

Diagnostic Performance of ChatGPT-o1 and DeepSeek-V3 in Expert-Validated Simulated Ear, Nose, and Throat Scenarios: A Comparative Accuracy Study

Nazlim Hilal Taraf¹, Burcu Vural Camalan², Sumeyra Doluoglu³, Erhan Arslan⁴, Ahmet Ural⁵, Gulbin Demiroglu⁶, Atilla Halil Elhan⁷, Samet Özlügedik¹

- 1 M.D., University of Health Sciences Ankara Etlik City Hospital, Department of Otorhinolaryngology Head and Neck Surgery, Ankara, Turkey
- 2 M.D., University of Health Sciences Ankara Etlik City Hospital, Department of Otorhinolaryngology Head and Neck Surgery, Ankara, Turkey
- 3 M.D., University of Health Sciences Ankara Etlik City Hospital, Department of Otorhinolaryngology Head and Neck Surgery, Ankara, Turkey
- 4 Assoc.Prof., University of Health Sciences Ankara Etlik City Hospital, Department of Otorhinolaryngology Head and Neck Surgery, Ankara, Turkey
- 5 Professor Doctor, University of Health Sciences Ankara Etlik City Hospital, Department of Otorhinolaryngology Head and Neck Surgery, Ankara, Turkey
- 6 M.D., University of Health Sciences Ankara Etlik City Hospital, Department of Otorhinolaryngology Head and Neck Surgery, Ankara, Turkey
- 7 Professor Doctor, Ankara University Faculty of Medicine, Department of Department of Biostatistics, Ankara, Turkey
- 8 Professor Doctor, University of Health Sciences Ankara Etlik City Hospital, Department of Otorhinolaryngology Head and Neck Surgery, Ankara, Turkey

Cite this article as:

Hilal Taraf N, Vural Camalan B, Doluoglu S, Arslan E, Ural A, Demiroglu G, Halil Elhan A. Diagnostic Performance of ChatGPT-o1 and DeepSeek-V3 in Expert-Validated Simulated Ear, Nose, and Throat Scenarios: A Comparative Accuracy Study, Eur J Rhinol Allergy 2026;9(1): 1-9

E-mail:

erdiozdemir67@hotmail.com
Received: December 21, 2025
Accepted: February 05, 2026

DOI: 10.65396/ejra.1846059

Copyright@Author(s) -

Available

online at www.eurjrhinol.org

Content of this journal is

licensed under a Creative

Commons Attribution-

NonCommercial 4.0

International License.

Abstract

Objective: To compare the diagnostic accuracy of two advanced large language models (LLMs), ChatGPT-o1 and DeepSeek-V3, in expert-validated simulated otorhinolaryngology cases, and to assess subspecialty-specific performance and inter-rater agreement relative to that of human specialists.

Methods: A cross-sectional diagnostic accuracy study was conducted using 70 expert-validated clinical vignettes across five ENT subspecialties. Two academic otolaryngologists and two LLMs independently evaluated each case. All LLMs operated in deterministic mode (temperature = 0) with standardized single-pass prompting in isolated sessions. Diagnostic accuracy, inter-rater agreement (Cohen's κ), and subspecialty-specific performance were analyzed. A post hoc power analysis (Cohen's $h = 0.22$; $\alpha = 0.05$) assessed the ability to detect moderate effect sizes.

Results: Both LLMs achieved 90.0% diagnostic accuracy (63/70), with no significant difference between them ($p = 1.00$) and substantial inter-model agreement ($\kappa = 0.68$). Human evaluators achieved 97.1% and 92.9% accuracy, with fair inter-rater agreement ($\kappa = 0.26$). Subspecialty performance was highest in otology and pediatric ENT (100%) and in rhinology (92.3%), with greater variability in laryngology and head and neck surgery. Shared error patterns included overestimating malignancy in high-risk patients. A post hoc power analysis showed 78% power to detect moderate differences.



Conclusion: In controlled, vignette-based evaluations, ChatGPT-o1 and DeepSeek-V3 demonstrated diagnostic accuracy approaching expert-level performance across simulated ENT scenarios, with strong inter-model agreement and subspecialty-dependent variability. These findings highlight the potential of LLMs as diagnostic decision-support tools while underscoring the need for multimodal and real-world validation before clinical implementation.

Keywords: ChatGPT; DeepSeek; Diagnostic accuracy; Ear, nose, and throat; Large language models.

1. Introduction

Artificial intelligence (AI) has become increasingly integrated into clinical practice across multiple medical specialties, including radiology, dermatology, ophthalmology, and oncology. In otorhinolaryngology, AI-based systems have been applied to endoscopic image interpretation, voice and speech analysis, audiologic assessment, and the prediction of treatment outcomes in head and neck oncology. Collectively, these applications illustrate that ENT practice—historically dependent on visual, auditory, and narrative clinical data—represents a particularly suitable domain for evaluating AI-driven diagnostic support tools.

Alongside specialty-specific solutions, large language models (LLMs) have emerged as powerful general-purpose systems that can process clinical narratives, generate differential diagnoses, and synthesize guideline-based management strategies. Early studies found that ChatGPT achieved over 50% accuracy across all components of the United States Medical Licensing Examination (USMLE), exceeding the passing threshold for Steps 1–3.^{1,2} Within otolaryngology, LLM performance has varied by task complexity and subspecialty: ChatGPT showed moderate adherence to clinical practice guidelines,³ achieved diagnostic accuracy of 80%–82% in structured ENT case scenarios,^{4,5} and reached 57% accuracy on 2,576 board-style questions.⁶ Additional comparative evaluations with UpToDate, guideline-based tools, or multidisciplinary tumor board decisions have further highlighted both the promise and the limitations of LLMs in ENT practice.^{7–10} Given that otolaryngology relies heavily on structured symptom-driven narratives, LLMs are a plausible candidate for text-based diagnostic support.

Alongside the development of proprietary models, open-source LLM architectures have rapidly evolved. DeepSeek-V3, introduced in late 2024, is a transformer-based, multilingual model trained on extensive scientific and domain-specific corpora and optimized through reinforcement learning and context-aware reasoning strategies. Early independent benchmarks show performance comparable to or exceeding that of Western proprietary models across several scientific and medical tasks. However, its diagnostic performance has not been evaluated in otorhinolaryngology, and no direct comparison with ChatGPT has been reported. Because ENT diagnosis depends heavily on coherent symptom narratives that require precise contextual interpretation, determining DeepSeek-V3's performance in this setting remains an important open question.

Despite the growing literature on AI applications in otolaryngology, several methodological gaps persist. Most prior work evaluated

single LLMs in isolation,^{3–10} focused on limited question formats, or examined guideline adherence rather than diagnostic accuracy. Direct comparisons between LLM architectures are scarce, and no prior study has systematically evaluated a proprietary LLM (ChatGPT-o1) against a contemporary open-source model (DeepSeek-V3) across multiple ENT subspecialties using standardized, expert-validated clinical scenarios. Scenario-based evaluation, grounded in expert-confirmed diagnoses, serves as a valid proxy for outpatient diagnostic decision-making and is an appropriate foundational step before integrating multimodal clinical data such as endoscopic imaging, audiometry, or radiology. Therefore, the present study aimed to compare the diagnostic accuracy and inter-model agreement of ChatGPT-o1 and DeepSeek-V3 across 70 expert-validated, simulated otorhinolaryngology cases representing five major subspecialties. By addressing a clear methodological gap and establishing initial performance benchmarks, this study provides one of the first structured comparisons of two major LLM architectures in ENT clinical diagnosis. It offers a foundation for future multimodal AI evaluations in the field.

1. Methods

1.1. Study Design and Setting

This study was designed as a cross-sectional analysis of diagnostic accuracy conducted in May 2025. The term *cross-sectional* denotes a single-time-point evaluation of a fixed set of simulated clinical vignettes by all evaluators—two large language models (LLMs) and two academic otolaryngologists. The aim was to assess diagnostic accuracy and inter-rater agreement using standardized, expert-validated case scenarios representing common otorhinolaryngology presentations. Because no real patient data were involved, institutional ethical approval was not required.

1.2. Development of Simulated Clinical Cases

A total of 70 simulated otorhinolaryngology cases were constructed to reflect typical outpatient presentations across five subspecialties. These subspecialties (otology, head and neck surgery, rhinology, laryngology, and pediatric ENT) were selected because they represent the core clinical domains of otorhinolaryngology practice and account for the majority of outpatient encounters in academic and community settings. The distribution was as follows: otology (n=17), head and neck surgery (n=19), rhinology (n=13), laryngology (n=13), and pediatric ENT (n=8). Each vignette included patient demographics, chief complaint, history of present illness, relevant physical examination findings, and any essential ancillary investigations.

All scenarios were initially drafted by an otolaryngologist with ten years of clinical experience, using standard clinical patterns, widely used textbooks, and contemporary guideline recommendations. The objective was to ensure clinical realism and maintain an appropriate level of diagnostic complexity without oversimplification.

1.3. Expert Validation and Establishment of Reference Diagnoses

Each vignette was independently validated by a senior otolaryngologist

with more than 25 years of experience. The reviewer assessed clarity, clinical accuracy, and fidelity to real-world outpatient practice. Reference (final) diagnoses were assigned using established clinical guidelines and consensus disease definitions. Although multi-expert adjudication (e.g., the Delphi method) may further enhance diagnostic reliability, single-expert validation is commonly accepted in vignette-based diagnostic accuracy research and is acknowledged as a methodological limitation.

These expert-defined diagnoses served as the benchmark against which all human and LLM responses were evaluated.

1.4. LLM Evaluation Procedure and Model Selection

Two LLMs were evaluated: ChatGPT-o1 and DeepSeek-V3. At the time of data collection (May 2025), these were the most advanced publicly available text-only LLMs from their respective development paradigms—OpenAI's proprietary instruction-tuned model (ChatGPT-o1, released December 2024) and an emerging high-performance open-source architecture (DeepSeek-V3, released December 2024). Their inclusion reflects the state of text-based LLMs in mid-2025, establishing baseline performance benchmarks before the widespread clinical deployment of multimodal architectures such as GPT-4o or Gemini 1.5 Pro. This temporal context is important, as the findings represent text-based diagnostic capabilities that may serve as a reference point for future multimodal evaluations. Although subsequent model iterations may have been released after the study period, the versions evaluated were the most advanced publicly available text-only LLMs at the time of data collection and thus provide a valid benchmark for comparison.

Each vignette was submitted using an identical prompt:

"You are an experienced otorhinolaryngology consultant evaluating a patient in a clinical setting. Based on the following information, provide the single most likely diagnosis."

To prevent contextual contamination, each case was entered into a separate, newly initiated session. All model outputs were generated within 48 hours to avoid version drift.

Prompt delivery and model settings were standardized to ensure reproducibility. Each vignette was presented with a single fixed diagnostic prompt in a new, isolated session to prevent contextual carryover or hidden memory effects. Both LLMs operated in deterministic mode (temperature=0) with no stochastic sampling parameters (e.g., top-p, nucleus sampling), follow-up queries, or answer refinement. Automatic self-correction and chain-of-thought prompting were not enabled. Each case was evaluated once per model, without repeated runs or majority-vote aggregation, reflecting real-time clinical use, where consistent and stable outputs are essential. All model outputs were generated within 48 hours to maintain version consistency.

1.5. Human Expert Evaluation

Two academic otolaryngologists (with 15 and 20 years of experience) independently reviewed all 70 vignettes. They were blinded to one another's responses, to LLM outputs, and to the reference diagnoses.

The case order was individually randomized for each evaluator to minimize sequence bias and learning effects.

1.6. Outcome Measures

The primary outcome was diagnostic accuracy, defined as the proportion of correct diagnoses matching the expert-validated reference diagnosis.

Secondary outcomes included: analysis of incorrect responses; inter-rater agreement across all evaluator pairs (LLM–LLM, physician–physician, and LLM–physician); and subspecialty-specific diagnostic performance. All responses were dichotomized as correct or incorrect based on an exact diagnostic match.

1.7. Statistical Analysis

All statistical analyses were conducted using IBM SPSS Statistics v30.0. Categorical variables were compared using the Fisher–Freeman–Halton exact test, which is appropriate for multi-row contingency tables with small expected cell counts. Within-group paired comparisons (e.g., ChatGPT-o1 vs. DeepSeek-V3) were performed using the McNemar test. A two-tailed significance threshold of $p < 0.05$ was applied, as no directionality was hypothesized a priori.

Inter-rater agreement was assessed using Cohen's kappa (κ) with 95% confidence intervals. Kappa values were interpreted according to the Landis and Koch classification: <0.20: slight agreement, 0.21–0.40: fair agreement, 0.41–0.60: moderate agreement, 0.61–0.80: substantial agreement, 0.81–1.00: almost perfect agreement. Effect sizes and confidence intervals were reported for all comparative outcomes to characterize the precision of estimates.

1.8. Data Management and Quality Control

All vignette texts, reference diagnoses, and evaluator responses were recorded in a secure electronic database. A randomly selected 10% sample was subjected to double-entry verification to ensure data accuracy and integrity. Because the study was based entirely on fictional, anonymized clinical scenarios, no confidentiality or ethical concerns were present.

3. Results

3.1. Study Overview

Seventy expert-validated simulated otorhinolaryngology cases were analyzed by two large language models (ChatGPT-o1 and DeepSeek-V3) and two academic otolaryngologists. Case distribution across the five subspecialties was balanced: otology (24.3%), head and neck surgery (27.1%), rhinology (18.6%), laryngology (18.6%), and pediatric ENT (11.4%). Most scenarios represented mild-to-moderate diagnostic complexity, reflecting typical outpatient presentations (**Table 1, Figure 1**).

3.2. Overall Diagnostic Accuracy

Both ChatGPT-o1 and DeepSeek-V3 correctly diagnosed 63 of 70 cases, yielding an accuracy of 90.0% (95% CI, 80.5–95.9). No statistically

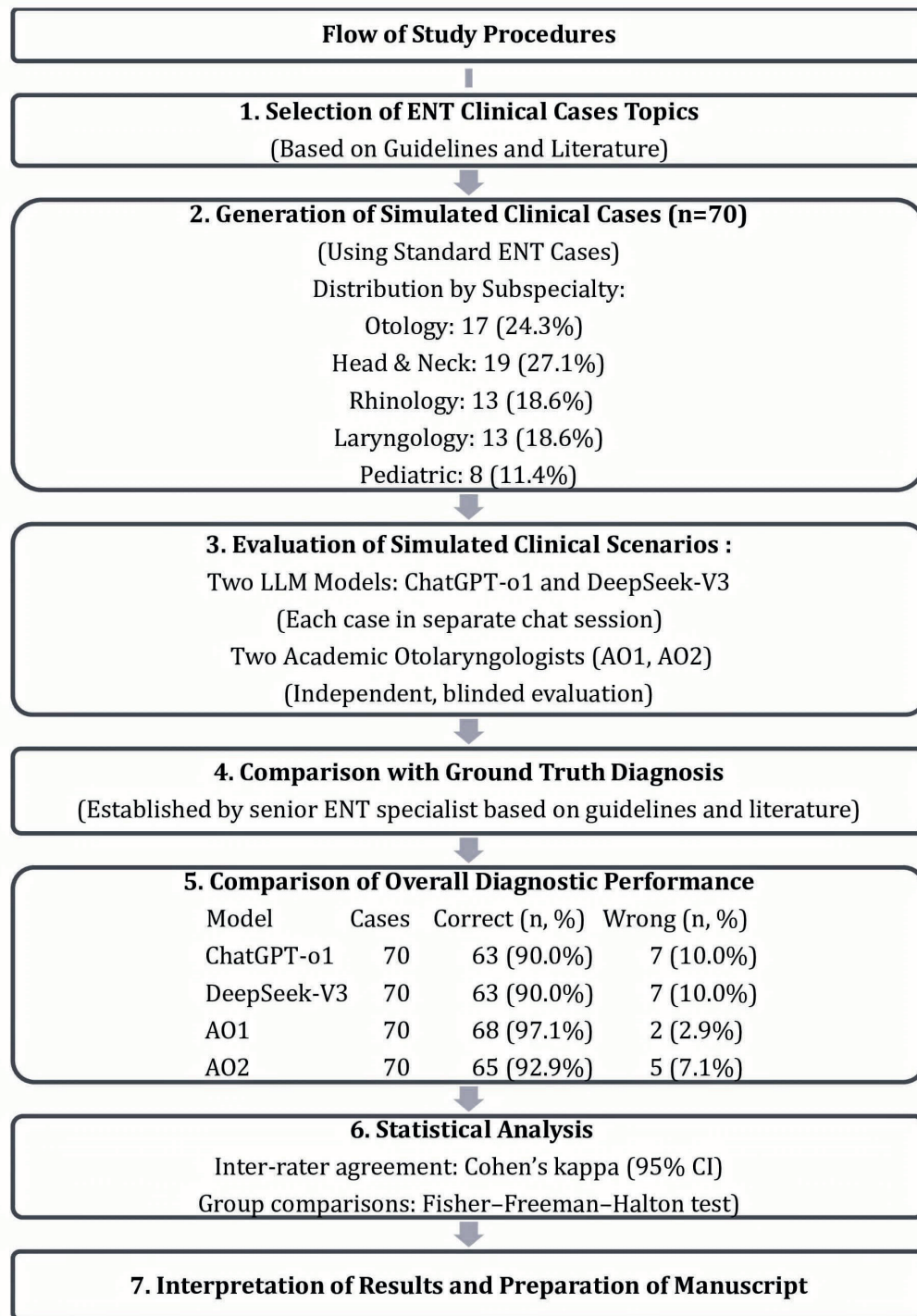


Figure 1. Study Flow Diagram

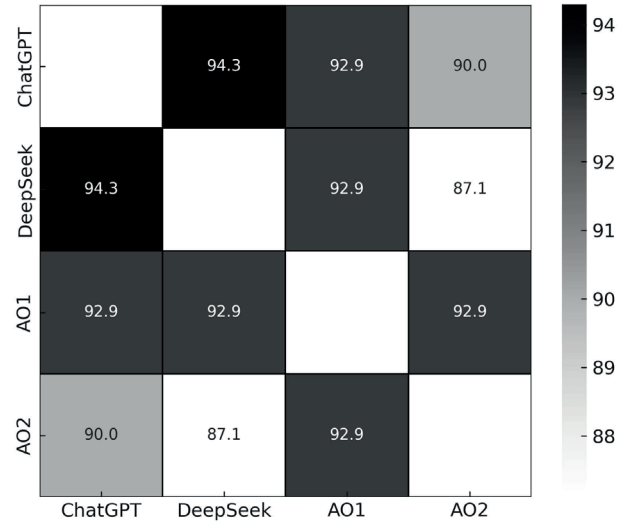
Flow diagram showing evaluation of 70 simulated otolaryngology cases across five subspecialties. Cases were independently evaluated by two LLMs (ChatGPT-o1 and DeepSeek-V3, separate sessions) and two academic otolaryngologists (blinded evaluation). Diagnostic accuracy: 90.0% (both LLMs), 97.1% (AO1), 92.9% (AO2), with no significant differences ($p > 0.05$). LLM = large language model; AO = academic otolaryngologist; ENT = Ear, Nose, and Throat; CI = confidence interval

Table 1. Characteristics of Simulated Clinical Scenarios (n=70)

Subspecialty Distribution	n (%)
Otology	17 (24.3)
Head and Neck Surgery	19 (27.1)
Rhinology	13 (18.6)
Laryngology	13 (18.6)
Pediatric ENT	8 (11.4)
Age groups	n (%)
Pediatric (0-18 years)	12 (17.1)
Adult (19-64 years)	43 (61.4)
Elderly (≥65 years)	15 (21.4)
Sex	n (%)
Male	45 (64.3)
Female	25 (35.7)
Case Complexity	n (%)
Simple (single clear diagnosis)	48 (68.6)
Moderate (multiple differential)	18 (25.7)
Complex (rare or multisystem)	4 (5.7)
Primary Presentation Type	n (%)
Inflammatory/ Infectious	18 (25.7)
Neoplastic	16 (22.9)
Functional/Neurologic	14 (20.0)
Traumatic	8 (11.4)
Congenital	6 (8.6)
Other	8 (11.4)
Associated Risk Factors Present	n (%)
Smoking history	12 (17.1)
Recent surgery/intubation	8 (11.4)
Pediatric age (<5 years)	7 (10.0)
Immunocompromised/Comorbidities	5 (7.1)
None specified	38 (54.3)

ENT = Ear Nose and Throat

significant difference was observed between the two models ($p = 1.00$; Fisher–Freeman–Halton test). Inter-model agreement was substantial, with a kappa coefficient of $\kappa = 0.68$ (95% CI, 0.39–0.97; $p < 0.001$). Only four cases were discordant between the two LLMs (**Table 2, Figure 2**).

**Figure 2. Agreement Matrix Among ChatGPT-o1, DeepSeek-V3 and Academic Otolaryngologists.**

Heatmap showing pairwise agreement rates (%) between all evaluators across 70 cases. Darker shades indicate higher agreement. Highest agreement between the two LLMs (94.3%), lowest between DeepSeek-V3 and AO2 (87.1%). All pairwise agreement rates exceeded 87%. LLM = large language model; AO = academic otolaryngologist.

3.3. Human Expert Performance

The two otolaryngologists demonstrated high diagnostic accuracy, with 97.1% (68/70) and 92.9% (65/70), respectively. Despite similar accuracy levels ($p = 0.37$, McNemar test), inter-rater agreement was fair ($\kappa = 0.26$; 95% CI, 0.00–0.70), indicating variability in clinical interpretation even among experienced specialists (**Table 3**).

3.4. Comparisons Between LLMs and Human Evaluators

Pairwise comparisons revealed no statistically significant differences between LLMs and human specialists. Diagnostic accuracy did not differ significantly between ChatGPT-o1 and AO1 ($p = 0.18$), ChatGPT-o1 and AO2 ($p = 0.75$), DeepSeek-V3 and AO1 ($p = 0.18$), or DeepSeek-V3 and AO2 ($p = 0.72$). Concordance exceeded 87% across all evaluator pairs, with the highest agreement observed between the two LLMs (94.3%) (**Figure 2**).

3.5. Subspecialty-Specific Accuracy

Diagnostic performance showed variability across subspecialties (**Table 4**). Both LLMs achieved perfect accuracy in otology and pediatric ENT cases, while rhinology cases were also diagnosed with high accuracy (92.3% for both models). In laryngology, ChatGPT-o1

Table 2. Diagnostic Concordance and Discordance Between ChatGPT-o1 and DeepSeek-V3 (n=70)

	DeepSeek Misdiagnosis	DeepSeek Correct	Total	
ChatGPT Misdiagnosis	5 (7.1%)	2 (2.9%)	7 (10.0%)	
ChatGPT Correct		2 (2.9%)	61 (87.1%)	63 (90.0%)
Total	7 (10.0%)	63 (90.0%)	70 (100%)	

Table 3. Diagnostic Accuracy and Agreement Between Two Academic Otolaryngologists (AO1 and AO2) (n=70)

	AO2 Misdiagnosis	AO2 Correct	Total
AO1 Misdiagnosis	1 (1.4%)	1 (1.4%)	2 (2.9%)
AO1 Correct	4 (5.7%)	64 (91.4%)	68 (97.1%)
Total	5 (7.1%)	65 (92.9%)	70 (100%)

Table 4. Diagnostic Accuracy of ChatGPT-o1 and DeepSeek-V3 by ENT Subspecialties

Subspecialty	n	ChatGPT Correct n (%)	DeepSeek Correct n (%)	p-value**	Kappa Coefficient
Otology	17	17 (100.0%) 95% CI: 80.5–100%	17 (100.0%) 95% CI: 80.5–100%	NA	NA
Head & Neck Surgery	19	16 (84.2%) 95% CI: 60.4–96.6%	14 (73.7%) 95% CI: 48.8–90.9%	0.5	$\kappa=0.68$, $p=0.002$
Rhinology	13	12 (92.3%) 95% CI: 64.0–99.8%	12 (92.3%) 95% CI: 64.0–99.8%	1.00	$\kappa=1.00$, $p<0.001$
Pediatric ENT	8	8 (100.0%) 95% CI: 80.5–100%	8 (100.0%) 95% CI: 80.5–100%	NA	NA
Laryngology	13	10 (76.9%) 95% CI: 46.2–95.0%	12 (92.3%) 95% CI: 64.0–99.8%	0.5	$\kappa=0.43$, $p=0.05$
Overall p-value*	70	$p=0.19$	$p=0.09$		

*: McNemar test. **: Fisher–Freeman–Halton exact test.

ENT = Ear Nose and Throat;

NA: Kappa could not be calculated because all cases were correctly classified by both models.

correctly identified 76.9% of cases, whereas DeepSeek-V3 achieved 92.3%. Head and neck surgery cases demonstrated a similar spread, with ChatGPT-o1 reaching 84.2% accuracy and DeepSeek-V3 73.7%. Despite these numerical differences, none of the subspecialty-level comparisons reached statistical significance ($p>0.05$) (**Figure 3**). Detailed confusion matrices for these challenging subspecialties are presented in **Tables 5 and 6**.

3.6. Error Pattern Analysis

Across all evaluators, 81.4% (57/70) of cases were correctly diagnosed by all four individuals. No case was misclassified by all evaluators simultaneously. Both LLMs shared five identical errors, most commonly involving overestimation of malignancy risk in smokers, suggesting a systematic tendency toward malignancy-biased reasoning in ambiguous scenarios. Human evaluators made fewer total errors,

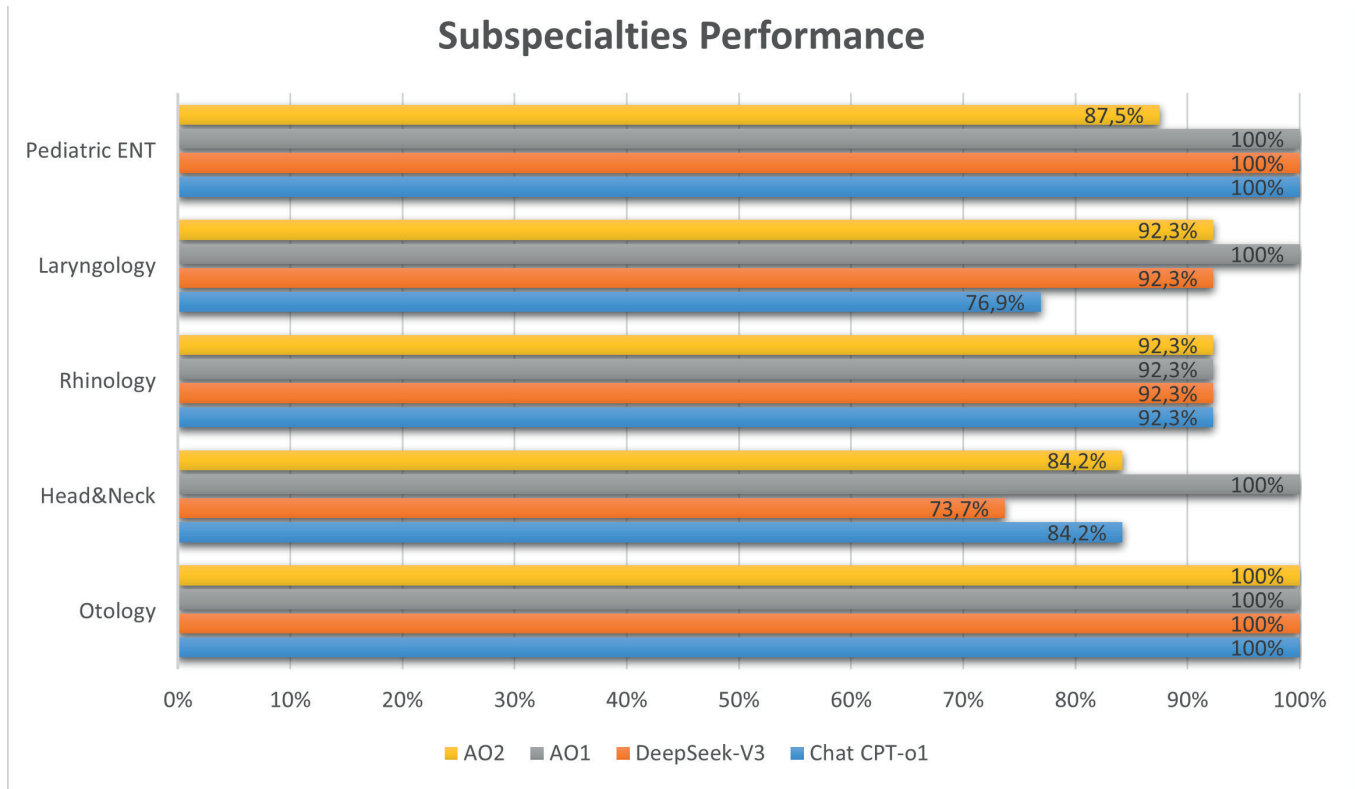


Figure 3. Diagnostic Accuracy Across ENT Subspecialties

Diagnostic accuracy rates across five subspecialties. All evaluators achieved 100% accuracy in otology and pediatric ENT. Greatest variation in head-neck surgery (DeepSeek: 73.7%) and laryngology (ChatGPT: 76.9%). Human experts maintained more consistent performance across subspecialties. ChatGPT-o1, DeepSeek-V3, AO1, AO2.

LLM = large language model; AO = academic otolaryngologist.

Table 5. Confusion Matrix of ChatGPT-o1 and DeepSeek-V3 in Head and Neck Surgery Cases (n=19)

	DeepSeek Misdiagnosis	DeepSeek Correct	Total
ChatGPT Misdiagnosis	3 (15.8%)	0 (0.0%)	3 (15.8%)
ChatGPT Correct	2 (10.5%)	14 (73.7%)	16 (84.2%)
Total	5 (26.3%)	14 (73.7%)	19 (100%)

Table 6. Confusion Matrix of ChatGPT-o1 and DeepSeek-V3 in Laryngology Cases (n=13)

	DeepSeek Misdiagnosis	DeepSeek Correct	Total
ChatGPT Misdiagnosis	1 (7.7%)	2 (15.4%)	3 (23.1%)
ChatGPT Correct	0 (0.0%)	10 (76.9%)	10 (76.9%)
Total	1 (7.7%)	12 (92.3%)	13 (100%)

but their mistakes were more heterogeneous, particularly when distinguishing congenital from acquired conditions in pediatric cases (**Supplementary File 1**).

3.7. Post Hoc Power Analysis

A post hoc power analysis using the observed effect size (Cohen's $h=0.22$; $\alpha=0.05$) demonstrated a statistical power of 0.78 for detecting medium-sized differences in diagnostic accuracy.

4. Discussion

The present study provides a contemporary evaluation of two advanced large language models, ChatGPT-o1 and DeepSeek-V3, in simulated otorhinolaryngology practice. Prior ENT-AI studies have reported diagnostic accuracies ranging from 57% to 82%,³⁻⁶ generally focusing on single-model assessments or limited task formats. In contrast, the current work offers one of the first direct comparisons between two distinct LLM architectures within standardized, clinically grounded scenarios, representing a clear advancement over earlier single-system investigations.⁷⁻⁹ The overall accuracy observed for both models aligns with recent evidence suggesting that modern LLMs can approach human-level diagnostic performance in controlled, text-based evaluations,^{10,11} although domain-optimized models have demonstrated even higher accuracy than both GPT-4 and expert clinicians in certain medical specialties.^{12,13} Importantly, high diagnostic accuracy achieved in text-based scenarios should not be interpreted as equivalent to real-world diagnostic safety, as routine otorhinolaryngology practice relies on multimodal clinical inputs and contextual judgment that cannot be fully captured by vignette-based assessments.

A further distinctive contribution of this study is the subspecialty-level analysis. High accuracy was observed in otology, pediatric ENT, and rhinology, whereas greater variability appeared in laryngology and head and neck surgery—patterns consistent with prior ENT literature.^{3,14,15} These differences likely reflect subspecialty-specific disparities in training data representation and the inherent diagnostic complexity of certain clinical domains. Both models also exhibited trends previously described, such as over-diagnosis of malignancy in high-risk patients, suggesting heightened sensitivity to specific clinical features but also highlighting the potential for systematic bias in ambiguous scenarios.⁴ Such tendencies further underscore the need for cautious interpretation when extrapolating text-based diagnostic performance to real-world clinical decision-making.

Inter-rater agreement analysis revealed substantial concordance between the two LLMs, indicating that different model architectures may converge toward similar diagnostic reasoning patterns when exposed to standardized inputs.¹⁶ Agreement among human evaluators was lower, reinforcing existing evidence that variability in clinical judgment persists even among experienced clinicians.³ Notably, the relatively low inter-rater agreement between human experts does not indicate incorrect clinical performance, but rather reflects differences in diagnostic reasoning and interpretative approaches despite comparable overall accuracy. Non-significant between-group

differences should therefore be interpreted cautiously, as the study was reasonably powered to detect moderate effect sizes but may not have been sufficient to identify small differences in diagnostic performance. Recent studies have further explored the role of LLMs in head and neck cancer decision-making, including guideline-based staging, treatment planning, and comparisons with multidisciplinary tumor boards.^{17,18} For example, ChatGPT demonstrated 72% concordance with tumor board recommendations in head and neck cancer cases,¹⁷ and both ChatGPT and DeepSeek showed similar adherence to guideline-based staging and treatment planning frameworks.¹⁸ These findings support the potential utility of LLMs as adjunctive tools within multidisciplinary care pathways, rather than as replacements for clinician-led decision-making.

Beyond methodological considerations, the clinical translation of LLMs into otorhinolaryngology practice will require robust regulatory validation and safety frameworks. In the United States and the European Union, AI-based diagnostic support tools are subject to oversight by the Food and Drug Administration (FDA) and the European Medicines Agency (EMA), respectively, with regulatory classification dependent on intended use and level of clinical autonomy. Current frameworks emphasize prospective clinical validation, post-market surveillance, and transparent algorithmic explainability prior to deployment in patient care settings. Until such validation pathways are completed, LLMs should be regarded as supplementary educational or decision-support tools rather than autonomous diagnostic systems, and their use must remain under direct clinician supervision.

4.1. Limitations

This study has several limitations that should be considered when interpreting its findings. First, the evaluation relied exclusively on text-based clinical vignettes, which, despite careful construction and expert validation, do not capture the inherently multimodal nature of real-world otorhinolaryngology practice. In routine clinical settings, diagnostic decision-making is informed by endoscopic imaging, audiologic testing, radiologic studies, and voice recordings, and the performance of text-only large language models may differ when applied to such multimodal inputs supported by newer AI architectures. Similar concerns have been raised regarding the clinical translation of text-based models, particularly in relation to hallucination risk and incomplete data representation.³

Second, the use of simulated vignettes limits ecological validity. Although standardized and grounded in evidence-based clinical patterns, vignettes cannot fully reflect the complexity, variability, and contextual nuances of real patient encounters. In addition, reference diagnoses were established by a single senior otolaryngologist rather than through multi-expert adjudication or Delphi-based consensus. While this approach is consistent with prior vignette-based ENT-AI studies, reliance on a single adjudicator may introduce subjectivity, particularly in diagnostically ambiguous subspecialties such as head and neck surgery and laryngology, and may therefore limit diagnostic reliability.

Third, the study cohort consisted of 70 cases representing common

outpatient presentations, with rare, highly complex, or multisystem conditions intentionally excluded. Although this sample size exceeds that of several previous ENT–AI investigations, it does not encompass the full clinical spectrum of otorhinolaryngology practice, thereby limiting generalizability to more specialized or high-complexity clinical settings.

Finally, established clinical resources such as UpToDate continue to provide more consistently accurate and evidence-based recommendations,⁹ underscoring that large language models should currently be regarded as supportive decision-assistance tools rather than autonomous diagnostic systems. Collectively, these limitations highlight the need for future prospective validation studies incorporating real patient data, multimodal inputs, and multi-expert adjudication frameworks prior to clinical deployment.

5. Conclusion

This study demonstrates that ChatGPT-o1 and DeepSeek-V3 can achieve diagnostic performance approaching that of experienced otolaryngologists in simulated clinical scenarios. By providing the first direct comparison between two distinct LLM architectures, together with subspecialty-specific performance evaluation and

inter-rater agreement analysis, the study offers unique insights into model behavior across different ENT domains. Future investigations incorporating multimodal clinical data, real patient encounters, and multi-expert validation frameworks will be essential to determine the clinical applicability and ensure the safe and effective integration of large language models into otorhinolaryngology practice.

Declaration of Interest Statement

Conflict of Interest: The authors declare no conflicts of interest.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Ethics Approval: Formal ethics committee approval was not required as this study involved only simulated clinical scenarios without real patient data or human subject involvement.

Data Availability: The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
2. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? *JMIR Med Educ*. 2023;9:e45312.
3. Tessler I, Wolfovitz A, Alon EE, et al. ChatGPT's adherence to otolaryngology clinical practice guidelines. *Eur Arch Otorhinolaryngol*. 2024;281(7):3829–3834.
4. Teixeira-Marques F, Medeiros N, Nazaré F, et al. Exploring the role of ChatGPT in clinical decision-making in otorhinolaryngology: a ChatGPT designed study. *Eur Arch Otorhinolaryngol*. 2024;281(4):2023–2030.
5. Vaira LA, Lechien JR, Abbate V, et al. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. *Otolaryngol Head Neck Surg*. 2023;170(6):1492–1503.
6. Hoch CC, Wollenberg B, Lüers JC, et al. ChatGPT's quiz skills in different otolaryngology subspecialties. *Eur Arch Otorhinolaryngol*. 2023;280(9):4271–4278.
7. Buhr CR, Smith H, Huppertz T, et al. Assessing unknown potential—quality and limitations of different large language models in the field of otorhinolaryngology. *Acta Otolaryngol*. 2024;144(3):237–242.
8. Makhoul M, Melkane AE, Khoury PE, Hadi CE, Matar N. A cross-sectional comparative study: ChatGPT 3.5 versus diverse levels of medical experts in the diagnosis of ENT diseases. *Eur Arch Otorhinolaryngol*. 2024 May;281(5):2717–2721.
9. Karimov Z, Allahverdiyev I, Agayarov OY, Demir D, Almuradova E. ChatGPT vs UpToDate: comparative study of usefulness and reliability of Chatbot in common clinical presentations of otorhinolaryngology-head and neck surgery. *Eur Arch Otorhinolaryngol*. 2024 Apr;281(4):2145–2151.
10. Wang L, Li J, Zhuang B, et al. Accuracy of large language models when answering clinical research questions: systematic review and network meta-analysis. *J Med Internet Res*. 2025;27:e64486.
11. McDuff D, Schaeckermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. *Nat Mach Intell*. 2025;7:155–165.
12. Goh E, Gallo R, Hom J, et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Netw Open*. 2024 Oct 1;7(10):e2440969.
13. Long C, Subburam D, Lowe K, et al. ChatENT: augmented large language model for expert knowledge retrieval in otolaryngology. *Otolaryngol Head Neck Surg*. 2024;171(3):1042–1051.
14. Bellingher JR, Kwak MW, Ramos GA, Mella JS, Mattos JL. Quantitative Comparison of Chatbots on Common Rhinology Pathologies. *Laryngoscope*. 2024 Oct;134(10):4225–4231.
15. Qu RW, Qureshi U, Petersen G, Lee SC. Diagnostic and Management Applications of ChatGPT in Structured Otolaryngology Clinical Scenarios. *OTO Open*. 2023 Aug 22;7(3):e67.
16. Lorenzi A, Pugliese G, Maniaci A, et al. Reliability of large language models for advanced head and neck malignancies management: comparison between ChatGPT 4 and Gemini Advanced. *Eur Arch Otorhinolaryngol*. 2024;281(9):5001–5006.
17. Pamuk E, Bilen YE, Külekçi Ç, Kuşçu O. ChatGPT-4 vs. multi-disciplinary tumor board decisions for the therapeutic management of primary laryngeal cancer. *Acta Otolaryngol*. 2025 Aug;145(8):714–719.
18. Vural Camalan B, Doluoglu S, Taraf NH, Gunay MM, Ozlugedik S. ChatGPT versus DeepSeek in head and neck cancer staging and treatment planning: guideline-based study. *Eur Arch Otorhinolaryngol*. 2025 Sep;282(9):4815–4824.