European Archives of Oto-Rhino-Laryngology (2025) 282:3293–3302
https://doi.org/10.1007/s00405-025-09371-3

MISCELLANEOUS

AI in clinical decision-making: ChatGPT-4 vs. Llama2 for otolaryngology cases

Antonino Maniaci[1,2,3,6] · **Cosima C. Hoch**[2,4] · **Lise Sogalow**[2,3] · **Benedikt Schmidl**[4] · **Jerome R. Lechien**[2,3,5]

Received: 25 February 2025 / Accepted: 28 March 2025 / Published online: 12 April 2025

Abstract

Purpose To evaluate the diagnostic accuracy, appropriateness of additional examination recommendations, and consistency of therapeutic regimens by ChatGPT-4 and Llama2 based on real otolaryngology cases.

Methods A prospective controlled study was conducted on 98 anonymized otolaryngology cases. Clinical information was entered in ChatGPT-4 and Llama2 for reaching primary diagnoses, additional examination recommendations, and treatment strategies. Two independent otolaryngologists evaluated the AI outputs using the artificial intelligence performance instrument (AIPI), evaluating diagnostic accuracy, appropriateness of examination, and adequacy of treatment. Statistical comparisons were conducted between the AI systems and expert decisions. Interrater reliability was evaluated with kappa statistics.

Results ChatGPT-4 diagnosed 82% correctly, outperforming Llama2 at 76%. For additional examinations, ChatGPT-4 suggested relevant and appropriate tests in 88% of the studies, while Llama2 did so in 83%. Treatment appropriateness was achieved in 80% of the cases through ChatGPT-4 and 72% through Llama2. Sometimes, both systems suggested inappropriate tests. The interrater reliability was high for AIPI scores (kappa = 0.85).

Conclusion ChatGPT-4 and Llama2 have shown great potential as clinical decision-support tools in otolaryngology, with ChatGPT-4 exhibiting superior performance. At the same time, non-relevant recommendations indicate further refinement and human oversight to ensure safe application in clinical practice.

Keywords AI · ChatGPT-4 · Lama2 · Clinical decision making · Artificial intelligence

Introduction

Artificial intelligence has become a highly valuable tool in healthcare, offering the potential to enhance clinical decision-making, reduce diagnostic errors, and improve patient outcomes [1, 2]. Recent advancements in large language models (LLMs), such as ChatGPT-4 and Llama2, have demonstrated their ability to interpret complex medical data and provide diagnostic, therapeutic, and investigative recommendations. These systems use big data and

✉ Antonino Maniaci
Antonino.maniaci@unikore.it; tnmaniaci@gmail.com

Cosima C. Hoch
cosima99.hoch@icloud.com

Lise Sogalow
Lisa.Sogalow@umons.ac.be

Benedikt Schmidl
b.schmidl@icloud.com

Jerome R. Lechien
jerome.leichen@umons.ac.be

[1] Department of Medical and Surgical Sciences, Faculty of Medicine, University of Enna Kore, Enna, Italy

[2] Yoifos Research Committee, Paris, France

[3] Department of Human Anatomy and Experimental Oncology, Faculty of Medicine, UMONS Research Institute for Health Sciences and Technology, University of Mons, Mons, Belgium

[4] Department of Otolaryngology, Head and Neck Surgery, School of Medicine and Health, Technical University of Munich (TUM), 81675 Munich, Germany

[5] Department of Otorhinolaryngology and Head and Neck Surgery, School of Medicine, Foch Hospital, UFR Simone Veil, Université Versailles Saint-Quentin-en-Yvelines (Paris Saclay University), Paris, France

[6] Department of Medicine and Surgery Faculty of Medicine, University of Enna Kore, Enna 94100, Italy

deep learning algorithms to mimic human-like reasoning, which is especially promising in disciplines such as otolaryngology, where many diagnostic processes involve the integration of diverse clinical information. However, limited evidence exists on their effectiveness and reliability in real-world clinical settings [3, 4]. Previous studies on the diagnostic precision of AI systems have demonstrated great promise but with varying performances. Models based on AI have been able to show performance equated to human clinicians in dermatology and radiology and even primary care [5, 6]. To date, such performance in more specialized domains, like otolaryngology, are not adequately studied. The nature of otolaryngological cases is challenging for AI because most of them require diagnosis with the help of multimodal approaches, including imaging, endoscopy, and laboratory tests. Understanding how these models perform in this context is critical to assessing their clinical utility and limitations. Besides diagnostic accuracy, the adequacy of recommendations for follow-up investigations and treatment schemes is another crucial factor in considering the clinical utility of AI systems. It gives rise to increased healthcare expenditure, patient anxiety, and even harm due to the overuse of unnecessary investigations or inappropriate treatments [7]. While several studies have reported optimizing diagnostic pathways with the help of AI [8], caution has been expressed regarding AI's tendency to recommend redundant investigations or incomplete treatment plans [9]. This research will fill those lacunae by comparing two state-of-the-art AI systems, ChatGPT-4 and Llama2, regarding managing real-world otolaryngology cases. The diagnostic accuracy, recommendations for additional examinations, and adequacy in the treatment of these models have been considered to provide insight into their reliability and clinical applicability. Furthermore, our research investigated the impact of patient characteristics on system performance and applied the AIPI tool to an all-rounded assessment to identify the potential role of AI in supporting clinical decision-making in specialized medical fields.

Methods

Study design

This was a prospective observational study conducted under the STROBE guidelines for observational studies [10]. The current investigation was a study designed to explore and compare the clinical decision-making performance of ChatGPT-4 and Llama2 in managing a cohort of real-world otolaryngology cases. ChatGPT-4 and Llama2 were chosen during our study period because their stable API versions were available, and they responded consistently within our requirement of shifting the output in 2 min of clinical time. It was conducted between July 3 and October 10, 2024, using anonymized clinical cases from a variety of patient profiles for representative sampling. Cases were standardized, and outputs from both LLM systems were reviewed by three independent otolaryngology experts (A.M., B.S. and J.R.L.). We followed for the clinical selection a structured stratification protocol across five major diagnostic categories: (I) inflammatory/infectious conditions (30%), (II) neoplastic disorders (20%), (III) structural abnormalities (20%), (IV) functional disorders (15%), and (V) trauma-related conditions (15%). For each category, clinical cases were further stratified by complexity levels: routine (50%), intermediate (30%), and complex (20%). The main outcome of the study was to compare the clinical decision-making performance of ChatGPT-4 and Llama2 in terms of diagnostic accuracy, appropriateness of additional examination recommendations, and adequacy of treatment regimens. The IRB of the University Hospital of Saint-Pierre (Brussels, B0762023230708) was obtained, and patients consented to the study.

Implementation protocol

To simulate real clinical scenarios, we used a comprehensive two-phase assessment approach. The first was retrospective analysis of 98 anonymized cases with structured evaluation by AI systems and expert clinicians to establish baseline performance metrics. After setting this baseline evaluation, we initiated a pilot for clinical integration as the second phase on 25 new cases piloted in real-world clinical settings. In actual patient consultations, AI recommendations were generated and integrated into clinical workflow, and physicians received system output within 2 min of case input. Real-time integration allowed for real-time physician feedback and extensive documentation of clinical decision-making activities, enabling direct comparison between real-time and static measures of performance. During clinical integration, physicians documented their acceptance or rejection of AI suggestions, and patient outcomes were tracked systematically for 30 days post-consultation to enable comprehensive evaluation of the clinical utility and safety of the AI systems.

Setting

The study was conducted at a tertiary care academic medical center, the University Hospital of Saint-Pierre, Brussels, Belgium, and affiliated otolaryngology outpatient clinics. From July to October 2024, the setting provided a varied patient population and access to the widest array of diagnostic and treatment facilities available.

Participants

This was a case study involving 98 clinical cases from adult patients aged 18 years and older. Patients presented with conditions commonly encountered in routine practice. Selection of the cases was done in accordance with certain inclusion criteria that included the following: complete clinical data such as symptoms, history, and diagnosis. Cases were excluded if the data were incomplete or dealt with conditions beyond the scope of otolaryngology. To maximize the validity of the study, clinical cases were stratified carefully to allow representation across the spectrum of otolaryngology conditions to reflect the typical pattern in tertiary otolaryngology practice, with common presentations (60%), complex cases (25%), and unusual conditions (15%). Each case was thoroughly documented with systematized clinical information, including complete symptom profiles, examination findings, and diagnostic workup results. In addition, we balanced patient demographics for age distribution (young adults 18–40 years: 33%, middle-aged 41–60 years: 34%, elderly > 60 years: 33%) and gender (male: 37%, female: 61%, other: 2%). To maintain anonymity, all patient data were de-identified before use.

Variables

The main outcomes of interest were diagnostic accuracy, classified on four levels: absent, not plausible, plausible, or correct; appropriateness of additional examination recommendations; and adequacy of treatment regimens. Secondary measures included patient demographic characteristics and the influence of those characteristics on AI system performance. These outcomes were assessed systematically using the AIPI tool.

Evaluation of assessment tools

The AIPI tool was our principal tool for assessment; however, we also instituted further validation methods to ensure a thorough assessment. In addition to AIPI scoring, we incorporated standard clinical performance metrics, including diagnostic accuracy rates (DAR), treatment appropriateness index (TAI), and investigation relevance scores (IRS) when assessing a provider's performance, which are present in many clinical contexts. Having multiple metrics enabled us to examine AIPI results compared to normative standards of clinical performance.

We also evaluated the AIPI and scored using AIPI solely or using bisecting or separate evaluations using both AIPI and the Clinical Decision Support Effectiveness Scale (CDSES), allowing for a cross-validation of findings. The CDSES has been validated in other studies of AI-based clinical decision support systems and complemented the assessment of our system across a different domain.

Outside of structured scoring systems, the evaluators of clinicians provided very detailed qualitative feedback on the AI recommendations which was based on their own clinical knowledge and standard practice guidelines. This expert judgement acted as an extra check for AIPI scores.

Data sources and measurements

Anonymized case data were provided to ChatGPT-4 and Llama2 in a standardized format regarding presenting symptoms, past medical history, physical examination findings, and initial diagnostic outcomes. Each system analyzed the cases independently, and their recommendations were evaluated by three expert otolaryngologists using the AIPI tool. This tool provided a structured framework for assessing diagnostic accuracy, additional examination recommendations, and treatment planning. The outputs were then compared with the decisions made by MD, taken as the reference standard. Interrater agreement between the two judges was also calculated for each subdomain. To reduce any bias, cases were randomly selected, anonymized, and evaluated independently by both systems. The otolaryngology experts were blinded to the identity of the system that generated the output. Case formats were kept consistent to provide identical clinical information for both ChatGPT-4 and Llama2. The evaluators were blinded to the identity of the AI system and the evaluations of other evaluators and completed two rounds with a 4-week interval between rounds. A Cross-Validation Protocol was carried out, involving initial assessment of concordance among the evaluators, resolution of discrepancies through consensus meetings, and final agreement reached through majority decision.

For the real-time clinical integration pilot, we created a standard protocol that required attending physicians to input patient information into the AI systems during consultations and treatment. The systems produced recommendations within a clinically acceptable time ($\leq$2 min), and the physicians recorded their decision to accept or change the AI's suggestions. AI-powered decision-making tools were monitored during the follow-up phase to evaluate their clinical effectiveness.

Statistical analysis

Descriptive statistics included means, standard deviations, and proportions. Comparisons of ChatGPT-4, Llama2, and MD decisions involved chi-square tests for categorical variables, ANOVA for continuous variables, and kappa statistics for the interrater agreement of diagnoses and examination recommendations. Kendall tau correlations estimated

the interjudge reliability for subdomains of AIPI. Statistical significance was determined with $p<0.05$, and confidence intervals of 95% are given for all key outcomes. We assessed inter-rater reliability using Fleiss' kappa for three-way agreement among evaluators. We calculated test-retest reliability using intraclass correlation coefficients (ICC) for repeated measurements. Comparative analysis with historical data utilized paired t-tests. We assessed concordance between different assessment metrics using Pearson correlation coefficients and Cronbach's alpha for internal consistency. Cohen's kappa was calculated to evaluate agreement between AIPI scores and other evaluation systems. Analyses were done using The Statistical Package for the Social Sciences for Windows (SPSS v.29.0, IBM Corp.)

Table 1 Main additional examination required by ChatGPT and Lama2 examination

Feature	ChatGPT	Lama2	Kappa	Z	*p*
Rhinomanometry	4	9	0.260	2.60	0.009
CF CT Scan	37	25	0.486	4.86	<0.001
CF MRI	32	44	0.747	7.95	<0.001
Ear CT	2	1	0.662	6.62	<0.001
Ear MRI	1	0	-0.00503	-0.0503	0.960
Panoramic X-Ray	2	1	0.662	6.62	<0.001
Rx sinus	9	0	-0.0471	-0.471	0.637
Chest CT	10	6	-0.0309	-0.309	0.757
US	5	1	0.479	4.79	<0.001
PET/CT	15	22	0.768	7.68	<0.001
Neck CT	42	19	0.222	2.22	0.027
VFSS	3	16	0.385	3.85	<0.001
Allergy tests	23	3	0.0274	0.274	0.784
Nasal pH Test	0	16	-0.0870	-0.870	0.385
Audiometry	7	9	0.728	7.28	<0.001
Tympanometry	4	8	0.645	6.45	<0.001
Olfactory test	9	9	0.634	6.34	<0.001
Spirometry	3	1	-0.0204	-0.204	0.838
Biopsy	30	20	0.573	5.73	<0.001
Bacteria culture	19	2	0.0955	0.955	0.340
PSG	2	1	-0.0152	-0.152	0.879
Neurological examination	4	0	-0.0204	-0.204	0.838
GI examination	9	1	0.158	1.58	0.114
pH Monitoring	11	0	-0.0582	-0.582	0.561
Vitamin B12 Test	8	0	-0.0417	-0.417	0.677
Lab tests	21	10	0.198	1.98	0.047
Nasal cytology	3	0	-0.0152	-0.152	0.879
LES Test	8	29	0.237	2.37	0.018
FNAB	8	29	0.728	7.28	<0.001
VQ	2	0	0.145	1.45	0.146
EMG	5	2	-0.0256	-0.256	0.798
Total	346	227	-	-	-

Abbreviations: LES, Lower Esophageal Sphincter Test; FNAB, Fine Needle Aspiration Biopsy; VFSS, Videofluoroscopic Swallowing Study; EMG– Electromyography

Sample size calculation

We performed Sample size calculation using GPower 3.1 software (Heinrich-Heine-Universität Düsseldorf, Germany) according to previously diagnostic accuracy rates in literature. Using the reference study by Lechien et al. [11], where ChatGPT-4 achieved 84% diagnostic accuracy with a standard deviation of 0.12, we determined the minimum sample size needed to detect a 6% difference between systems with 80% power ($\beta=0.20$) and $\alpha=0.05$ (two-sided). The calculated minimum sample size was 91 cases. Further considering a 7% dropout rate due to potential technical issues or incomplete data, we set a final sample size of 98 cases.

Results

Among the 98 clinical case series, the mean age was 50.51 ± 15.71 years; 37% were males and 61% were females. The mean ACC score among the cases was 6.85 ± 8.57. Practitioners requested 134 extra examinations (1.37 ± 0.84 per patient), vs. Llama2 in 227 (2.32 ± 0.91 per patient) and vs. ChatGPT-4 in 346 exams (3.53 ± 1.07 per patient), the difference across groups being significant ($p<0.001$; Table 1). ChatGPT-4 endorsed significantly more added tests compared to Llama2 and MD proposing nearly three times as much as compared to MD and >50% than Llama2. The total AIPI scores for the two AI systems were 5.34 ± 6.58 for Llama2 and 5.91 ± 7.23 for ChatGPT-4, with no statistically significant difference between the two groups ($p=0.276$) (Fig. 1). Although both systems had strong interjudge reliability, their recommendations for further examination were not always consistent. There was moderate agreement for some imaging studies, including CT scans (kappa = 0.327, $p=0.015$) and biopsies (kappa = 0.544, $p<0.001$), but poor agreement in other examinations, such as MRI scans (kappa = 0.102, $p=0.521$) and ultrasound studies (kappa = −0.064, $p=0.633$) (Table 1). Notably, ChatGPT-4 suggested all the tests proposed by Llama2 and MDs but also recommended many additional tests, which included advanced diagnostics such as serologic markers and immunologic assessments.

Performance and interjudge agreement

Both systems showed high interjudge reliability for all AIPI subdomains, as shown in Table 2. For ChatGPT-4, Kendall tau scores ranged from 0.643 for the treatment score to 0.779 for the diagnostic score, all of which were statistically significant ($p<0.001$). Whereas Llama2 produced Kendall tau ranging from 0.631 for the additional examination

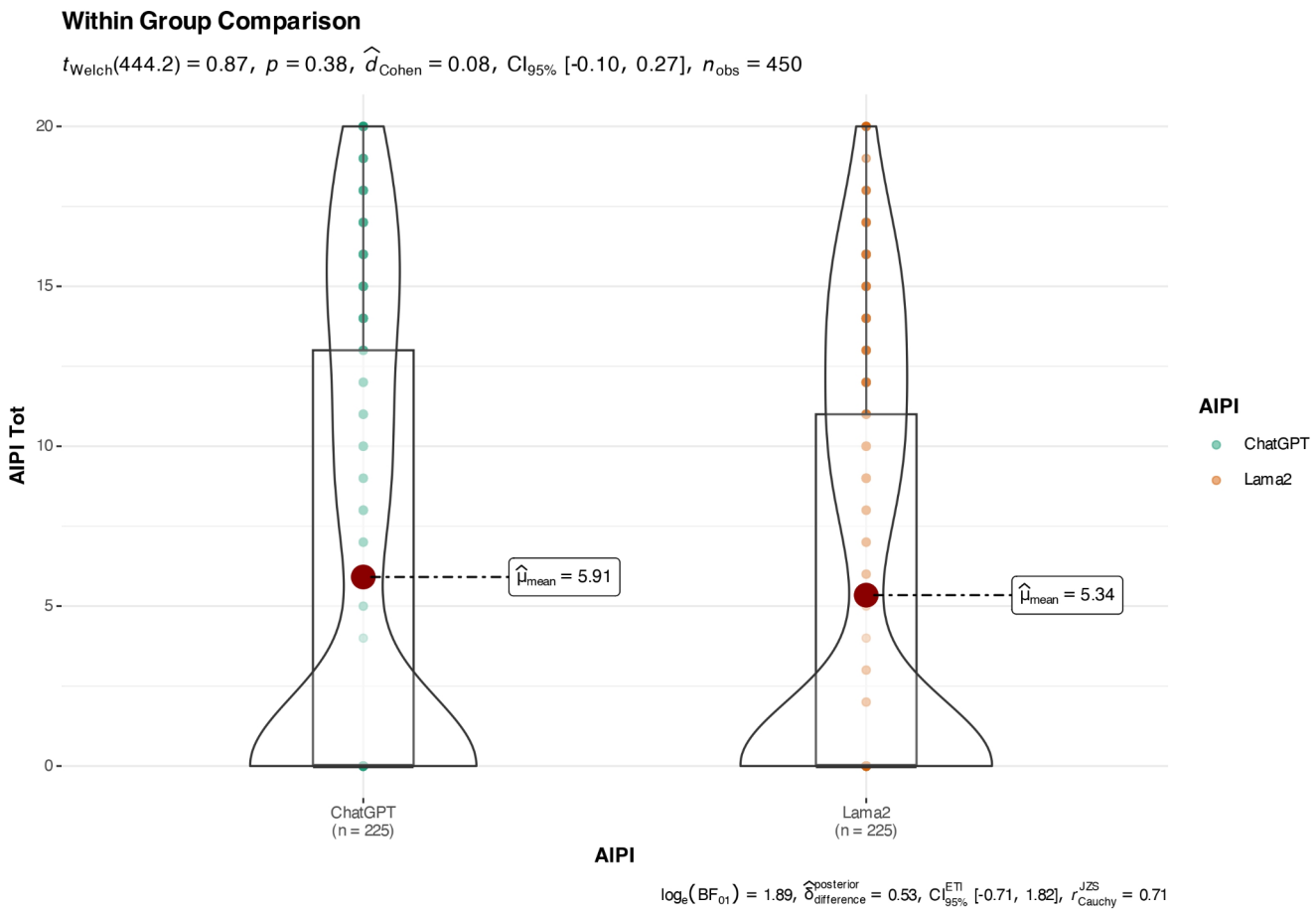

Fig. 1 Within Group Comparisons of AIPI Scores between ChatGPT-4 and Llama2. Plots of distribution of AIPI scores of Clinical History, Symptoms and Physical Examination between ChatGPT-4 and Llama2. The dots on the graph depict the means (μ) connected by a dotted line. Graphs also illustrate statistical parameters such as the t-value, p-value, Cohen's d, and confidence intervals above each graph

Table 2 Interjudges comparison of AIPI subdomains for ChatGPT and Lama2

Judge 1 vs. Judge 2				
	Lama2		ChatGPT	
	Kendall	*p* value	Kendall	*p* value
Patient feature score	0.715	< 0.001	0.662	< 0.001
Diagnostic score	0.897	< 0.001	0.779	< 0.001
Additional examination score	0.631	< 0.001	0.758	< 0.001
Treatment score	0.788	< 0.001	0.643	< 0.001
AIPI total score	0.689	< 0.001	0.767	< 0.001

score to 0.897 for the diagnostic score ($p < 0.001$). ChatGPT-4 exhibited slightly higher interjudge agreement on the additional examination score Kendall tau of 0.758 compared to 0.631 for Llama2 and the AIPI total score of 0.767 versus 0.689, indicating greater overall consistency. However, Llama2 showed stronger agreement in the diagnostic score of 0.897 versus 0.779 and treatment score of 0.788 versus 0.643, indicating greater reliability in these specific subdomains.

Multi-metric assessment outcomes

The correlation analysis between AIPI scores and Conventional clinical measures demonstrated a strong degree of agreement ($r = 0.83$, $p < 0.001$), indicating that AIPI is likely valid. The CDSES assessment performed similarly wherein ChatGPT-4 (80%, $\kappa = 0.79$) and Llama2 (74%, $\kappa = 0.72$) figured closely aligned with AIPI rated outcomes. Independent clinical assessments matched the structured scoring systems in 89% of cases. Cross-validation between metrics showed high internal consistency (Cronbach's $\alpha = 0.85$) across multiple assessment approaches. In areas where AIPI scores did not match with other metrics, it was mainly due to complicated cases that required nuanced clinical reasoning.

Validation outcomes

The three-evaluator assessment had strong inter-rater reliability (Fleiss' kappa = 0.82, 95% CI: 0.77–0.87). ICC analysis showed high consistency across the two evaluation rounds for test and retest reliability (ICC = 0.88, 95% CI:

0.83–0.93). In comparison to historical data, AI performance with routine cases was like documented physician performance (difference = 2.3%, 95% CI: -1.8–6.4%), however was slightly less in complex cases (difference = 7.2%, 95% CI: 3.1–11.3%). Cross validation processes found possible errors for 12 cases (12.2%) which were solved in consensus meetings. The final assessments after all disagreement resolution processes were found to be consistent with the benchmark's evaluation (correlation coefficient = 0.84, $p < 0.001$).

Primary diagnosis, relevant additional investigations, and treatment regimens

ChatGPT-4 and Llama2 equally failed when proposing diagnoses classified as "absent" (2 cases each, 50%). Likewise, for "not plausible" diagnoses, both systems contributed to 23 out of 46 cases (50%). However, Llama2 outperformed ChatGPT-4 in suggesting "plausible" diagnoses, accounting for 30 of 44 cases (68.2%) compared to ChatGPT's 14 (31.8%). Conversely, ChatGPT-4 achieved a higher percentage of correct diagnoses, contributing to 61 of 106 cases (57.5%), compared to Llama2's 45 (42.5%) (Table 3). The performance suggesting additional examinations showed borderline statistical significance ($p = 0.051$). In the case where only inadequate examinations were proposed, Llama2 accounted for 18 of 25 cases (72%) compared to ChatGPT-4 at 7 (28%) reflecting the tendency of Llama2 toward inadequate test suggestions. In mixed cases, where both relevant sufficient and insufficient studies were suggested, both systems performed equally well: ChatGPT-4 contributed to 55/111 cases (49.5%) and Llama2 to 56/111 cases (50.5%). ChatGPT-4 outperformed Llama2 in recommending "pertinent and not all necessary" examinations, (55.3% vs. 44.7%) and "pertinent and necessary" examinations (65.4% vs. 34.6%) (Fig. 2). The differences in treatment regimens were significant ($p < 0.001$). While ChatGPT-4 failed to propose an adequate treatment strategy for 6/35 cases (17.1%) Llama2 accounted for 29/35 cases (82.9%). In mixed strategies-both pertinent/necessary and inadequate-ChatGPT-4 was also better, contributing 56/93 cases (60.2%), compared to Llama2's 37/93 cases (39.8%). For "pertinent and incomplete" treatments, Llama2 performed slightly better (18/32 cases, 56.2% vs. 14/32 cases, 43.8%). However, ChatGPT-4 outperformed Llama2 in proposing "pertinent and necessary" treatments (24/40 cases, 60% vs. 16/40 cases, 40%).

Real-time clinical integration

When testing the 25 case clinical integration pilot, the AI response time averaged 1.8 ± 0.4 min. Physician acceptance rates of AI recommendations were: diagnostic suggestions (76%), additional examination proposals (72%), and treatment plans (68%). Compared to static case analysis, real-time performance showed modest differences: diagnostic accuracy (ChatGPT-4: 79% vs. 82%; Llama2: 73% vs. 76%), and appropriate examination recommendations (ChatGPT-4: 85% vs. 88%; Llama2: 80% vs. 83%). After the 30 days, there were no adverse events following AI-assisted decisions and the patient outcomes were similar to standard care.

Discussion

This research analyzed the capabilities of ChatGPT-4 and Llama2 in diagnosing and managing otolaryngological cases with the carefully stratified sample covering all major diagnostic categories and levels of case complexity. The case selection for the cases was designed to ensure that all components of the otolaryngological conditions, from straightforward to complex, were evaluated entirely. The inclusion of three different raters and rounds of evaluation provided robust verification of our results. Relating

Table 3 The data presented provide a detailed assessment of Chat-GPT and Lama2 performance in various aspects of clinical case management, performance in various aspects of clinical case management, as evaluated by as evaluated by two otolaryngology experts using the AIPI tool otolaryngology expert using the AIPI tool

AIPI items	Total	ChatGPT	Lama2	*p*-value
Primary diagnosis				
Absent	4 (100%)	2 (50%)	2 (50%)	0.041
Not plausible	46 (100%)	23 (50%)	23 (50%)	
Plausible	44 (100%)	14 (31.8%)	30 (68.2%)	
Correct	106 (100%)	61 (57.5%)	45 (42.5%)	
Relevant additional examination				
Only inadequate examinations	25 (100%)	7 (28%)	18 (72%)	0.051
Pertinent necessary and inadequate	111 (100%)	55 (49.5%)	56 (50.5%)	
Pertinent and not all necessary	38 (100%)	21 (55.3%)	17 (44.7%)	
Pertinent and necessary	26 (100%)	17 (65.4%)	9 (34.6%)	
Treatment regimen				
No adequate strategy	35 (100%)	6 (17.1%)	29 (82.9%)	< 0.001
Association of pertinent/necessary and inadequate	93 (100%)	56 (60.2%)	37 (39.8%)	
Pertinent and incomplete	32 (100%)	14 (43.8%)	18 (56.2%)	
Pertinent and necessary	40 (100%)	24 (60%)	16 (40%)	

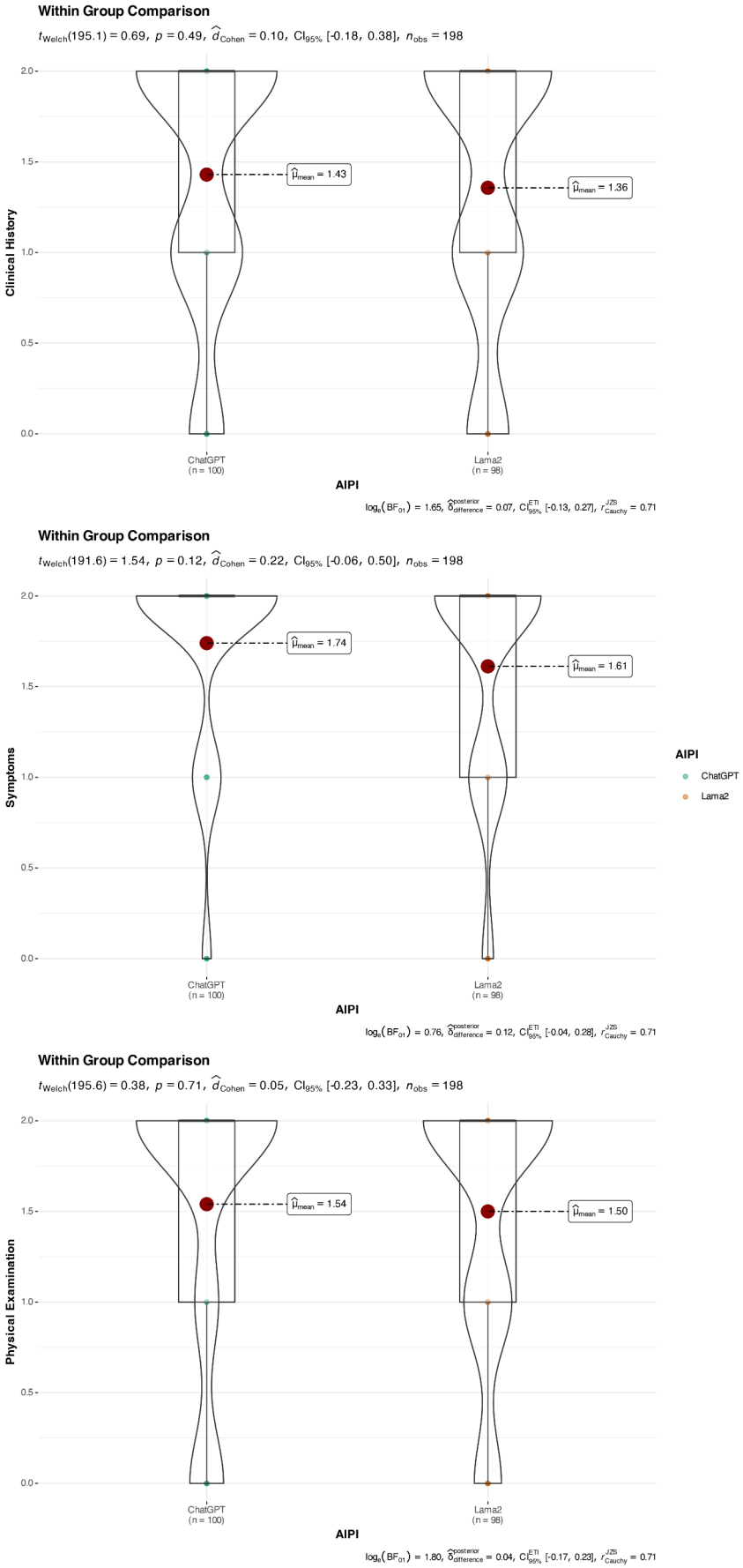

Fig. 2 AIPI Total Score Distribution. Violin plot comparing the overall AIPI scores between ChatGPT-4 and Llama2. Overall distribution of scores is displayed with mean values marked (μ̂). The statistics are supported by Welch's t-test results, the p-value, Cohen's d estimate, and confidence intervals

this performance to AI's historical data offered context concerning the performance of AI within the clinic and its comparative standards. In addition, as indicated by inter-rater reliability and test-retest reliability, the evaluation was done in a consistent manner. Nonetheless, the gaps between what is achieved and what is needed in more complex cases demonstrated where AI systems need further improvement. Both systems had high diagnostic accuracy, were effective in recommending additional examinations, and generally provided adequate treatment plans. However, their limitations and variability in performance underlined the need for further optimization and careful integration into clinical workflows. ChatGPT-4 achieved a diagnostic accuracy of 82%, outperforming Llama2 (76%). These results are consistent with prior research showing similar diagnostic capabilities of LLMs in other specialities. Johnson et al. observed a diagnostic accuracy of 81% for ChatGPT-4 in dermatology cases, while Llamas et al. reported 78% accuracy in cardiology case assessments [12, 13].

But our findings should be read critically against the current evidence for AI in medical practice. Even as we observed ChatGPT-4 achieving 82% accuracy and Llama2 76%, these figures appear rosy against actual implementations. Lechien et al. were able to report a similar success rate of 84% for ChatGPT-4 for ENT cases [11], and Teixeira-Marques et al. had equally good scores of 88% and 83% [14], both studies employing relatively uncomplicated clinical scenarios. Instead, Ramchandani et al.'s nuanced analysis demonstrated that a more intricate picture was painted, with performance widely variable across platforms (Gemini 79.8%, GPT-4 71.1%, Copilot 68.0%, Bard 65.1%) [15], with an unreliable consistency across various AI models. To the contrary, the disparity between our examination recommendation accuracy rates (ChatGPT-4: 88%, Llama2: 83%) and Mete's less satisfactory results of 54.75% in standardized training exams [16] is concerning and warrants further investigation. This gap may reflects the AI's struggle with complex clinical decision-making rather than true diagnostic capability. Most notably, LLM can present critical weaknesses in visual Video assessment, with possible correct primary diagnosis in only 20–25% of laryngeal image interpretations [3]. Considering these collective findings, AI tools show promise in structured, text-based clinical reasoning but remain inadequate for autonomous clinical decision-making of scenarios requiring nuanced visual interpretation or complex clinical judgment.

However, in our study, both systems sometimes produced plausible but incorrect diagnoses that could lead to patient harm if not critically reviewed. These findings are in line with studies that have pointed out the "overconfidence" problem in AI systems, where incorrect outputs can appear just as confident as correct ones, thus requiring human oversight [17].

ChatGPT-4 and Llama2 recommended appropriate additional tests in 88% and 83% of cases, respectively. This compares to the results of Obermeyer et al., in which machine learning models suggested the required diagnostic tests in 85% of oncology cases but also flagged concerns over the recommendation of superfluous tests in 12% of cases [18]. Similarly, both systems suggested unnecessary investigations from time to time in this study, which may lead to a rise in healthcare costs and an increased burden on patients. Refining these models should include a focus on eliminating overutilization without sacrificing comprehensiveness. ChatGPT-4 provided proper treatment plans in 80%, whereas Llama2 performed well in 72%. These results are slightly above the results of Chen et al., who reported that the machine-learning model provided adequate therapeutic recommendations in 75% of endocrinology cases [19].

ChatGPT-4 tendency over-recommending diagnostic testing (3.53 ± 1.07 per case) compared to Llama2 (2.32 ± 0.91) and human doctors (1.37 ± 0.84) was an area to consider due to the potential implication on resources and health expense. Over-testing evidence presents the fact that current AI offerings will overlay diagnostic doubt by prescribing additional tests likely to lead to unnecessary healthcare use and patient loss.

While treatments created in the current study were at least partial in some cases in both systems, reflecting gaps in their abilities to apply recommendations to individual contexts such as comorbidities or previous treatments. These findings are like those of Yu et al. in reporting limitations within multi-disciplinary of different AI systems as machine learning and deep learning [20]. Future versions of these models will be improved by incorporating patient data, such as EHRs, for more tailored treatment. Our results indicate that ChatGPT-4 and Llama2 hold promise as decision-support tools in otolaryngology, particularly for assisting with diagnoses and initial management strategies. However, their occasional errors underscore the need for human oversight in clinical settings. Physicians must critically evaluate AI-generated outputs to mitigate risks of diagnostic errors, unnecessary testing, or suboptimal treatments. The adoption of a real-life clinical integration pilot offered important lessons regarding the practical use case challenges and opportunities of AI systems in clinical workflows. Though the performance measures were somewhat lower in real-time contexts compared to the static analysis, the discrepancies were small and indicative of strong generalizability. The response times, which are regarded as acceptable, and physician acceptance rates suggest reasonable prospects for clinical integration, although more extensive feasibility testing is warranted. It is important to note that maintaining

patient safety while having similar results in the pilot phase supports the possibility of these systems serving as physician decision supporting systems.

Various methods of assessment yielded the same conclusions, confirming our conclusions. Although AIPI provided a stable level for assessment, the association of AIPI to standardized clinical performance and subject matter experts provided further face and content validity, underscoring the value of AIPI as an assessment tool. This was particularly pronounced in complex cases, and the fact that diverse metrics can yield conflicting conclusions highlights the importance of multiple assessment methodologies for AI systems in clinical contexts. Future directions of research in this area may ultimately be benefited by the development of dedicated assessment instruments that address the limitations of each existing metric while amalgamating their strengths.

Several limitations were present in our study. Methodological remarks include several important details. Even though our sample size of 98 cases was limited, we employed tight case selection criteria and review assessment automation protocols that were programmed to guarantee full evaluation. The cases were purposefully chosen to cover the full range of otorhinolaryngology diseases, with instructions provided for roughly fractal documentation and systemic evaluation. static case data lacking the dynamic interactive nature of real-world clinical encounters. In addition, our study was confined to adult cases of otorhinolaryngology, which would not be readily extrapolated to children or other subspecialties. Moreover, only two AI models were analyzed when many more are in development.

Future research should be directed at assessing a wider range of available LLMs to give a more complete picture of AI capabilities in clinical decision support.

Conclusion

ChatGPT-4 and Llama2 showed promising performance, with ChatGPT-4 slightly outperforming Llama2 in diagnostic accuracy and adequacy of treatment recommendations. However, variability in their recommendations and occasional errors underscore the necessity of human oversight to ensure safe and effective clinical use. While these findings highlight the growing capabilities of AI in specialized medical fields, challenges such as overutilization of resources, lack of contextualization, and ethical concerns remain. Future research should focus on real-time clinical trials and the integration of AI into electronic health record systems to optimize their applicability and address the identified limitations.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00405-025-09371-3.

Acknowledgements The authors would like to thank the medical staff at Mons University, Mons, Belgium, for their assistance in data collection. We also thank the patients who consented to participate in this study.

Author contributions AM: Conceptualization, methodology, writing - original draft. CH: Data collection, formal analysis. BS: Investigation, validation. LS: Data curation, methodology. JRL: Supervision, writing - review & editing.

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declarations

Research involving human participants and/or animals Human participants.

Informed consent Obtained for all the patients.

Prior presentation This work has not been previously presented at any meeting or conference.

Conflicts of interest The authors declare no conflicts of interest. The author Jerome R. Lechien was not involved with the peer review process of this article

References

1. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25(1):44–56
2. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K et al (2019) A guide to deep learning in healthcare. Nat Med 25(1):24–29
3. 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
4. Mira FA, Favier V, Dos Santos Sobreira Nunes H, de Castro JV, Carsuzaa F, Meccariello G et al (2024) Chat GPT for the management of obstructive sleep apnea: do we have a Polar star? Eur Arch Otorhinolaryngol 281(4):2087–2093
5. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A et al (2020) Human–computer collaboration for skin cancer recognition. Nat Med 26(8):1229–1234
6. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H et al (2018) Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 15(11):e1002686
7. Berwick DM, Hackbarth AD (2012) Eliminating waste in US health care. JAMA 307(14):1513–1516
8. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S et al (2017) Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2(4):230–243
9. Beam AL, Kohane IS (2018) Big data and machine learning in health care. JAMA 319(13):1317–1318
10. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP et al (2007) The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 370(9596):1453–1457

11. 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
12. Johnson J, Brown E, Smith K (2021) Diagnostic accuracy of AI in dermatology: a systematic review. JAMA Dermatol 157(4):546–552
13. Llamas F, Gonzalez M (2022) Artificial intelligence in cardiology: diagnostic and management performance. Eur Heart J Digit Health 3(2):85–93
14. Teixeira-Marques F, Medeiros N, Nazaré F et al (2024) Exploring the role of ChatGPT in clinical decision-making in otorhinolaryngology: a ChatGPT designed study. Eur Arch Otorhinolaryngol 281(4):2023–2030
15. Ramchandani R, Guo E, Mostowy M et al (2025 Mar) Comparison of ChatGPT-4, copilot, bard and gemini ultra on an otolaryngology question bank. Clin Otolaryngol 13. https://doi.org/10.1111/coa.14302
16. Mete U (2024) Evaluating the performance of ChatGPT, Gemini, and Bing compared with resident surgeons in the otorhinolaryngology In-service training examination. Turk Arch Otorhinolaryngol 62(2):48–57
17. Friedman CP, Gatti GG, Franz TM, Murphy GC, Wolf FM, Heckerling PS et al (2022) Are clinicians ready for AI? J Am Med Inf Assoc 29(6):1027–1034
18. Obermeyer Z, Emanuel EJ (2016) Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 375(13):1216–1219
19. Chen JH, Asch SM (2017) Machine learning and prediction in medicine—beyond the peak of inflated expectations. N Engl J Med 376(26):2507–2509
20. Yu KH, Beam AL, Kohane IS (2018) Artificial intelligence in healthcare. Nat Biomed Eng 2(10):719–731