Accuracy of AI-Powered Large Language Models in the Analysis of Laryngeal Images☆

*,‖,[1]Camille F. Legrain, *,‖Lise Sogalow, *,†,‡,‖Antonino Maniaci, and *,‡,§,¶,‖Jérôme R. Lechien, *§*Mons,* ¶*Brussels, Belgium,* †*Enna, Italy, and* ‡‖*Paris, France*

SUMMARY: Objective. To investigate the performance of three artificial intelligence (AI)-powered large language models (LLMs) as adjunctive tools in analyzing clinical pictures of common phoniatric disorders.

Methods. Medical history, symptoms, and videolaryngostroboscopic images of adult outpatients with laryngological disorders consulting in laryngology were presented to ChatGPT-4o, DeepSeek, and Claude-3.7-Sonnet for primary and differential diagnoses, management, and treatment. The accuracy of LLMs was assessed by two blinded laryngologists using the artificial intelligence performance instrument (AIPI). Errors of LLMs were documented.

Results. Thirty-nine patients completed the evaluations [19 (48.7%) females]. The mean age was 55.4 ± 19.6 years. Of the 52 identified phoniatric disorders, glottic insufficiency (n = 8, 20.5%), laryngopharyngeal reflux disease (n = 7, 17.9%), Reinke edema (n = 6, 15.4%), and vocal fold nodules (n = 5, 12.8%) were the most prevalent conditions. The performances of LLMs were low, with correct primary and differential diagnoses ranging from 15.4% to 28.2%, and 15.4% to 25.6%, respectively. AIPI scores were comparable across LLMs, except for treatment recommendations, with DeepSeek demonstrating significantly lower scores than Claude-3.7 and ChatGPT-4o ($P = 0.019$). The judges reported a high intraclass consistency (ICC) in the AIPI assessment [ICC = 0.888; 95% confident interval (0.831–0.932)]. The conditions associated with the highest inaccurate diagnosis across LLMs were reflux (n = 21), glottic insufficiency (n = 20), Reinke edema (n = 19), and vocal fold cyst (n = 16). ChatGPT-4o reported more accurate scores for difficult clinical cases ($r_s = 0.410$; $P = 0.008$).

Conclusion. The performance of AI-powered LLMs is mild for providing accurate analysis of phoniatric cases based on clinical findings and laryngostroboscopic images. Future studies are needed when updated LLM versions will consider video analyses rather than only clinical images.

Keywords: ChatGPT–Artificial intelligence–Otolaryngology–Large Language Model.

INTRODUCTION

The development of artificial intelligence (AI)-powered large language models (LLMs) is emerging worldwide with easy and full access to populations and practitioners.[1,2] In otolaryngology-head and neck surgery, the accuracy of LLMs was investigated for supporting primary and differential diagnoses of common or very rare clinical cases,[3–5] for patient education,[5,6] in searching scientific references,[7,8] in proofreading manuscripts,[9] or as humanitarian outreach support.[10] A recent systematic review investigating the performance of LLMs as adjunctive clinical tools reported a primary diagnosis accuracy of ChatGPT and Claude-Sonnet ranging from 45.7% to 80.2% for general otolaryngology cases.[11] In laryngology and phoniatrics, only two studies investigated the accuracy of ChatGPT in the analysis of real general clinical cases, while the accuracy of other emergent LLMs, such as Claude-Sonnet and DeepSeek, the Chinese LLM, was never investigated.[12,13] The lack of studies investigating the capability of multiple LLMs for analyzing clinical laryngeal images may be attributed to the lack of possibility to upload images in some initial versions of LLMs; this issue currently being resolved.

The objective of this study was to investigate the performance of three LLMs, ChatGPT-4o, Claude-3.7-Sonnet, and DeepSeek in the analysis of clinical pictures of common laryngeal conditions.

METHODS

Patients and setting

Fifty patients consulting at the Division of Laryngology of the Gembloux Medical Center and CHU Saint-Pierre (Brussels, Belgium) for primary laryngeal symptoms were consecutively recruited from September 2024 to December 2024. The patient data (eg, demographics, history, symptoms, medication, physical examination) and the laryngostroboscopic pictures were prospectively collected by a researcher who retrospectively entered the anonymized data into the application programming interface of

Accepted for publication October 15, 2025.

☆ Funding: None.

From the *Department of Surgery, University of Mons, Mons, Belgium; †Department of Medicine and Surgery, Kore University, Enna, Italy; ‡Research Committee, Young Otolaryngologists of the International Federation of Otorhinolaryngological Societies (IFOS – Yo-IFOS), Paris, France; §Division of Laryngology and Broncho-esophagology, Department of Otolaryngology-Head Neck Surgery, EpiCURA Hospital, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium; ¶Department of Otorhinolaryngology and Head and Neck Surgery, CHU Saint-Pierre, Brussels, Belgium; and the ‖Research Committee of Young Otolaryngologists of the International Federation of Otorhinolaryngological Societies (IFOS), Paris, France

[1] Dr Legrain and Dr Sogalow have similarly contributed and are joined as co-first authors.

Address correspondence and reprint requests to Jérôme R. Lechien, Department of Surgery, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium. E-mail: Jerome.Lechien@umons.ac.be

Journal of Voice, Vol xx, No xx, pp. xxx–xxx

0892-1997

https://doi.org/10.1016/j.jvoice.2025.10.021

DeepSeek (DeepSeek, Hangzhou, China), Chatbot Generative Pre-trained Transformer (ChatGPT-4o; OpenAI, San Francisco, USA), and Claude-3.7-Sonnet (Anthropic, San Francisco, USA). Patients were included if the information related to the medical record, clinical examination, laryngostroboscopic images, and any additional examinations were fully available. Patients with incomplete data, without identified laryngeal disorders, or lacking laryngostroboscopic picture findings were excluded. Because the uploading of video is not allowed in some LLMs, the authors have included pictures from videolaryngostroboscopy in the present study. Note that in case of mobility disorders, such as vocal fold paralysis or posterior glottic stenosis, two images were provided with an explanation that both images consisted of adduction and abduction of the vocal folds.

Chatbots were systematically queried to provide an analysis of the clinical cases and the related videolaryngostroboscopic images using standardized questions (Figure 1). The primary researcher (CL) used the following standardized sentences after the description of each case: *What are your primary and differential diagnoses?*; *What is your analysis and disorder detected on the clinical laryngeal images?*; *What are your additional examinations to find the diagnosis (management plan)?*; *What are your treatment(s) for the primary diagnosis?* The responses of the LLMs were collected in a database. For each case, the laryngeal configuration included images with vocal folds in abduction and adduction. The study was approved by the institutional review board of CHU Saint-Pierre (CHUSP, n°BE0762023230708). Prospective recruited patients consented to participate. This study adhered to the STrengthening the Reporting of OBservational Studies in Epidemiology guidelines for observational studies to ensure transparency and replicability of our findings.[14]

Large language model consistency

The diagnosis of laryngeal conditions was performed prior to LLM analysis by two board-certified laryngologists, considering patient history, symptoms, videolaryngostroboscopy findings, and additional findings, including evaluations of voice quality, additional examinations, or histopathological findings. The responses from LLMs were then independently evaluated by two practitioners for consistency and performance analysis.

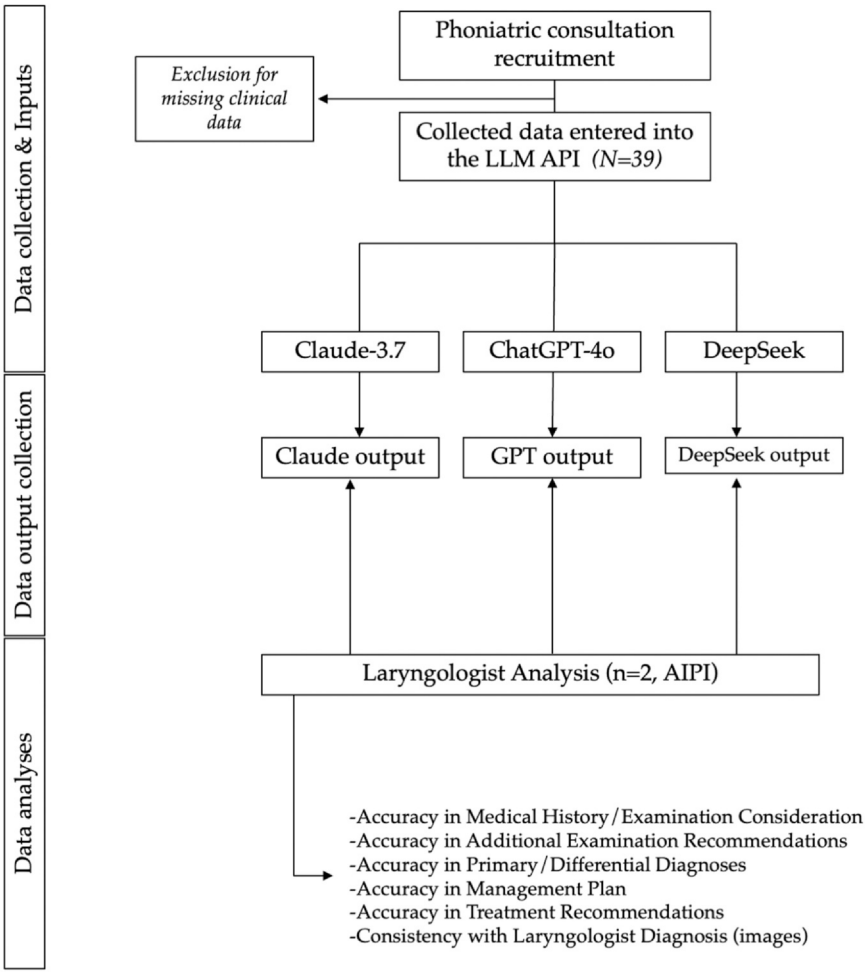

FIGURE 1. Chart flow. AIPI, artificial intelligence performance instrument; LLM, large language model.

The consistency of LLM analysis of clinical images was assessed by the laryngologists (agreement) using a 5-point Likert scale, ranging from 1 (very low consistency) to 5 (very high consistency).[12]

The performance of ChatGPT-4 in providing accurate primary and differential diagnoses, suggesting additional examinations, and recommending treatments was assessed using the Artificial Intelligence Performance Instrument (AIPI) (Figure 2).[15] AIPI is a 9-item validated instrument for the performance analysis of generative AI chatbots, covering medical and surgical history; symptoms; physical examination; diagnosis; additional examinations; management plans; and treatments in the overall management of real clinical cases. AIPI is subdivided into the following sub-scores, each associated with common items: patient feature score (/6), diagnosis score (/7), additional examination score (/5), and treatment score (/3). The final AIPI score ranges from 0 (inadequate management) to 20 (excellent management).[15] Errors in ChatGPT-4 responses were documented, and correct information was subsequently entered into the application programming interface as human feedback.

Statistical analyses

Statistical analyses were performed using the Statistical Package for the Social Sciences for Windows (*SPSS* version 29,0; IBM Corp, Armonk, NY, USA). To minimize selection bias, we used consecutive sampling, where every presenting patient meeting the inclusion criteria was invited to participate until the predetermined sample size was reached. The comparison of AIPI and 5-point Likert scale scores across LLMs was performed with the Kruskal-Wallis test regarding the Shapiro-Wilk results for variables ($P < 0.05$). The interrater reliability for the AIPI score assigned by the judges was evaluated with intraclass consistency (ICC). The number of additional examinations indicated by DeepSeek, ChatGPT-4o, Claude-3.7-Sonnet, and the laryngologist were compared with the Kruskal-Wallis test. A significance level of $P < 0.05$ was used.

RESULTS

The data of 39 patients were presented to DeepSeek, ChatGPT-4o, and Claude-3.7-Sonnet. Eleven patients were excluded because of incomplete clinical data. The mean age of patients was 55.4 ± 19.6 years. There were 19 (48.7%) females and 20 (51.3%) males, respectively (Table 1). Among the 52 identified disorders, laryngopharyngeal reflux disease (n = 7, 17.9%), Reinke's edema (n = 6, 15.4%), glottic insufficiency (n=5, 12.8%), and vocal fold nodules (n = 5, 12.8%) were the most common diagnoses. Some images inputted into the different LLMs are available in

Outcomes of Artificial Intelligence Performance Instrument (AIPI)	**Practitioner evaluation**			Item score	Subscores
1. Consideration of medical and surgical history in the AI management:	Fully (2)	Partly (1)	Not (0)	…../2	Patient feature score …...../6
2. Consideration of symptoms of patients in the AI management	Fully (2)	Partly (1)	Not (0)	…../2	
3. Consideration of physical findings reported by practitioner(s)	Fully (2)	Partly (1)	Not (0)	…../2	
4. The differential diagnoses provided by AI are:	Complete and plausible (3) Incomplete but plausible (2) Incomplete and not plausible for one or several (1) Absent (0)			…../3	Diagnosis score …...../7
5. The primary diagnosis of AI was:	Correct (3) Plausible (2) Not plausible (1) Absent (0)			…../3	
6. The management plan of AI included potential physical/additional examinations for determining the diagnosis		Yes (1)	No (0)	…../1	
7. The additional examinations proposed by AI are/include	All pertinent and necessary examinations (3) All pertinent but partialy necessary examinations (2) An association of pertinent, necessary, and inadequate examinations (1) An association of inadequate examinations (0)			…../3	Additional Examination Score …...../5
8. AI identified the most relevant additional examination to perform first	Yes (1) No, AI provided a list without stratification (0)			…../1	
9. The treatments proposed by AI are/include	All pertinent and necessary therapeutic findings (3) All pertinent but incomplete therapeutic findings (2) An association of pertinent, necessary, and inadequate therapeutic findings (1) No adequate therapeutic approach (0)			…../3	
				Total AIPI	……./20

FIGURE 2. Artificial intelligence performance instrument. AIPI is a 9-item validated instrument for the performance analysis of generative AI chatbots, covering medical and surgical history; symptoms; physical examination; diagnosis; additional examinations; management plans; and treatments in the overall management of real clinical cases. AIPI is subdivided into the following sub-scores, each associated with common items: patient feature score (/6), diagnosis score (/7), additional examination score (/5), and treatment score (/3). The final AIPI score ranges from 0 (inadequate management) to 20 (excellent management). AI, artificial intelligence.

TABLE 1.
Patient Features

Outcomes	Patients (N = 39)
Age (mean, SD)	55.4 ± 19.6
Gender (N, %)	
Female	19 (48.7)
Male	20 (51.3)
Primary and secondary diagnoses	
Glottic insufficiency/aging voice	8 (20.5)
Laryngopharyngeal reflux disease	7 (17.9)
Reinke edema	6 (15.4)
Vocal cord nodules	5 (12.8)
Vocal cord cyst	4 (10.3)
Vocal cord paralysis	3 (7.7)
Vocal cord atrophy	3 (7.7)
Vocal cord sulcus	3 (7.7)
Vocal cord scar	2 (5.1)
Synechia	2 (5.1)
Leukoplakia	2 (5.1)
Vocal cord hypomobility	1 (2.6)
Vocal cord paresis	1 (2.6)
Vocal cord polyp	1 (2.6)
Vocal cord hemorrhage	1 (2.6)
Glottic stenosis	1 (2.6)
Granuloma	1 (2.6)
Postintubation granuloma	1 (2.6)

Abbreviations: N, number; SD, standard deviation.

Figure 3. The mean complexity score was 2.4 ± 1.1. Regarding the laryngologist evaluations, 7 (17.9%), 14 (35.9%), 6 (15.4%), and 12 (30.8%) clinical cases were judged difficult (4/5), moderately difficult (3/5), slightly difficult (2/5), and not difficult (1/5). There was no clinical situation judged as very difficult (5/5).

Performance analysis

The mean AIPI scores of DeepSeek, ChatGPT-4o, and Claude-3.7-Sonnet are reported in Table 2. Regarding AIPI item and total scores, only the treatment score reported significant differences across LLMs, with DeepSeek demonstrating a significantly lower therapeutic recommendation score compared to ChatGPT-4o and Claude-3.7-Sonnet ($P = 0.019$). The mean AIPI scores of DeepSeek, Claude-3.7-Sonnet, and ChatGPT-4o ranged from 11.82 to 12.31, which consist of low scores. The detailed investigation of diagnosis and management performances of LLMs are reported in Table 3. The performances of LLMs were low, with correct primary and differential diagnoses ranging from 15.4% to 28.2%, and 15.4% to 25.6%, respectively. ChatGPT-4o and Claude-3.7-Sonnet recommended only relevant additional examinations in 5 (12.8%) and 1 (2.6%) of cases, respectively, while DeepSeek did not recommend relevant additional examinations ($P = 0.035$; Table 3). The mean numbers of indicated additional examinations per patient were significantly higher for DeepSeek (5.21), ChatGPT-4o (4.28), and Claude-3.7-Sonnet (6.28) compared with laryngologist (0.28; $P = 0.001$). According to current guidelines or recommendations, the management plan was consistent in 35.9% to 41.0% of cases, while treatments were judged pertinent and necessary in 2.6% to 23.1% of cases. The conditions associated with the highest inaccurate diagnosis (cumulative mistakes) across LLMs were laryngopharyngeal reflux disease (n = 21), glottic insufficiency (n = 20), Reinke's edema (n = 19), vocal fold cyst (n = 16), and vocal fold nodules

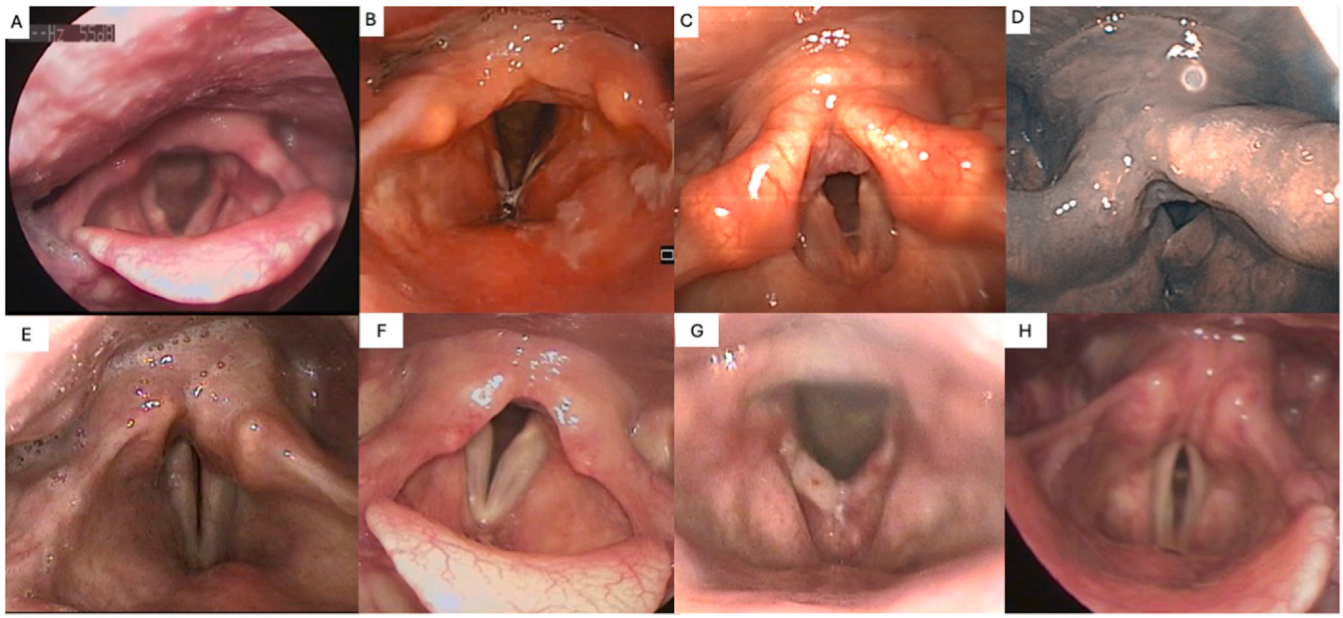

FIGURE 3. Some images entered into the large language model interface. Cyst (A), leukoplakia with ulcerative lesions (B), Reflux with posterior commissure granulation (C), Reinke edema (D), Bilateral vocal fold paralysis (image showed in adduction attempt (E), vocal fold sulcus (expiration time, F), anterior glottic synechia (G), bilateral vocal fold atrophy (H).

TABLE 2.
Accuracy Scores of Large Language Models

AIPI management outcomes	DeepSeek	ChaGPT-4o	Claude-3.7	*P* value
1. Consideration of medical history (/2)	1.59 ± 0.59	1.74 ± 0.59	1.54 ± 0.55	0.114
2. Consideration of symptoms (/2)	1.49 ± 0.60	1.82 ± 0.45	1.51 ± 0.60	0.004
3. Consideration of physical examination findings (/2)	1.46 ± 0.55	1.10 ± 0.91	1.41 ± 0.55	0.902
Patient feature score (/6)	4.54 ± 1.52	4.67 ± 1.36	4.46 ± 1.57	0.897
4. Differential diagnosis (/3)	1.87 ± 0.73	1.71 ± 0.89	1.90 ± 0.82	0.700
5. Primary diagnosis (/3)	1.56 ± 0.79	1.64 ± 0.96	1.59 ± 0.94	0.476
6. Management plan (/1)	0.36 ± 0.49	0.41 ± 0.50	0.41 ± 0.50	0.866
Diagnosis score (/7)	3.79 ± 1.64	3.77 ± 2.03	3.90 ± 1.92	0.859
7. Additional examinations (/3)	1.49 ± 0.51	1.54 ± 0.72	1.56 ± 0.60	0.397
8. The most relevant additional examination (/1)	0.33 ± 0.48	0.44 ± 0.50	0.38 ± 0.49	0.648
Additional examination score (/4)	1.82 ±;0.79	1.97 ± 1.06	1.95 ± 0.86	0.874
9. Treatment (/3)	1.67 ± 0.58	1.90 ± 0.79	1.85 ± 0.67	0.019
AIPI total score (/20)	11.82 ± 3.50	12.31 ± 4.21	12.15 ± 4.26	0.871
Consistency score (5-point Likert Scale)	2.36 ± 1.06	2.46 ± 1.45	2.10 ± 1.43	0.399

Abbreviations: AIPI, artificial intelligence performance instrument.

(n = 13). The judges reported a high ICC in the AIPI assessment [ICC = 0.888; 95% confident interval (0.831–0.932)].

Association analyses

The Spearman correlation analysis reported that DeepSeek and Claude-3.7-Sonnet performances were not significantly correlated with the complexity of cases. The complexity of cases was significantly moderately correlated with the ChatGPT-4o primary diagnosis score (r_s = 0.457; P = 0.003), and AIPI diagnosis subscores (r_s = 0.410; P = 0.008), meaning that ChatGPT-4o reported more accurate scores for difficult clinical cases.

DISCUSSION

As highlighted by Agarwal et al in their study on blood physiology, the accuracy and reliability of ChatGPT-4o, DeepSeek, and Claude-3.7 remain underexplored in

TABLE 3.
Number and Proportion Comparisons of Large Language Model Performances

AIPI Outcomes	DeepSeek	ChatGPT-4o	Claude-3.7	Chi-square
Primary Diagnosis (N (%))				
Correct (3)	6 (15.4)	10 (25.6)	11 (28.2)	
Plausible (2)	11 (28.2)	8 (20.5)	2 (5.1)	
Not plausible (1)	21 (53.8)	18 (46.2)	25 (64.1)	
Absent (0)	1 (2.6)	3 (7.7)	1 (2.6)	0.112
Differential diagnosis				
Correct (3)	6 (15.4)	8 (20.5)	10 (25.6)	
Plausible (2)	21 (53.8)	15 (38.5)	16 (41.0)	
Not plausible (1)	10 (25.6)	13 (33.3)	12 (30.8)	
Absent (0)	2 (5.1)	3 (7.7)	1 (2.6)	0.710
Relevant additional examination				
Pertinent and necessary (3)	0 (0.0)	5 (12.8)	1 (2.6)	
Pertinent and not necessary (2)	19 (48.7)	11 (28.2)	21 (53.8)	
Pertinent, necessary, and inadequate (1)	20 (51.3)	23 (59.0)	16 (41.0)	
Only inadequate examinations (0)	0 (0.0)	0 (0.0)	1 (2.6)	0.035
Treatment				
Pertinent and necessary (3)	1 (2.6)	9 (23.1)	4 (10.3)	
Pertinent and incomplete (2)	25 (64.1)	18 (46.2)	27 (69.2)	
Association of pertinent/necessary and inadequate	12 (30.8)	11 (28.2)	6 (15.4)	
No adequate strategy (0)	1 (2.6)	1 (2.6)	2 (5.1)	0.648
Management plan (0–1)				
Pertinent (1)	14 (35.9)	16 (41.0)	16 (41.0)	
Not pertinent (0)	25 (64.1)	23 (59.0)	23 (59.0)	0.866

Abbreviation: AIPI, artificial intelligence performance instrument.

specialized medical fields, a gap that also applies to the field of phoniatrics.[16] The findings of this study suggest that DeepSeek, ChatGPT-4o, and Claude-3.7-Sonnet demonstrate comparable and low clinical values in the analysis of phoniatric cases. With a mean score ranging from 11.82 to 12.31, the performance of the three LLMs outperforms those reported in the current literature for general otolaryngological cases. In a prospective case series of 100 real clinical cases, our group reported that ChatGPT-4 was associated with a mean AIPI score of 13.7, 13.3, and 15.1 for laryngological and swallowing, head and neck, and otological cases, respectively.[4] Interestingly, in this study, the ChatGPT-4 primary diagnoses were consistent with those of the practitioner in 67%, 65%, and 86% of otological, rhinological, and head and neck cases, respectively. Similarly to our study, ChatGPT-4 reported the lowest primary diagnosis rate in laryngology/swallowing disorders with 38% of consistent diagnoses.[4] Note that the clinical cases presented to ChatGPT-4 did not include clinical images in this study. Maniaci et al investigated how clinical images affect ChatGPT-4's diagnostic performance in laryngology and swallowing disorders.[12] Their findings revealed that ChatGPT-4 achieved lower AIPI scores when analyzing cases that included images compared to text-only clinical scenarios.[12] Despite different case profiles between our studies—with our cohort having a higher proportion of phoniatric cases, particularly vocal fold benign lesions—the AIPI scores for ChatGPT-4 and ChatGPT-4o were comparable (12.31 versus 12.45).

Recently, Chiesa-Estomba et al explored the accuracy of ChatGPT-4o in the identification of malignant vocal fold diagnoses based on the analysis of videolaryngostroboscopy.[13] In this study, including 20 patients, the authors observed that ChatGPT-4o identified the primary diagnosis in 30% of cases, while proposing malignancies as one of the top three diagnoses in 90% of cases.[13] With a comparable ICC to ours (0.890), Chiesa-Estomba et al reported a mean consistency score for image analysis of 2.36 ± 1.13, which is comparable to ours (2.46).

To date, most studies have focused their investigation of LLM accuracy on ChatGPT versions, which may be attributed to its large mediatization at the launching time (November 2022).[5,17,18] However, it was suggested that other LLMs, such as Claude-Sonnet may have similar or higher performance in clinical diagnosis-making.[19,20] The present study evaluated the diagnostic accuracy of two previously under-investigated LLMs in phoniatric cases: DeepSeek (a newly developed Chinese LLM) and Claude-3.7-Sonnet. Both LLMs have received limited attention in the current literature. Our results suggested that DeepSeek reported subtly lower performance than Claude-3.7-Sonnet and ChatGPT-4o, particularly for treatment recommendations, and when considering the recommendation of accurate additional diagnoses. Prasad et al evaluated the performance of DeepSeek-R1 and ChatGPT-4 in providing information for five common procedures in otolaryngology (adenotonsillectomy, tympanoplasty, endoscopic sinus surgery, parotidectomy, and total laryngectomy), revealing that both LLMs were associated with limitations in precision, comprehensiveness, and nuanced clinical reasoning, reporting key limitations of DeepSeek.[21] A recent study similar to ours, conducted by Hasnain et al, evaluated the performance of DeepSeek, ChatGPT-4o, and Claude 3.5 in image-based cases of conjunctivitis. It found that DeepSeek provided more precise and detailed information, with a reduced hallucination rate of 7%, compared to 13% for ChatGPT. In the same study, Claude achieved 100% accuracy in binary classification tasks, significantly outperforming ChatGPT's 62.5% accuracy. These findings partially support our own observations, although DeepSeek's performance in our cohort remained slightly below that of Claude 3.7 and ChatGPT-4o.[22] Sogalow et al compared the accuracy of five LLMs (ChatGPT-4o, Gemini-2.0-Flash, Claude-Sonnet-3.7, DeepSeek-R1, and Mistral-Large2) in the management of 63 humanitarian clinical situations.[20] In this study, ChatGPT-4o and Claude-3.7-Sonnet outperformed DeepSeek-R1 in most AIPI item scores and total score, which supports a potential lower accuracy of DeepSeek compared to the American LLMs.[20] Schmidl et al similarly reported that Claude-3-Sonnet demonstrated comparable and superior accuracy than ChatGPT-4 in the treatment recommendations and diagnostic work-up, respectively.[23]

The accuracy assessment of three widely used LLMs, including the poorly investigated DeepSeek and Claude-3.7-Sonnet, on real image-based phoniatric cases is the primary strength of this study. The analysis by two experienced practitioners reporting high interrater reliability is an additional strength. The low number of patients, the inability to upload videos on some LLMs (eg, DeepSeek), and the related lack of consideration of videolaryngostroboscopy rather than fixed clinical images are the primary limitations of the study. Despite these limitations, our findings may raise awareness on the mild accuracy of ChatGPT-4o, Claude-3.7-Sonnet, and DeepSeek in the analysis of images of benign phoniatric conditions.

CONCLUSION

The performance of AI-powered LLMs is mild for providing accurate analysis of phoniatric cases based on clinical findings and laryngostroboscopic images. Future studies are needed to assess the accuracy of updated LLM versions, potentially considering video analyses rather than only clinical images.

Author contributions

Camille F. Legrain: design, acquisition of data, data analysis & interpretation, drafting, final approval, and accountability for the work; final approval of the version to be published; 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. **Antonino Maniaci**: design, acquisition of data, data analysis & interpretation, drafting, final approval, and accountability for the work; final approval

of the version to be published; 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. **Lise Sogalow:** design, acquisition of data, data analysis & interpretation, drafting, final approval, and accountability for the work; final approval of the version to be published; 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. **Jerome R. Lechien:** design, acquisition of data, data analysis & interpretation, drafting, final approval, and accountability for the work; final approval of the version to be published; 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.

Declaration of Competing Interest

The author has no financial interest in the subject under discussion.

Acknowledgments

None.

References

1. Boussina A, Krishnamoorthy R, Quintero K, et al. Large language models for more efficient reporting of hospital quality measures. *NEJM AI.* 2024;1. https://doi.org/10.1056/aics2400420.
2. Blease CR, Locher C, Gaab J, Hägglund M, Mandl KD. Generative artificial intelligence in primary care: an online survey of UK general practitioners. *BMJ Health Care Inform.* 2024;31:e101102. https://doi.org/10.1136/bmjhci-2024-101102.
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. (DOI:). *Ear Nose Throat J.* 2025. https://doi.org/10.1177/01455613251335385.
4. Lechien JR, Naunheim MR, Maniaci A, et al. Performance and consistency of ChatGPT-4 versus otolaryngologists: a clinical case series. *Otolaryngol Head Neck Surg.* 2024;170:1519–1526. https://doi.org/10.1002/ohn.759.
5. Lechien JR, Rameau A. Applications of ChatGPT in otolaryngology-head neck surgery: a state of the art review. *Otolaryngol Head Neck Surg.* 2024;171:667–677. https://doi.org/10.1002/ohn.807.
6. Patel EA, Fleischer L, Filip P, et al. The use of artificial intelligence to improve readability of otolaryngology patient education materials. *Otolaryngol Head Neck Surg.* 2024;171:603–608. https://doi.org/10.1002/ohn.816.
7. 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 Otorhinolaryngol.* 2024. https://doi.org/10.1007/s00405-023-08441-8.
8. Frosolini A, Franz L, Benedetti S, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. *Eur Arch Otorhinolaryngol.* 2023;280:5129–5133. https://doi.org/10.1007/s00405-023-08205-4.
9. Lechien JR, Gorton A, Robertson J, Vaira LA. Is ChatGPT-4 accurate in proofread a manuscript in otolaryngology-head and neck surgery? *Otolaryngol Head Neck Surg.* 2023. https://doi.org/10.1002/ohn.526.
10. Lechien JR. Expanding the capacity of general practitioners in Sub-Saharan Africa with artificial intelligence. *Otolaryngol Head Neck Surg.* 2025. https://doi.org/10.1002/ohn.1335.
11. Filali Ansary R, Lechien JR. Clinical decision support using large language models in otolaryngology: a systematic review. *Eur Arch Otorhinolaryngol.* 2025. https://doi.org/10.1007/s00405-025-09504-8.
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;171:1106–1113. https://doi.org/10.1002/ohn.897.
13. Chiesa-Estomba CM, Andueza-Guembe M, Maniaci A, et al. Accuracy of ChatGPT-4o in text and video analysis of laryngeal malignant and premalignant diseases. *J Voice.* 2025. https://doi.org/10.1016/j.jvoice.2025.03.006.
14. von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med.* 2007;147:573–577. https://doi.org/10.7326/0003-4819-147-8-200710160-00010.
15. 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;281:2063–2079. https://doi.org/10.1007/s00405-023-08219-y.
16. Agarwal M, Sharma P, Wani P. Evaluating the accuracy and reliability of large language models (ChatGPT, Claude, DeepSeek, Gemini, Grok, and Le Chat) in answering item-analyzed multiple-choice questions on blood physiology. *Cureus.* 2025;17:e81871. https://doi.org/10.7759/cureus.81871.
17. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health.* 2023;11:1166120. https://doi.org/10.3389/fpubh.2023.1166120.
18. Lu JG, Song LL, Zhang LD. Cultural tendencies in generative AI. *Nat Hum Behav.* 2025. https://doi.org/10.1038/s41562-025-02242-1.
19. Lechien JR, Maniaci A. Large language models as adjunctive tools for diagnosing rare diseases in otolaryngology: a controlled study. Oral Communication at the Annual Meeting of the American Academy of Otolaryngology Head and Neck Surgery, October 12, Indianapolis, USA. *Rev J Otolaryngol Head Neck Surg.* 2025.
20. Sogalow L, Victoor L, Khalife M, Lechien JR. Evaluating five AI-powered language models as otolaryngology clinical support tools in rural Kenya. *Laryngoscope.* 2025.
21. Prasad S, Langlie J, Pasick L, Chen R, Franzmann E. Evaluating advanced AI reasoning models: ChatGPT-4.0 and DeepSeek-R1 diagnostic performance in otolaryngology: a comparative analysis. *Am J Otolaryngol.* 2025;46:104667. https://doi.org/10.1016/j.amjoto.2025.104667.
22. Hasnain M, Aurangzeb K, Alhussein M, Ghani I, Mahmood MH. AI in conjunctivitis research: assessing ChatGPT and DeepSeek for etiology, intervention, and citation integrity via hallucination rate analysis. *Front Artif Intell.* 2025;8:1579375. https://doi.org/10.3389/frai.2025.1579375.
23. Schmidl B, Hütten T, Pigorsch S, et al. 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.* 2024;281:6099–6109.