavailable examinations (Table 3). Despite a moderate primary diagnosis score, the percentage of consistent diagnoses with the otolaryngologist was low, reaching 25.4% of cases for ChatGPT-4o. However, combining plausible and correct primary diagnoses, ChatGPT-4o and Claude-Sonnet-3.7 reported 63.5% and 65.1% of plausible or correct primary diagnoses, respectively (Table 3). The proportion of correct or adequate clinical responses was calculated between the two evaluators, demonstrating that ChatGPT-4o had the highest number of correct differential diagnoses: 16 (25.4%), followed by DeepSeek-R1 with 6 (9.5%; p=0.008). Deep seek and Mistral reported the lowest proportions of adequate management plans according to judges (p=0.002; Table 3).

Table 4: Interrater Reliability.

	95% CI		
ICC Outcomes	Cronbach	Minimum	Maximum
AIPI			
ChatGPT-4o	0.841	0.776	0.893
Gemini-2.0-Flash	0.858	0.801	0.905
Claude-Sonnet-3.7	0.835	0.769	0.89
DeepSeek-R1	0.836	0.77	0.89
Mistral	0.839	0.774	0.892

Abbreviations: AIPI: Artificial Intelligence Performance Instrument; CI: Confidence Interval, ICC: Intraclass Correlation Coefficient

The ICC ranged from 0.835 to 0.858, supporting adequate interrater reliability for the AIPI evaluations (Table 4).

Discussion

Since the launch of ChatGPT in November 2022, the number

of studies investigating the usefulness and performance of LLMs as adjunctive tools in clinical settings is increasing, with most studies evaluating ChatGPT performance in common or rare cases of Western country otolaryngology consultations [1,2,5,9,11-17] This study is the first investigation exploring the usefulness and accuracy of five commonly used LLMs for common ear, nose, and throat disorders found in a clinical setting where otolaryngological care is not provided by an otolaryngologist.

Consistently with our recent preliminary report,6 our results support that some LLMs, particularly ChatGPT-4o and Claude-Sonnet-3.7, may be interesting adjunctive clinical tools in Sub-Saharan regions, reaching 63.5% and 65.1% of plausible or correct primary diagnoses for ChatGPT-4o and Claude-Sonnet-3.7, respectively. A recent systematic review investigating the accuracy of LLMs for real clinical cases reported that primary diagnostic accuracy ranged from 45.7% to 80.2% across different LLMs, with Claude often outperforming ChatGPT [17].

The proportions of accurate and plausible primary diagnoses found in the present paper corroborate the findings of the literature.

Regarding differential diagnosis, our findings suggested superior performance for ChatGPT-4o over other LLMs, with 85.7% of correct and plausible answers. Importantly, the gap between ChatGPT-4o and other LLMs was much higher when considering only correct differential diagnoses (25.4% *vs.* 3.2% to 9.5%), which did not corroborate the literature examining the performance of several LLMs for differential diagnoses [17] Indeed, the differential diagnosis accuracy of ChatGPT-3.5 and 4.0 ranged from 28.3% to 90% in studies using AIPI, with laryngology conditions reporting the lowest accuracy (28.3%) [11,12,13,17] Moreover, it has been found that Claude-3.5-Sonnet reported significantly higher accuracy rates for the differential diagnosis of rare diseases compared to ChatGPT-4o.13 The performance of ChatGPT-4o appears therefore optimal in a context of humanitarian outreach without material and medical resources.

The management plan is a key point to consider in Sub-Saharan regions without facility access. In this study, the management plan evaluation included the need to address patients to a University Hospital versus the possibility to treat them in the rural setting. In most cases the LLMs suggested the referral of numerous patients that were unnecessary. Although recent studies suggested low performance of Gemini in the management of real clinical cases in otolaryngology [13,17], the present study found that Gemini-2.0-Flash and Claude-Sonnet-3.7 reached the highest rates of adequate management plans (49.2% and 39.7%) in a context of humanitarian outreach. Both LLMs reported highest consistency with otolaryngologists in referring patients to the University hospital. The high performance of Claude corroborates the literature showing its superiority to other LLMs for proposing a realistic management plan for otolaryngological cases [13]. However, the results of Gemini-2.0-Flash were unexpected regarding the recent studies comparing Gemini to other LLMs, demonstrating its inferiority to ChatGPT [13,15].

The high rate of referral to university hospitals corroborates the high rates for recommending additional examinations, which are however not available in rural settings (e.g magnetic resonance, impedance-pH testing, scintigraphy). In this study, both judges observed that LLMs did not consider the shortage of technologies in their outputs, explaining the rates (0%) of adequate additional examination recommendations (Table 3). This observation may suggest that LLMs recommend investigating each potential differential diagnosis, whereas a practitioner prioritizes a more rational diagnosis [2].

The assessment of LLM usefulness in humanitarian outreach settings and the consideration of five different LLMs available on smartphones are the primary strengths of this study. This study was required to confirm preliminary results [6], suggesting that AI-powered LLMs may be an adjunctive clinical tool in Sub-Saharan settings without permanent otolaryngologists. Contrary to this preliminary study, the LLM accuracy was evaluated by two independent practitioners who reported high ICC, while 5 versus 1 LLMs have been evaluated. To date, there is no study assessing the accuracy of Deep seek and Mistral in the otolaryngology literature, which is an additional originality of the present paper. The Chinese LLM was launched in July 2023. The French Mistral LLM, which was considered as the first European LLM, was launched in June 2023. Their young age may support their lowest results, while future studies are needed to investigate the accuracy of future updated versions.

The study was conducted in rural low-income countries, having limited access to medical materials, possibilities of additional examinations and, consequently, confirmation of all diagnoses. This point is the primary limitation of the present study, while the clinical management of these cases, including the primary diagnosis confirmation, was carried out by a single board-certified otolaryngologist with a human-related risk of errors. The low number of patients is an additional limitation, which can be addressed in the continuation of this study in future outreaches.

Conclusion

Large language models, including ChatGPT-4o and Claude-Sonnet-3.7, may be effective adjunctive clinical tools in humanitarian otolaryngology outreach, achieving 63.5% and 65.1% of plausible or correct primary diagnoses respectively. ChatGPT-4o showed the highest overall performance, particularly in establishing diagnoses. However, all LLMs failed to adequately consider the resource limitations in rural settings, commonly recommending unavailable examinations and unnecessary referrals.

References

1. Warrier A, Singh R, Haleem A, Zaki H, Eloy JA. The Comparative Diagnostic Capability of Large Language Models in Otolaryngology. Laryngoscope. 2024;134(9):3997-4002.

2. Lechien JR, Naunheim MR, Maniaci A, Radulesco T, Saibene AM, Chiesa-Estomba CM, et al. Performance and Consistency of ChatGPT-4 Versus Otolaryngologists: A Clinical Case Series. Otolaryngol Head Neck Surg. 2024;170(6):1519-26.

3. Lechien JR, Saxena S, Vaira LA, Hans S, Maniaci A. Artificial Intelligence-Assisted Diagnosis of an Unusual Cause of Periodic Epistaxis: A Case Report. Ear Nose Throat J. 2025:1455613251335385.

4. Darbari Kaul R, Zhong W, Liu S, Azemi G, Liang K, Zou E, Sacks PL, et al. Development of an Open-Source Algorithm for Automated Segmentation in Clinician-Led Paranasal Sinus Radiologic Research. Laryngoscope. 2025.

5. 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.

6. Lechien JR. Expanding the Capacity of General Practitioners in Sub-Saharan Africa with Artificial Intelligence. Otolaryngol Head Neck Surg. 2025;173(4):1024-27.

7. Mulwafu W, Fagan JJ, Mukara KB, Ibekwe TS - ENT Outreach in Africa: Rules of Engagement - OTO Open. 2018;2(2):2473974X18777220.

8. Jayawardena ADL, Kahue CN, Cummins SM, Netterville JL - Expanding the Capacity of Otolaryngologists in Kenya through Mobile Technology - OTO Open. 2018;2(1):2473974X18766824.

9. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med. 2007;147(8):573-7.

10. Lechien JR, Maniaci A, Gengler I, Hans S, Chiesa-Estomba CM, Vaira LA. Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the Artificial Intelligence Performance Instrument (AIPI). Eur Arch Otorhinolaryngol. 2024;281(4):2063-79.

11. Lechien JR, Georgescu BM, Hans S, Chiesa-Estomba CM. ChatGPT performance in laryngology and head and neck surgery: a clinical case-series. Eur Arch Otorhinolaryngol. 2024;281(1):319-33.

12. Maniaci A, Chiesa-Estomba CM, Lechien JR. ChatGPT-4 Consistency in Interpreting Laryngeal Clinical Images of Common Lesions and Disorders. Otolaryngol Head Neck Surg. 2024;171(4):1106-13.

13. Lechien JR, Maniaci A. Large Language Models as Adjunctive Tools for Diagnosing Rare Diseases in Otolaryngology: A Controlled Study. Otolaryngol Head Neck Surg, In press. 2025.

14. Schmidl B, Hütten T, Pigorsch S, Stögbauer F, Hoch CC, Hussain T, et al. Assessing the use of the novel tool Claude 3 in comparison toChatGPT4.0 as an artificial intelligence tool in the diagnosis and therapy of primary head and neck cancer cases. Eur Arch Otorhinolaryngol. 2024;281(11):6099-6109.

15. Lorenzi A, Pugliese G, Maniaci A, Lechien JR, Allevi F, Boscolo-Rizzo P, et al. Reliability of large language models for advanced head and neck malignancies management: a comparison between ChatGPT 4 and Gemini Advanced. Eur Arch Otorhinolaryngol. 2024;281(9):5001-06.

16. 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.

17. Filali Ansary R, Lechien JR. Clinical decision support using large language models in otolaryngology: a systematic review. Eur Arch Otorhinolaryngol. 2025;282(8):4325-34.