AMERICAN ACADEMY OF
OTOLARYNGOLOGY–
HEAD AND NECK SURGERY
FOUNDATION

Otolaryngology–
Head and Neck Surgery
2025, Vol. 00(00) 1–9
© 2025 American Academy of
Otolaryngology–Head and Neck
Surgery Foundation.
DOI: 10.1002/ohn.70070
http://otojournal.org
WILEY

# Comparative Accuracy, Stability, and Correctability of Large Language Models in Otolaryngology and Pharmacovigilance

**Filippo Bruno, MPharm[1]*, Lise Sogalow, MD[1]*, Bertrand Blankert, MPharm, PhD[2], and Jerome R. Lechien, MD, PhD, FACS[1,3,4,5]**

## Abstract

*Objective.* To compare the clinical and pharmacovigilance performance, stability, and correctability of 3 large language models (LLMs) in otolaryngology outpatient care.

*Study design.* Prospective case series.

*Setting.* Multicenter University Hospitals.

*Methods.* Consecutive adults (August-October 2024) with established primary diagnoses were entered into ChatGPT-4o, Gemini-1.5-Pro, and Claude-3.5-Sonnet using only history and physical examination findings (no complementary tests) via standardized prompts. Two blinded otolaryngologists rated clinical accuracy with the Artificial Intelligence Performance Instrument (AIPI); 2 blinded pharmacists rated pharmacological information on a 5-point Likert scale. Errors were fed back to models and all cases were re-queried one month later. Interrater reliability used ICC; stability used Cronbach's $\alpha$. Group differences used Kruskal-Wallis.

*Results.* Fifty-one patients with 60 diagnoses across otolaryngology subspecialties were consecutively recruited (38 females (74.5%); mean age of 42.4 ± 17.4 years). All LLMs recommended significantly more additional examinations than practitioners ($P = .001$), with a significant increase of the number of recommended additional examinations after regenerated inputs for ChatGPT-4o and Claude-3.5-Sonnet, respectively. Claude-3.5-Sonnet and ChatGPT-4o outperformed Gemini-1.5-Pro for AIPI-clinical management ($P = .001$) and pharmacovigilance findings ($P = .001$). The physicians (ICC = 0.853) and the pharmacists (ICC = 0.991) demonstrated an almost perfect interrater reliability. All LLMs demonstrated an almost perfect clinical stability ($\alpha = 0.831$-$0.856$), though human feedback did not significantly reduce misdiagnosis rates in subsequent interactions.

*Conclusion.* In outpatient ENT cases using clinical features alone, ChatGPT-4o and Claude-3.5-Sonnet deliver higher clinical and pharmacovigilance performance than Gemini-1.5-Pro, with almost perfect interrater reliability and stable outputs. Re-querying after feedback did not improve accuracy, questioning short-term correctability.

## Keywords

accuracy, artificial intelligence, head neck surgery, Large Language Model, otolaryngology, otorhinolaryngology, stability

Received June 3, 2025; accepted October 22, 2025.

The development of artificial intelligence-powered large language models (LLMs) is emerging in medicine and surgery with easy and full access to populations.[1] The accessibility and popularity of LLMs may encourage patients to use them for education before or after a medical consultation,[2,3] while practitioners may consider LLMs as an adjunctive clinical tool for the management of complex clinical cases.[4,5] Regarding the widespread use of LLMs, it is important to evaluate their performance in providing medical information to both patient and physician, including diagnosis and treatment accuracy and stability. The number of studies evaluating the accuracy of LLMs is increasing in otolaryngology-head and neck surgery,[6] but most of them focus on ChatGPT,[6] with a few studies[7-9] comparing ChatGPT to

[1]Department of Surgery, Research Institute for Language Science and Technology, University of Mons, Mons, Belgium
[2]Laboratory of Pharmaceutical Analysis, Faculty of Medicine and Pharmacy, Research Institute for Health Sciences and Technology, University of Mons-UMONS, Mons, Belgium
[3]Department of Otolaryngology–Head and Neck Surgery, Foch Hospital, School of Medicine, Paris Saclay University, Phonetics and Phonology Laboratory (UMR 7018 CNRS, Université Sorbonne Nouvelle/Paris 3), Paris, France
[4]Department of Otolaryngology, Elsan Polyclinic de l'Atlantique, Poitiers, France
[5]Department of Otolaryngology–Head and Neck Surgery, CHU Saint-Pierre, Brussels, Belgium

*Dr Sogalow and Mr Bruno have similarly contributed to the study design and can be joined as co-first authors. Pr Blankert and Pr Lechien have similarly contributed to the study design and can be joined as co-last authors.

**Corresponding Author:**
Jerome R. Lechien, MD, PhD, FACS, Department of Otolaryngology–Head and Neck Surgery, Elsan Hospital, Poitiers, France.
Email: Jerome.Lechien@umons.ac.be

another LLM. In the same vein, studies rarely evaluated the stability of LLM outputs over time,[6] which is an important parameter in determining the performance of an LLM. The manufacturers of these models commonly present the reinforcement learning from Human Feedback as an important technique to align the outputs of LLMs, which suggest potential correction of LLM's mistakes after human feedback. However, to the best of our knowledge, whether such mechanisms translate into real-time improvements at the user level has never been investigated in otolaryngology-head and neck surgery.

The aim of this prospective longitudinal study was to investigate the accuracy, stability, and correctability properties of 3 widely used LLMs in the management of cases in otolaryngology-head and neck surgery.

## Methods

### Patients and Setting

Adult patients consulting in the departments of Otolaryngology-Head and Neck Surgery of the Dour Medical Center (Dour, Belgium) and Saint-Pierre University Hospital (Brussels, Belgium) were consecutively recruited from August 2024 to October 2024. All patients were recruited in outpatient clinics during routine ENT consultations. The following data were prospectively collected by the senior otolaryngologist (JRL): age, gender, clinical history, comorbidities, symptoms, clinical examination findings, ongoing treatments, additional examinations, primary diagnosis, and treatments (eg, medication, doses, and duration). All patients underwent a complete otolaryngological examination, including otoscopy, nasolaryngoscopy, oral cavity examination, and neck palpation. Only patients with complete information, including an established primary diagnosis, were recruited. The exclusion criteria consisted of incomplete data records (ongoing disease diagnosis exploration), non-native French-speaking patients, and those who declined to participate. We used consecutive sampling to recruit patients across all ENT subspecialties during a fixed recruitment window. No formal power calculation was performed, given the exploratory design of this study.

The study was approved by the institutional review board of CHU Saint-Pierre (CHUSP, n°BE0762023230708). All participants prospectively enrolled were aged 18 years or older and provided written informed consent.

This study adhered to the STROBE guidelines for observational studies to ensure transparency and replicability of our findings.[10]

### Data Collection and Anonymization

Data from complete medical records were anonymized and systematically entered into the Chatbot Generative Pre-trained Transformer (ChatGPT-4o; OpenAI), Gemini-1.5-Pro (Google) and Claude-3.5-Sonnet (Anthropic) interfaces for primary and differential diagnoses, management plans, and treatments. Based on the treatment findings, LLMs were interrogated for drug posology (recommended doses and duration), and the 5 most prevalent adverse events. The primary researcher (FB) used the following standardized sentences after the description of each case: *What are your primary and differential diagnoses?*; *What are your additional examinations to find the diagnosis (management plan)?*; *What are your treatment(s) for the primary diagnosis?*; *What are the recommended drug doses and duration?*; and *What are the 5 most common adverse events of this medication?*. The responses of the LLMs were collected in a database.

### Large Language Model Performance Analysis

The database with the LLM outputs was submitted to 2 sets of judges (otolaryngologists and pharmacists). Consistent with previous study,[2] practitioner judges used all possible national and international consensus statements or guidelines to establish the appropriate management. The performance of LLMs in the management of otolaryngological cases was assessed in a blinded manner by 2 physicians using the Artificial Intelligence Performance Instrument (AIPI), which is a validated and reliable instrument used to assess the performance and consistency of artificial intelligence chatbots.[11]

AIPI consists of 9 items assessing the ability of LLMs to consider medical and surgical history; symptoms; physical examination; diagnosis; additional examinations; management plan; and treatments in the overall management of real clinical cases. AIPI is subdivided into the following sub-scores associating 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).[11]

Regarding the pharmacological analysis, the 2 pharmacist judges independently assessed the LLM response through a 5-point Likert scale ranging from 1 (*inconsistent information*) to 5 (*perfectly consistent information*). The lists of the most prevalent adverse events of medications were established with the *Centre Belge d'Information Pharmacothérapeutique* (CBIP) database, which is the national pharmacovigilance guide for all medications.

### Stability and Correctability of Large Language Models

The errors of LLMs were systematically corrected through explanations entered into the API with the correct response. Each medical record data and information were re-entered into the APIs of the 3 LLMs one month after the initial task, maintaining the same sentences and information. A second analysis of accuracy was performed by practitioner and pharmacist judges using the same tools (AIPI and 5-point Likert scale).

Similar to the first round, the errors of LLMs were collected.

### Statistical Analyses

Statistical analyses were performed using the Statistical Package for the Social Sciences for Windows (SPSS version 29.0; IBM Corp.). 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. Additional examinations indicated by otolaryngologists and the 3 LLMs were coded with predefined numbers in a matrix, which facilitated the evaluation of consistency between the physician's findings versus those of the LLMs. The mean numbers of additional examinations per patient indicated by LLMs and practitioners were compared with the Kruskal-Wallis test.

The comparison of AIPI and 5-point Likert scale scores across LLMs was carried out with the Kruskal-Wallis test. The proportions of errors were compared across LLMs with chi-square tests. The expert interrater reliability was assessed for the AIPI and 5-point Likert scale scores with the intraclass correlation coefficient (ICC). The stability of LLM outputs was evaluated through Cronbach-$\alpha$. The consistency was based on the Landis and Koch classification: <0.20 slight agreement; 0.21 to 0.40 fair agreement; 0.41 to 0.60 moderate agreement; 0.61 to 0.80 substantial agreement; 0.81 to 1.00 almost perfect agreement. This interpretation was applied to interrater reliability for the AIPI (ICC) and to stability analyses (Cohen's kappa values). A significance level of $P < .05$ was used.

## Results

Fifty-one patients were consecutively recruited, including 38 (74.5%) females and 13 (25.5%) males (**Figure 1**). The mean age was 42.4 ± 17.4 years. The primary diagnoses are described in **Table 1**. Sixty diagnoses were made, with 9 patients reporting 2 disorders. There were 29 (48.3%) laryngological or head and neck cases, 19 (31.7%) rhinological cases, and 12 (20.0%) otological cases.

### Clinical Performances of Large Language Models

At baseline, ChatGPT-4o, Gemini-1.5-Pro, and Claude-3.5-Sonnet recommended mean numbers of additional

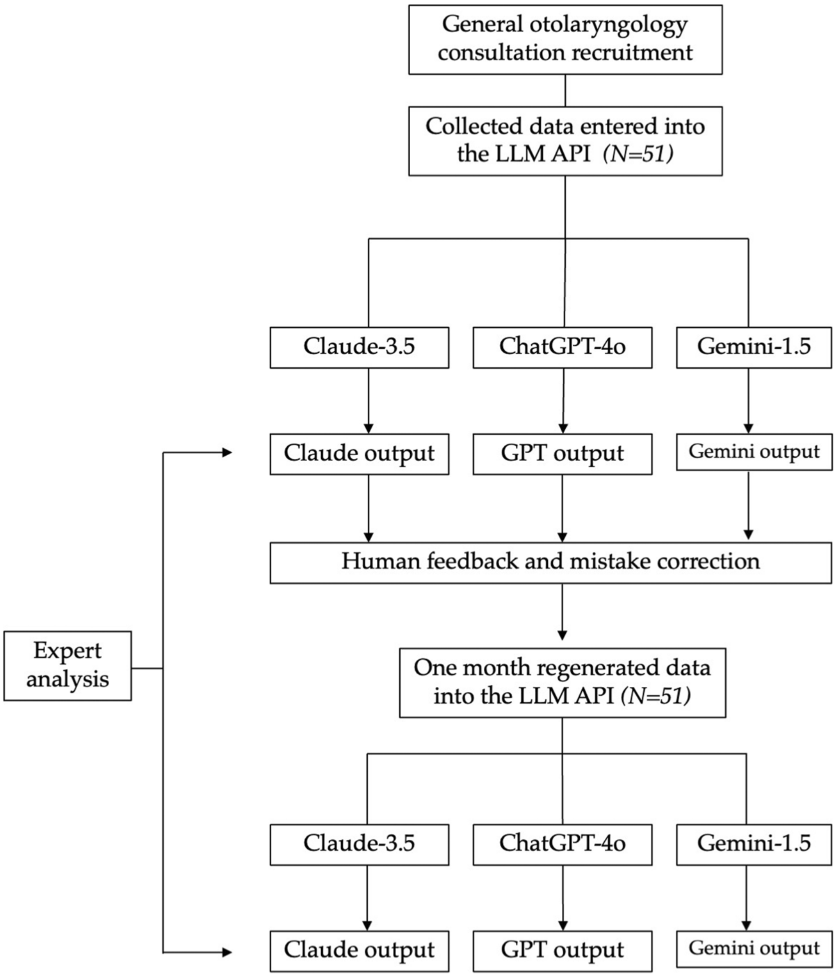
General otolaryngology consultation recruitment
Collected data entered into the LLM API (N=51)
Claude-3.5
ChatGPT-4o
Gemini-1.5
Claude output
GPT output
Gemini output
Human feedback and mistake correction
Expert analysis
One month regenerated data into the LLM API (N=51)
Claude-3.5
ChatGPT-4o
Gemini-1.5
Claude output
GPT output
Gemini output

**Figure 1.** Chart flow.

**Table 1.** Demographic and Clinical Findings

| Outcomes | Patients (n = 51) |
|---|---|
| Age (mean, SD) | 42.4 ± 17.4 |
| Gender (N, %) | |
| Female | 38 (74.5) |
| Male | 13 (25.5) |
| Primary and secondary diagnoses | |
| Rhinology | |
| Postviral olfactory dysfunction | 4 (6.6) |
| Chronic rhinosinusitis without nasal polyps | 4 (6.6) |
| Allergic rhinitis | 2 (3.3) |
| Acute rhinosinusitis | 2 (3.3) |
| Odontogenic maxillary rhinosinusitis | 2 (3.3) |
| Rhinitis medicamentosa | 1 (1.6) |
| Vasomotor rhinitis | 1 (1.6) |
| Nasal cavity foreign body | 1 (1.6) |
| Recurrent anterior epistaxis | 1 (1.6) |
| Nasopharyngeal reflux disease | 1 (1.6) |
| Acute post-viral anosmia | 1 (1.6) |
| Laryngology-head and neck | |
| Laryngopharyngeal reflux disease | 10 (19.6) |
| Laryngopharyngeal hypersensitivity syndrome | 2 (3.3) |
| Recurrent aphthous stomatitis | 1 (1.6) |
| Lingual candidiasis | 1 (1.6) |
| Viral pharyngitis | 1 (1.6) |
| Gastroesophageal reflux disease | 1 (1.6) |
| Salivary lithiasis | 1 (1.6) |
| Chlamydia pneumoniae tracheolaryngitis | 1 (1.6) |
| Retrograde cricopharyngeal dysfunction | 1 (1.6) |
| Essential laryngeal tremor | 1 (1.6) |
| Spasmodic dysphonia | 1 (1.6) |
| Vocal cord granuloma | 1 (1.6) |
| Asthmatic chronic cough | 1 (1.6) |
| Hyperthyroidism | 1 (1.6) |
| Secondary oropharyngeal syphilis | 1 (1.6) |
| Vocal fold scars | 1 (1.6) |
| Idiopathic bilateral vocal cord paralysis | 1 (1.6) |
| Idiopathic left facial paresis | 1 (1.6) |
| Otology | |
| Eustachian tube dysfunction | 2 (3.3) |
| Chronic otitis media | 1 (1.6) |
| External otitis | 1 (1.6) |
| Suppurative otitis media | 1 (1.6) |
| Left tonsil abscess | 1 (1.6) |
| Recurrent cholesteatoma | 1 (1.6) |
| Fungal otitis externa | 1 (1.6) |
| Cerumen impaction | 1 (1.6) |
| Herpes zoster (external ear duct) | 1 (1.6) |
| Refractory external otitis with stenosis | 1 (1.6) |
| Sudden sensorineural hearing loss | 1 (1.6) |

Abbreviation: SD, standard deviation.

examinations of 2.29 ± 1.38, 2.80 ± 1.39, and 2.53 ± 1.55 per patient, respectively, all being significantly higher compared to the mean number of additional examinations recommended by practitioners (1.04 ± 1.13; $P = .001$). After regenerated inputs, the mean number of additional examinations recommended by ChatGPT-4o significantly increased to 2.80 ± 1.67 ($P = .046$). A similar observation was found for Claude-3.5-Sonnet, with a significant increase in the mean number of additional examinations per patient from 2.53 ± 1.55 to 3.37 ± 1.15 ($P = .001$).

The mean AIPI item, sub-, and total scores of ChatGPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro are described in **Table 2**. ChatGPT-4o and Claude-3.5-Sonnet demonstrated significantly higher performance scores for the patient feature and treatment subscores, and the total AIPI compared to Gemini-1.5-Pro. The 3 LLMs were comparable for the differential and primary diagnosis scores (**Table 2**). The 2 physicians demonstrated an almost perfect interrater reliability for the AIPI analysis (ICC = 0.853; 95% CI = 0.816-0.885). The proportion of correct primary diagnoses and plausible differential diagnoses are reported in **Table 3**. In terms of proportions of adequate responses, LLM demonstrated significant differences only for treatment recommendations with Claude-3.5-Sonnet and ChatGTP-4o outperforming Gemini-1.5-Pro.

## *Pharmacovigilance Performances of Large Language Models*

The mean pharmacovigilance scores of ChatGPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro were 3.90 ± 0.49, 3.85 ± 0.61, and 3.09 ± 0.71, respectively. ChatGPT-4o and Claude-3.5-Sonnet outperformed Gemini-1.5-Pro in terms of pharmacovigilance findings ($P = .001$). The 2 pharmacists demonstrated an almost perfect interrater reliability for the 5-point Likert scale score (ICC = 0.991; 95% CI = 0.987 to 0.993).

## *Stability of Large Language Models*

The stability of LLMs through regenerated inputs is reported in **Table 3**. Regarding clinical management (AIPI scores), ChatGPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro demonstrated comparable almost perfect stability through regenerated inputs with Cronbach's $\alpha$ ranging from 0.831 to 0.856. The stability analysis of regenerated inputs for pharmacovigilance demonstrated high stability for ChatGPT-4o and Claude-3.5-Sonnet (**Table 4**). Gemini-1.5-Pro demonstrated stability of outputs corresponding to slight agreement, which was related to the high number of items scored as 0 by the judges because Gemini-1.5-Pro provided no response to the entered questions at some time points during the analysis.

## *Errors, Human Feedback, and Reinforcement Learning*

The detailed misdiagnoses at baseline and after human feedback are described in **Table 5**. Gemini-1.5-Pro

**Table 2.** Clinical Performances of Large Language Models

| AIPI management outcomes | ChatGPT-4.0 | Gemini-1.5 | Claude-3.5 | *P*-value |
|---|---|---|---|---|
| 1. Consideration of medical history (/2) | 1.8 ± 0.4 | 1.6 ± 0.6 | 1.9 ± 0.3 | .046 |
| 2. Consideration of symptoms (/2) | 1.9 ± 0.3 | 1.7 ± 0.5 | 1.8 ± 0.3 | .017 |
| 3. Consideration of physical examination findings (/2) | 1.8 ± 0.4 | 1.7 ± 0.5 | 1.8 ± 0.3 | NS |
| Patient feature score (/6) | 5.5 ± 0.8 | 4.9 ± 1.3 | 5.5 ± 0.6 | .036 |
| 4. Differential diagnosis (/3) | 2.6 ± 0.5 | 2.3 ± 0.5 | 2.5 ± 0.6 | NS |
| 5. Primary diagnosis (/3) | 2.7 ± 0.5 | 2.5 ± 0.7 | 2.70 ± 0.5 | NS |
| 6. Management plan (/1) | 0.8 ± 0.4 | 0.7 ± 0.4 | 0.8 ± 0.3 | NS |
| Diagnosis score (/7) | 6.0 ± 1.0 | 5.5 ± 1.4 | 6.0 ± 1.0 | NS |
| 7. Additional examinations (/3) | 1.9 ± 0.6 | 1.7 ± 0.6 | 1.8 ± 0.7 | NS |
| 8. The most relevant additional examination (/1) | 0.4 ± 0.4 | 0.3 ± 0.4 | 0.3 ± 0.4 | NS |
| Additional examination score (/4) | 2.3 ± 0.9 | 1.9 ± 0.8 | 2.1 ± 1.0 | NS |
| 9. Treatment (/3) | 2,0 ± 0.7 | 1.1 ± 0.4 | 1.9 ± 0.5 | .001 |
| AIPI total score (/20) | 15.8 ± 2.5 | 13.5 ± 3.0 | 15.5 ± 2.4 | .001 |

Abbreviations: AIPI, artificial intelligence performance instrument; NS, nonsignificant.

**Table 3.** Proportions of Correct or Adequate Clinical Responses

| | ChatGPT-4o | Gemini-1.5 | Claude-3.5 | |
|---|---|---|---|---|
| AIPI outcomes | n = 51 | n = 51 | n = 51 | *P*-value |
| Primary diagnosis (N (%)) | | | | |
| Correct | 36 (70.6) | 32 (62.8) | 41 (80.4) | |
| Plausible | 11 (21.6) | 10 (19.6) | 7 (13.7) | NS |
| Not plausible | 3 (5.9) | 9 (17.6) | 3 (5.9) | |
| Absent | 1 (2.0) | 0 (0) | 0 (0) | |
| Differential diagnosis | | | | |
| Correct | 28 (54.9) | 20 (39.2) | 33 (64.7) | |
| Plausible | 18 (35.3) | 26 (51.0) | 13 (25.5) | NS |
| Not plausible | 5 (9.8) | 5 (9.8) | 5 (9.8) | |
| Absent | 0 (0) | 0 (0) | 0 (0) | |
| Relevant additional examination | | | | |
| Pertinent and necessary | 11 (21.6) | 4 (7.8) | 8 (15.7) | |
| Pertinent and not necessary | 28 (54.9) | 30 (58.8) | 25 (49.0) | NS |
| Pertinent, necessary and inadequate | 11 (21.6) | 15 (29.4) | 16 (31.4) | |
| Only inadequate examinations | 1 (2.0) | 2 (3.9) | 2 (3.9) | |
| Treatment | | | | |
| Pertinent and necessary | 14 (27.5) | 0 (0) | 16 (31.4) | |
| Pertinent and incomplete | 26 (51.0) | 13 (25.5) | 20 (39.2) | .001 |
| Association of pertinent/necessary and inadequate | 10 (19.6) | 33 (64.7) | 14 (27.5) | |
| No adequate strategy | 1 (2.0) | 5 (9.8) | 1 (2.0) | |
| Management plant (0-1) | | | | |
| Pertinent | 42 (82.4) | 36 (70.6) | 42 (82.4) | NS |
| Not pertinent | 9 (17.6) | 15 (29.4) | 9 (17.6) | |

Abbreviation: NS, nonsignificant.

made 27 primary misdiagnoses at baseline and 22 after human feedback (regenerated inputs). ChatGPT-4o made 16 misdiagnoses at baseline and 13 after regenerated inputs, while Claude-3.5-Sonnet made 14 misdiagnoses at both baseline and regenerated inputs. The number of misdiagnoses by ChatGPT-4o and Claude-3.5-Sonnet was significantly lower compared to Gemini-1.5-Pro ($P = .001$). Despite the human feedback, the number of misdiagnoses did not significantly change across regenerated inputs. Some new misdiagnoses appeared at the regenerated timepoint in all LLMs (**Table 5**).

## Discussion

The emergence of LLM chatbots has great promise for enhancing healthcare practice and patient information. However, many grey areas persist regarding their accuracy, stability, and particularly their ability to integrate corrective feedback over time (short-term correctability), which remains insufficiently explored in otolaryngology-head and neck surgery. To the best of our knowledge, this study is the most comprehensive one prospectively assessing the accuracy, stability, and correctability properties of 3 widely used LLMs, spanning findings from practitioner consultation to the pharmacist office.

**Table 4.** Stability Analysis

| | | 95% CI | |
|---|---|---|---|
| Stability outcomes | Crohnbach | Minimum | Maximum |
| AIPI | | | |
| ChatGPT-4o | 0.831 | 0.754 | 0.892 |
| Gemini-1.5-Pro | 0.856 | 0.791 | 0.908 |
| Claude-3.5-Sonnet | 0.855 | 0.789 | 0.907 |
| Pharmacovigilance (5-point Likert scale) | | | |
| ChatGPT-4o | 0.980 | 0.965 | 0.989 |
| Gemini-1.5-Pro | 0.100 | −0.576 | 0.486 |
| Claude-3.5-Sonnet | 0.986 | 0.976 | 0.992 |

Abbreviations: AIPI, artificial intelligence performance instrument; IC, confident intervalle.

The results of our analysis support the superiority of ChatGPT-4o and Claude-3.5-Sonnet over Gemini-1.5-Pro in terms of clinical management of real cases and output stability through regenerated inputs. Most studies available in the literature focused on ChatGPT-3.5, 4.0, and 4o for assessing accuracy of LLMs, which limits our comparison with the literature.[6,12] The accuracy rates of ChatGPT for providing management plans, differential diagnoses, and therapeutic regimens for real clinical cases substantially varied across studies depending on input features (clinical case descriptions, videos, clinical and pathological images), and ChatGPT versions.[12] The ChatGPT differential diagnosis accuracy ranges from 28.3% to 90%, with the lowest accuracy (28.3%) found when considering laryngology images,[13] and the highest rates for studies evaluating the differential diagnosis accuracy in clinical cases without image interpretation (63.5% to 90%).[11,14] For additional examinations,

**Table 5.** Misdiagnoses and Reinforcement Learning Property

| | Cohort | Misdiagnoses | | | | | |
|---|---|---|---|---|---|---|---|
| Diagnosis findings | Total number | ChatGPT-4o | re-ChatGPT-4o | Gemini-1.5 | re-Gemini-1.5 | Claude-3.5 | re-Claude-3.5 |
| Rhinology | | | | | | | |
| CRSNP | 4 (6.6) | 2 (12.5) | 2 (15.4) | 2 (7.4) | 3 (13.6) | 1 (7.1) | 2 (14.3) |
| Nasopharyngeal reflux disease | 1 (1.6) | 1 (6.3) | 1 (7.7) | 1 (3.7) | 1 (4.5) | 1 (7.1) | 1 (7.1) |
| Odontogenic maxillary rhinosinusitis | 2 (3.3) | 1 (6.3) | 1 (7.7) | 2 (7.4) | 0 (0) | 1 (7.1) | 1 (7.1) |
| Postviral olfactory dysfunction | 4 (6.6) | 0 (0) | 1 (7.7) | 1 (3.7) | 1 (4.5) | 0 (0) | 0 (0) |
| Acute rhinosinusitis | 2 (3.3) | 2 (12.5) | 1 (7.7) | 2 (7.4) | 1 (4.5) | 2 (14.3) | 0 (0) |
| Recurrent anterior epistaxis | 1 (1.6) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| Otology | | | | | | | |
| Eustachian tube dysfunction | 2 (3.3) | 0 (0) | 0 (0) | 2 (7.4) | 2 (9.1) | 1 (7.1) | 1 (7.1) |
| Refractory external otitis with stenosis | 1 (1.6) | 0 (0) | 0 (0) | 1 (3.7) | 0 (0) | 0 (0) | 0 (0) |
| Chronic otitis media | 1 (1.6) | 1 (6.3) | 1 (7.7) | 0 (0) | 1 (4.5) | 0 (0) | 1 (7.1) |
| External otitis | 1 (1.6) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| Suppurative otitis media | 1 (1.6) | 1 (6.3) | 1 (7.7) | 1 (3.7) | 1 (4.5) | 1 (7.1) | 1 (7.1) |
| Laryngology-Head and Neck | | | | | | | |
| Laryngopharyngeal reflux disease | 10 (19.6) | 4 (25.0) | 2 (15.4) | 5 (18.5) | 5 (22.7) | 3 (21.4) | 3 (21.4) |
| Viral pharyngitis | 1 (1.6) | 1 (6.3) | 0 (0) | 1 (3.7) | 1 (4.5) | 0 (0) | 1 (7.1) |
| Gastroesophageal reflux disease | 1 (1.6) | 1 (6.3) | 1 (7.7) | 1 (3.7) | 1 (4.5) | 1 (7.1) | 0 (0) |
| Chlamydia Pneumoniae tracheolaryngitis | 1 (1.6) | 1 (6.3) | 0 (0) | 1 (3.7) | 1 (4.5) | 1 (7.1) | 1 (7.1) |
| Secondary oropharyngeal syphilis | 1 (1.6) | 0 (0) | 0 (0) | 1 (3.7) | 0 (0) | 0 (0) | 0 (0) |
| Retrograde cricopharyngeal dysfunction | 1 (1.6) | 0 (0) | 0 (0) | 1 (3.7) | 1 (4.5) | 0 (0) | 0 (0) |
| Spasmodic dysphonia | 1 (1.6) | 0 (0) | 0 (0) | 1 (3.7) | 0 (0) | 0 (0) | 1 (7.1) |
| Laryngopharyngeal hypersensitivity syndrome | 2 (3.3) | 0 (0) | 1 (7.7) | 2 | 1 (4.5) | 1 (7.1) | 0 (0) |
| Vocal cord granuloma | 1 (1.6) | 1 (6.3) | 1 (7.7) | 1 (3.7) | 1 (4.5) | 1 (7.1) | 1 (7.1) |
| Asthmatic chronic cough | 1 (1.6) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| Vocal fold scars | 1 (1.6) | 0 (0) | 0 (0) | 1 (3.7) | 1 (4.5) | 0 (0) | 0 (0) |
| Total number of misdiagnoses | 60 | 16 | 13 | 27 | 22 | 14 | 14 |

Abbreviations: CRSNP, Chronic Rhinosinusitis without Nasal Polyps; re, regenerated.

ChatGPT-3.5/4 recommended adequate additional examinations in 10% to 29% of cases,[6,12] which reflects a lower capability of ChatGPT to provide additional examination recommendation rather than primary and differential diagnoses. This ChatGPT limitation may be related to its inability to select the most appropriate examinations as highlighted in 5 studies with a significantly higher number of recommended additional examinations per patient from ChatGPT versus otolaryngologists.[13-17] The treatment recommendations of ChatGPT were consistent with otolaryngologists in 16.7% to 60% of cases.[12,15,16] Additional examination and treatment recommendations were compared between ChatGPT-4 and Claude-3- and 3.5-Sonnet in 2 studies.[7,18] Schmidl et al showed that ChatGPT-4 and Claude-3-Sonnet reported comparable results for oncological treatment recommendations and explanations, whereas Claude-3-Sonnet demonstrated higher accuracy for diagnostic work-up than ChatGPT-4.[7] Interestingly, Claude-3-Sonnet more effectively selected appropriate additional examinations than ChatGPT-4.[7] In another study, ChatGPT-4o and Claude-3.5-Sonnet were challenged for primary and differential diagnoses of rare conditions in otolaryngology.[18] Consistent with our results, both Claude-3.5-Sonnet and ChatGPT-4o recommended a significantly higher number of additional examinations compared to practitioners. Concerning the accuracy of the primary diagnosis, Claude-3.5-Sonnet reported significantly higher rates of correct diagnosis compared to ChatGPT-4 (54.3% *versus* 45.7%) in both rare and common otolaryngological diseases,[18] which does not corroborate our observation highlighting similar performances.

The error analysis revealed that Claude-3.5-Sonnet (n = 14) and ChatGPT-4o (n = 16) made a significantly lower rate of clinical mistakes compared to Gemini-1.5-Pro (n = 27), which strengthens the superiority of both LLMs over Gemini-1.5-Pro as clinical adjunctive tools. Interestingly, while the errors of LLMs were corrected through human feedback, the re-generated inputs demonstrated that none of the LLMs modified their outputs. This suggests that single-user corrective feedback is not incorporated into subsequent outputs in public APIs. In the deep learning field, reinforcement learning from human feedback (RLHF) helps fine-tune responses by incorporating clinician inputs at scale, but such mechanisms are applied during model training, not at the level of individual interactions.[19] Despite its importance, RLHF has been poorly investigated in medical and surgical disciplines. O'Reilly et al evaluated an independent cardiological deep neural network for detecting electrocardiogram abnormalities and reported a significant improvement after human feedback.[20] To the best of our knowledge, whether RLHF mechanisms translate into short-term correctability of LLMs has never been investigated in otolaryngology-head and neck surgery.

The use of LLMs by patients cannot be limited to medical consultation information, and it can extend to pharmacological findings of prescribed treatments. In pharmacology, medication-related harm, also referred as adverse drug events includes preventable or non-preventable harms caused by interventions related to medication use.[21] Preventable medication error can occur at any step from the physician prescribing medications to the patient receiving the medication from the pharmacist. While it is common for patients to request additional information from practitioners and pharmacists about potential adverse events of drugs, they may also forget the posology (doses and duration), which may encourage some patients to use LLMs for obtaining this information. The results of the present study suggest moderate-to-high pharmacovigilance performances of ChatGPT-4o and Claude-3.5-Sonnet. In a recent systematic review of 30 medical studies, Ong et al reported that generative AI and LLMs demonstrated high performance for listing and detecting common adverse events, which corroborates our observation.[22] Although the current literature dedicated to the performance of LLMs in clinical pharmacology remains limited., Huang et al evaluated the performance of ChatGPT in clinical pharmacy practice, especially prescription review, patient medication education, and adverse drug reaction recognition.[23] While ChatGPT reported excellent drug counselling for doses and intake features, its performance for medication education and adverse event information was significantly lower than that of the pharmacists.[23]

The investigation of pharmacological accuracy and reinforcement learning properties of LLMs are the primary strengths of this study given the very limited literature about these 2 important aspects. While judge assessment can be subjective and influenced by clinical experience, adherence to guidelines for evaluating the accuracy of LLMs' outputs is an additional strength, supporting the almost perfect interrater reliability between both practitioners and pharmacists.

A limitation of our design is that we did not stratify model performance by information depth. Future studies including complementary examinations could help determine whether diagnostic accuracy varies with increasing levels of available clinical data.

We acknowledge that the number of recommended ancillary tests cannot be considered a validated quality metric, as the optimal diagnostic work-up is disease-specific and varies among clinicians. In this exploratory analysis, this metric was reported descriptively to illustrate the tendency of LLMs to over-recommend investigations compared to practitioners, rather than as a measure of diagnostic accuracy.

The small sample size and the lack of subspecialty analysis (laryngology, head and neck, rhinology, and otology) are the primary limitations of the present study. Our sample included a disproportionately high number of laryngology and head and neck cases and did not include a range of cases that might be more commonly encountered in various general otolaryngological practice settings. It limits the external validity of our findings. In addition, although the AIPI provides a valuable framework for evaluating LLM performance, this tool may not fully consider the nuances of AI's role in clinical decision-making, particularly in reflecting the complex dynamics of real-world clinical scenarios.

## Conclusion

Large language models showed variable accuracy and stability, with ChatGPT-4o and Claude-3.5-Sonnet outperforming Gemini-1.5-Pro across clinical and pharmacovigilance tasks. Although none of the models integrated single-user corrective feedback into subsequent outputs, this likely reflects the limits of public APIs rather than an absence of reinforcement learning.

### Acknowledgments

The authors have nothing to report.

### Author Contributions

**Filippo Bruno**, design, acquisition of data, data analysis and interpretation, 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, Bertrand Blankert**, data analysis and interpretation, 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, data analysis and 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.

### Disclosures

**Competing interests:** None.

**Funding source:** None.

### References

1. 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(2):603-608. doi:10.1002/ohn.816
2. 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(6):1519-1526. doi:10.1002/ohn.759
3. Swisher AR, Wu AW, Liu GC, Lee MK, Carle TR, Tang DM. Enhancing health literacy: evaluating the readability of patient handouts revised by ChatGPT's large language model. *Otolaryngol Head Neck Surg*. 2024;171(6):1751-1757. doi:10.1002/ohn.927
4. 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(1):e101102. doi:10.1136/bmjhci-2024-101102
5. Lechien JR, Saxena S, Vaira LA, Hans S, Maniaci A. Artificial intelligence-assisted diagnosis of an unusual cause of periodic epistaxis: a case report. *Ear Nose Throat J*. Published online June 3, 2025. doi:10.1177/01455613251335385
6. Lechien JR, Rameau A. Applications of ChatGPT in otolaryngology-head neck surgery: a state of the art review. *Otolaryngol Head Neck Surg*. 2024;171(3):667-677. doi:10.1002/ohn.807.7
7. 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 Otrhinolaryngol*. 2024;281(11):6099-6109. doi:10.1007/s00405-024-08828-1
8. Lorenzi A, Pugliese G, Maniaci A, et al. Reliability of large language models for advanced head and neck malignancies management: a comparison between ChatGPT 4 and Gemini Advanced. *Eur Arch Otrhinolaryngol*. 2024;281(9):5001-5006. doi:10.1007/s00405-024-08746-2
9. Dronkers EAC, Geneid A, Al Yaghchi C, et al. Evaluating the potential of AI Chatbots in treatment decision-making for acquired bilateral vocal fold paralysis in adults. *J Voice*. 2024;S0892-1997(24):00059-6. doi:10.1016/j.jvoice.2024.02.020
10. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med*. 2007;Oct 16 147(8):573-577. doi:10.7326/0003-4819-147-8-200710160-00010
11. 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 Otrhinolaryngol*. 2024;281(4):2063-2079. doi:10.1007/s00405-023-08219-y
12. Filali Ansary R, Lechien JR. Clinical decision support using large language models in otolaryngology: a systematic review. *Eur Arch Otrhinolaryngol*. 2025;282(8):4325-4334. doi:10.1007/s00405-025-09504-8
13. Maniaci A, Chiesa-Estomba CM, Lechien JR. ChatGPT-4 Consistency in Interpreting Laryngeal Clinical Images of Common Lesions and Disorders. *Otolaryngol Head Neck Surg*. 2024;171(4):1106-1113. doi:10.1002/ohn.897
14. Lechien JR, Chiesa-Estomba CM, Baudouin R, Hans S. Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. *Eur Arch Otrhinolaryngol*. 2024;281(4):2105-2114. doi:10.1007/s00405-023-08326-w
15. Lechien JR, Georgescu BM, Hans S, Chiesa-Estomba CM. ChatGPT performance in laryngology and head and neck surgery: a clinical case-series. *Eur Arch Otrhinolaryngol*. 2024;281(1):319-333. doi:10.1007/s00405-023-08282-5
16. Radulesco T, Saibene AM, Michel J, Vaira LA, Lechien JR. ChatGPT-4 performance in rhinology: a clinical case series. *Int*

*Forum Allergy Rhinol*. 2024;14(6):1123-1130. doi:10.1002/alr.23323

17. Maniaci A, Lazzeroni M, Cozzi A, et al. Can chatbots enhance the management of pediatric sialadenitis in clinical practice? *Eur Arch Otrhinolaryngol*. 2024;281(11):6133-6140. doi:10.1007/s00405-024-08798-4
18. Lechien JR, Maniaci A. Large Language Models as adjunctive tools for diagnosing rare diseases in otolaryngology: a controlled study. Oral presentation at: The Annual Meeting of the American Academy of Otolaryngology Head Neck Surgery; October 12, 2025; Indianapolis, IN, USA.
19. Berry P, Dhanakshirur RR, Khanna S. Utilizing large language models for gastroenterology research: a conceptual framework. *Therap Adv Gastroenterol*. 2025;18:17562848251328577. doi:10.1177/17562848251328577
20. O'Reilly C, Oruganti SDR, Tilwani D, Bradshaw J. Model-driven analysis of ECG using reinforcement learning. *Bioengineering*. 2023;10(6):696. doi:10.3390/bioengineering10060696
21. Bates DW. Incidence of adverse drug events and potential adverse drug events: implications for prevention. *JAMA*. 1995;274(1):29-34.
22. Ong JCL, Chen MH, Ng N, et al. A scoping review on generative AI and large language models in mitigating medication-related harm. *npj Digital Medicine*. 2025;8(1):182. doi:10.1038/s41746-025-01565-7
23. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: a comparative study of ChatGPT and clinical pharmacists. *Br J Clin Pharmacol*. 2024;90(1):232-238. doi:10.1111/bcp.15896