Laryngoscope
Investigative Otolaryngology

ORIGINAL RESEARCH OPEN ACCESS

Evaluation of Large Language Models' Concordance With Guidelines on Olfaction

Ariana L. Shaari[1] | Anthony M. Saad[1] | Disha Patil[1] | Jennifer Yanik[1] | Wayne D. Hsueh[1] | Jean Anderson Eloy[1,2,3,4,5] | Andrey Filimonov[1]

[1]Department of Otolaryngology – Head and Neck Surgery, Rutgers New Jersey Medical School, Newark, New Jersey, USA | [2]Center for Skull Base and Pituitary Surgery, Neurological Institute of New Jersey, Rutgers New Jersey Medical School, Newark, New Jersey, USA | [3]Department of Neurological Surgery, Rutgers New Jersey Medical School, Newark, New Jersey, USA | [4]Department of Ophthalmology and Visual Science, Rutgers New Jersey Medical School, Newark, New Jersey, USA | [5]Department of Otolaryngology and Facial Plastic Surgery, Saint Barnabas Medical Center - RWJBarnabas Health, Livingston, New Jersey, USA

Correspondence: Ariana L. Shaari (arianashaari@gmail.com)

Received: 10 December 2024 | **Revised:** 2 March 2025 | **Accepted:** 7 March 2025

Funding: The authors received no specific funding for this work.

Keywords: artificial intelligence | large language models | olfaction | rhinology | surgical education

ABSTRACT

Objective: To assess the concordance of artificial intelligence (AI)-generated information with the 2022 International Consensus Statement on Allergy and Rhinology: Olfaction (ICAR-O).

Methods: Forty-two guidelines were extracted from the ICAR-O. Each guideline was converted into a question, which was presented to ChatGPT version 4.o and Google Gemini. Concordance was deemed an agreement between the AI response and the clinical recommendation. Credibility was granted if the AI platform provided a credible resource. Accuracy was graded on a Likert scale (0: entirely inaccurate information, 1: mix of accurate and inaccurate information, 2: entirely accurate information). Statistical analysis was performed.

Results: A total of 84 responses were generated. The mean accuracy of the ChatGPT and Gemini responses was 1.85 and 1.48 out of 2, respectively, indicating that the responses contained a mix of accurate and inaccurate information. ChatGPT responses were significantly more accurate than Gemini responses ($p = 0.001$). Of the ChatGPT responses, 78.57% ($N = 33$) were concordant with the ICAR-O guidelines and 100% ($N = 42$) cited a credible resource. Of the Gemini responses, 66.67% ($N = 28$) were concordant and 97.62% ($N = 41$) cited a credible resource. There were no significant differences in concordance ($p = 0.22$) or credibility ($p = 0.31$) between the AI platforms.

Conclusion: ChatGPT provided more accurate information than Gemini on olfaction. However, overall, both platforms did not consistently align with clinical guidelines. AI platforms require further evaluation before clinical implementation or use as educational adjuncts.

Level of Evidence: N/A.

1 | Introduction

Advances in artificial intelligence (AI) have ushered in pivotal changes in healthcare. Large language models (LLMs), such as ChatGPT version 4.o (OpenAI, San Francisco, CA, USA) and Google Gemini (formerly known as Google Bard) (Google, Mountain View, CA, USA) are powerful AI models that can generate images, interpret convoluted questions,

and interact with users through conversation [1–3]. The development of AI can be categorized into two main phases. AI 1.0 includes symbolic AI, in which human knowledge is manually encoded into rules and logic, and probabilistic models that use statistical reasoning to handle uncertainty [4]. These systems, while predictable, were limited in flexibility and required expertise to create [4]. AI 2.0 began with deep learning, which utilizes neural networks to learn from large datasets [5]. This shift brought breakthroughs in areas such as computer vision, natural language processing, and healthcare, enabling AI systems to perform specialized tasks such as classification and prediction with high accuracy [5]. ChatGPT version 3.0 was released in November 2022, reaching over 100 million users in 2 months [6]. Similarly, Google released Gemini in March 2023 and had over 300 million users by May 2024 [7]. Within otolaryngology, interest in LLMs has been steadily growing. Studies have evaluated the ability of LLMs to create educational materials for conditions such as benign paroxysmal positional vertigo, serve as a research aid, generate information on procedures such as rhinoplasty, and interpret videolaryngostroboscopy images of laryngological conditions [8–12].

Despite these advancements, LLMs' knowledge of the diagnosis and management of olfactory disorders/dysfunction (OD) remains relatively unexplored. OD are an evolving subject within otolaryngology. They often have a profoundly negative impact on patients' quality of life and can be associated with psychological issues such as anxiety and depression [13, 14]. New developments in treating OD have emerged, especially in the wake of the COVID-19 pandemic [15]. Previous literature suggests that the most recent version of ChatGPT, ChatGPT version 4.o, can appropriately answer questions about COVID-19-induced olfactory dysfunction [16]. Studies have demonstrated that other versions of ChatGPT, such as ChatGPT-4.0, are capable of citing and utilizing accurate resources in head and neck surgery; in contrast, the older version, ChatGPT-3.5, often invented false references [17]. This continuous growth in the reliability of AI platforms in handling and dispensing medical information beckons to be studied further and frequently, as it is not stagnant.

Recent studies have evaluated the quality of ChatGPT-generated information against clinical guidelines within otolaryngology, such as the International Consensus Statement on Allergy and Rhinology: Rhinosinusitis [18]. However, there is a gap in the literature comparing AI-generated information to other rhinologic guidelines. To consolidate the developing literature on OD, an international expert panel published the International Consensus Statement on Allergy and Rhinology: Olfaction (ICAR-O) in 2022 [18]. Given that the management and outcomes of OD are a frequently evolving area of study and the significant impact of OD on patients' quality of life, it is imperative to ensure that LLM-generated information aligns with the most relevant clinical guidelines on OD. No studies to date have specifically evaluated the concordance of LLMs with guidelines on olfaction. Therefore, this study aimed to fill that gap and evaluate the accuracy, credibility, and concordance of responses by two major LLMs to questions generated from ICAR-O 2022.

2 | Materials and Methods

The ICAR-O guidelines were accessed. Forty-two guidelines and their designated grade of evidence were extracted based on the recommendations in ICAR-O. According to the ICAR-O guidelines, an aggregate grade of evidence of A corresponds to well-designed randomized control trials (RCT), B to RCTs with minor limitations, C to observational studies, and D to expert opinions or case reports [18]. Two reviewers independently converted each guideline into a question. A final question was agreed upon through discussion. Each question was additionally prompted with "Cite 1 resource you used to answer this question" to evaluate the quality of resources that the LLM provides. Each question was independently inputted into ChatGPT version 4.o and Google Gemini on October 6th, 2024 (Newark, New Jersey). ChatGPT 4.o and Google Gemini were selected as the two LLMs to evaluate due to their popularity and ease of access. Responses were pasted verbatim from the AI platform into a shared document.

Concordance was defined as an agreement between the LLM response and the clinical guideline recommendation (1: Concordant, 0: Discordant). A resource was judged to be credible if it was sponsored by a federal agency, medical school, hospital, large professional or nonprofit organization, or written by a medical or scientific professional [19, 20]. The resource provided by the LLM was graded for credibility (1: Credible Resource, 0: Not-Credible Resource). Two reviewers independently reviewed each response for concordance and credibility; all responses were reviewed for accuracy by a board-certified rhinologist. Accuracy was graded on a 3-point Likert scale (0: entirely inaccurate information, 1: mix of accurate and inaccurate information, 2: entirely accurate information) [21, 22].

Statistical analysis for the categorical data of concordance and credibility was conducted using a z-score test for population proportions to determine if there was a significant difference in the proportion of concordant and credible responses between ChatGPT and Gemini. For accuracy, the difference in mean accuracy between ChatGPT and Gemini responses was compared using a two-sample t-test. All analysis was conducted at the 95% confidence level, with p-values less than 0.05 considered statistically significant.

3 | Results

Overall, a total of 84 responses were generated by ChatGPT and Gemini and assessed. Samples of the ICAR-O guideline, question, and LLM responses are displayed in Table 1.

Figure 1 displays the number of concordant responses by ChatGPT and Gemini. Of the 42 ChatGPT responses, 78.57% ($N = 33$) were concordant with the ICAR-O guidelines. Of the 42 Gemini responses, 66.67% ($N = 28$) were concordant. There was no significant difference found between the two LLM platforms regarding concordance ($p = 0.22$).

Figure 2 displays the number of credible resources provided by ChatGPT and Gemini. Of the 42 ChatGPT responses, 100%

TABLE 1 | Sample guideline, question, and artificial intelligence responses.

Policy	Question	ChatGPT Response	Google Gemini Response
Intervention:Olfactory Training should be considered in patients with Post-traumatic olfactory disorder/dysfunction.	Is olfactory training indicated in patients with post-traumatic Olfactory disorder/dysfunction? Cite 1 resource you used to answer this question.	Yes, **olfactory training** is indicated for patients with post-traumatic olfactory disorder/dysfunction. Olfactory training involves repeated exposure to specific odors over an extended period to stimulate and potentially restore olfactory function. Studies have shown that patients with post-traumatic olfactory loss can benefit from olfactory training, as it can improve odor detection and identification abilities over time. This non-invasive approach is particularly helpful in post-traumatic cases, where the olfactory pathways may have been disrupted but not entirely destroyed. It works by promoting neural plasticity and regeneration in the olfactory system, which can enhance recovery in some patients. Though the degree of improvement may vary, olfactory training is widely considered a safe and effective therapeutic option for post-traumatic olfactory dysfunction. *Source:* Damm, M., Pikart, L. K., Reimann, H., Burkert, S., Göktas, Ö., Haxel, B., Frey, S., & Hummel, T. (2014). **Olfactory training is helpful in postinfectious but not in posttraumatic olfactory loss.** *Laryngoscope, 124*(4), 826–831.	**Yes, olfactory training is indicated in patients with post-traumatic olfactory disorder/dysfunction.** This is because olfactory training can help stimulate and retrain the olfactory system, potentially improving olfactory function. **Resource:** * **Frontiers in Neurology:** https://pubmed.ncbi.nlm.nih.gov/29790420/ Olfactory training typically involves repeatedly sniffing specific odors over a period of time. While the effectiveness of olfactory training for post-traumatic olfactory dysfunction may vary, it is often considered a safe and relatively low-risk intervention that may improve olfactory function in some patients.
Intervention: Dupilumab may be recommended for patients with OD related to severe CRSwNP who have not improved despite other medical and surgical treatment options as part of a shared decision-making process.	Is dupilumab indicated in patients with Olfactory disorder/dysfunction related to chronic rhinosinusitis with nasal polyps? Cite 1 resource you used to answer this question.	Yes, **dupilumab** is indicated in patients with olfactory dysfunction related to **chronic rhinosinusitis with nasal polyps (CRSwNP)**. Dupilumab is a monoclonal antibody that targets the **IL-4 and IL-13** pathways involved in type 2 inflammation, which plays a significant role in CRSwNP. Clinical studies have demonstrated that dupilumab not only helps reduce nasal polyp size and improve nasal obstruction but also significantly improves olfactory function in patients with CRSwNP. Dupilumab is particularly useful for patients who do not respond adequately to intranasal corticosteroids or those who have recurrent polyps following surgery. Its use has been associated with improvements in quality of life, including better sense of smell, and it provides a systemic anti-inflammatory effect without the side effects associated with long-term oral corticosteroid use. *Source:* Bachert, C., Han, J. K., Desrosiers, M., Hellings, P. W., Amin, N., Lee, S. E., Smith, S. G., & Khan, A. H. (2019). **Dupilumab in patients with severe chronic rhinosinusitis with nasal polyps (LIBERTY NP SINUS-24 and LIBERTY NP SINUS-52): two multicentre, randomized, double-blind, placebo-controlled, phase 3 trials.** *Lancet, 394*(10,209), 1638–1650.	**Yes, dupilumab is indicated in patients with olfactory disorder/dysfunction related to chronic rhinosinusitis with nasal polyps.** This is based on recent clinical studies demonstrating its effectiveness in improving olfactory function in these patients. **Resource:** * **National Institutes of Health (NIH):** https://pubmed.ncbi.nlm.nih.gov/34628065/ Dupilumab is a monoclonal antibody that targets interleukin-4 and interleukin-13, which are involved in allergic inflammation. By blocking these cytokines, dupilumab can reduce inflammation in the nasal passages and sinuses, which can improve nasal breathing and potentially enhance olfactory function.

(Continues)

TABLE 1 | (Continued)

Policy	Question	ChatGPT Response	Google Gemini Response
Intervention: Omalizumab may be recommended for OD related to patients with severe CRSwNP who have not improved despite other medical and surgical treatment options as part of a shared decision-making process.	Is omalizumab indicated in patients with Olfactory disorder/dysfunction related to chronic rhinosinusitis with nasal polyps? Cite 1 resource you used to answer this question.	Yes, **omalizumab** is indicated in patients with olfactory dysfunction related to **chronic rhinosinusitis with nasal polyps (CRSwNP)**, particularly in those with **coexisting allergic rhinitis or elevated IgE levels**. Omalizumab is a monoclonal antibody that targets IgE, reducing allergic inflammation, which plays a role in the development and persistence of nasal polyps in some patients with CRSwNP. Clinical trials have shown that omalizumab can reduce the size of nasal polyps, improve nasal obstruction, and enhance olfactory function in patients with CRSwNP, especially in those with allergic comorbidities. It has been approved as an add-on therapy for patients with severe CRSwNP who do not respond adequately to intranasal corticosteroids or who have recurrent polyps despite surgery. Omalizumab is part of a growing number of biologic therapies offering targeted treatment to improve symptoms, including the sense of smell, by reducing underlying inflammation in CRSwNP. *Source:* Gevaert, P., Calus, L., Van Zele, T., et al. (2013). **Omalizumab is effective in allergic and nonallergic patients with nasal polyps and asthma.** *Journal of Allergy and Clinical Immunology, 131*(1), 110-116.e1.	**Yes, omalizumab is indicated in patients with olfactory disorder/dysfunction related to chronic rhinosinusitis with nasal polyps.** This is based on clinical studies that have demonstrated its effectiveness in improving olfactory function in these patients. **Resource:** * **National Institutes of Health (NIH):** https://pubmed.ncbi.nlm.nih.gov/34628065/ Omalizumab is a monoclonal antibody that targets immunoglobulin E (IgE), a protein involved in allergic reactions. By reducing IgE levels, omalizumab can help reduce inflammation in the nasal passages and sinuses, which can improve nasal breathing and potentially enhance olfactory function.
Intervention: More data are needed before clinicians can present this as a beneficial treatment option for their patients	Is α-lipoic indicated in the treatment of Olfactory disorder/dysfunction? Cite 1 resource you used to answer this question.	Yes, **alpha-lipoic acid (ALA)** is indicated as a potential treatment for certain types of **olfactory dysfunction**, particularly **post-viral olfactory loss**. Alpha-lipoic acid is an antioxidant that has been studied for its neuroprotective properties, and some research suggests it may help improve olfactory function by reducing oxidative stress and promoting nerve regeneration. A study by Hummel et al. (2002) found that alpha-lipoic acid had beneficial effects on patients with olfactory dysfunction following upper respiratory infections. The study suggested that ALA could help restore olfactory function by mitigating the damage to olfactory neurons caused by oxidative stress. Although ALA is not a universally established treatment, it has shown promise as an adjunct therapy, particularly for patients with post-infectious olfactory dysfunction. *Source:* Hummel, T., Heilmann, S., & Hüttenbrink, K. B. (2002). **Lipoic acid in the treatment of smell dysfunction following viral infection of the upper respiratory tract.** *Laryngoscope, 112*(11), 2076–2080.	**Yes, α-lipoic acid (ALA) may be indicated in the treatment of olfactory disorder/dysfunction, especially in cases related to post-infectious anosmia.** Studies have shown that ALA can improve olfactory function in patients with this condition. **Resource:** * **Lipoic acid in the treatment of smell dysfunction following viral infection of the upper respiratory tract:** https://pubmed.ncbi.nlm.nih.gov/12439184/ ALA is believed to work by reducing inflammation, promoting nerve growth, and acting as an antioxidant. However, more research is needed to fully understand its effectiveness and the optimal dosage for treating olfactory disorder/dysfunction. It is recommended to consult with a healthcare professional before starting ALA supplementation.

(Continues)

TABLE 1 | (Continued)

Policy	Question	ChatGPT Response	Google Gemini Response
Intervention: Surgical intervention for phantosmia is not recommended at this time, except in extremely rare cases. Referral to an expert in this field can be considered in cases that do not resolve with time, have failed multiple trials of medical therapy, and are life-threatening	Is surgical internvetion indicated for phantosmia? If so, in what cases? Cite 1 resource you used to answer this question.	**Surgical intervention** may be indicated for **phantosmia** (the perception of phantom smells) in certain cases, particularly when the condition is refractory to medical management or when there is an identifiable structural cause. Phantosmia is often caused by damage or irritation to the olfactory system, and surgical intervention is considered in the following scenarios: 1. **Structural abnormalities**: If imaging studies reveal an identifiable structural issue, such as a **tumor**, **nasal polyps**, or **olfactory groove meningioma**, that could be causing irritation or compression of the olfactory nerves, surgery may be necessary to correct the underlying issue and alleviate the phantosmia. 2. **Chronic rhinosinusitis or nasal polyps**: In patients with **chronic rhinosinusitis** or **nasal polyps**, endoscopic sinus surgery may be indicated to clear the obstruction and reduce inflammation, which could resolve the olfactory distortion causing phantosmia. 3. **Refractory cases**: In rare, refractory cases where no clear structural cause is found, but the phantosmia is severely affecting quality of life and not responding to conservative treatments (e.g., nasal corticosteroids or olfactory training), surgical options such as **olfactory bulb ablation** or **sectioning of the olfactory nerve** may be considered. However, these more extreme measures carry the risk of complete anosmia and are typically reserved for severe cases where other options have failed. Surgical intervention is generally considered only after thorough evaluation and when conservative treatments have been unsuccessful. *Source:* Leopold, D. A. (2002). **Olfactory distortion: Diagnosis and treatment.** *Chemical Senses, 27*(7), 611–615.	**Surgical intervention for phantosmia is generally not recommended.** While there have been rare cases where surgery has been used to treat phantosmia, it is often considered a last resort and is not always effective. **Resource:** * **National Institutes of Health (NIH):** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7689254/ In most cases, phantosmia is treated with medical interventions, such as medications to address underlying conditions or manage symptoms. If you are experiencing phantosmia, it is important to consult with a healthcare professional to determine the best course of treatment for your specific situation.

(Continues)

TABLE 1 | (Continued)

Policy	Question	ChatGPT Response	Google Gemini Response
Intervention: A validated olfactory QOL survey should be considered in individuals with CRS and in those who may have other diseases that impact olfaction.	Are validated olfactory quality of life surveys indicated in patients with chronic rhinosinusitis and in those who may have other diseases that impact olfaction? Cite 1 resource you used to answer this question.	Yes, **validated olfactory quality of life surveys** are indicated for patients with chronic rhinosinusitis (CRS) and other diseases that impact olfaction. These surveys, such as the **Questionnaire of Olfactory Disorders (QOD)** and the **Sino-Nasal Outcome Test (SNOT-22)**, are designed to assess the impact of olfactory dysfunction on patients' daily lives and overall quality of life. They are particularly useful in chronic conditions like CRS, where olfactory impairment is common, and in conditions like neurodegenerative diseases or post-traumatic anosmia. These surveys provide insights into how olfactory dysfunction affects aspects such as mood, social interactions, and food enjoyment, offering a comprehensive view that complements objective olfactory testing. They are valuable for guiding treatment decisions and tracking patient outcomes over time. *Source:* Litvack, J. R., Mace, J. C., & Smith, T. L. (2009). **Olfactory function and disease severity in chronic rhinosinusitis.** *American Journal of Rhinology & Allergy, 23*(2), 139–144.	**Yes, validated olfactory quality of life surveys are indicated in patients with chronic rhinosinusitis and those who may have other diseases that impact olfaction.** These surveys can provide valuable insights into the impact of olfactory dysfunction on a patient's daily life and overall well-being. **Resource:** * **National Institutes of Health (NIH):** https://pubmed.ncbi.nlm.nih.gov/33078172/ By assessing a patient's subjective experience of olfactory impairment, these surveys can help healthcare providers better understand the extent of the patient's disability and tailor treatment accordingly. Additionally, they can be used to track changes in olfactory function over time and evaluate the effectiveness of interventions. Some commonly used olfactory quality of life surveys include: * **Olfactory Disorders Questionnaire (QOD):** This survey assesses the impact of olfactory dysfunction on various aspects of daily life, such as social interactions, food enjoyment, and emotional well-being. * **Sensation-Specific Olfactory Dysfunction Questionnaire (SSODQ):** This survey focuses on the specific sensory experiences associated with olfactory dysfunction, such as changes in odor perception or the development of parosmia. * **Olfactory Function Index (OFI):** This survey assesses the impact of olfactory dysfunction on a patient's overall quality of life, including physical, social, and emotional domains.

(Continues)

TABLE 1 | (Continued)

Policy	Question	ChatGPT Response	Google Gemini Response
Intervention: Topical nasal corticosteroids are an option for OD in patients with CRSsNP before or after sinus surgery	Are topical nasal corticosteroids an option for olfactory dysfuncton in patients with chronic rhinosinusitis with nasal polyps, before or after sinus surgery? Cite 1 resource you used to answer this question.	Yes, **topical nasal corticosteroids** are an effective option for olfactory dysfunction in patients with **chronic rhinosinusitis with nasal polyps (CRSwNP)** both **before and after sinus surgery**. 1. **Before surgery**: Intranasal corticosteroids are used to reduce inflammation, shrink nasal polyps, and improve nasal airflow, which can help alleviate olfactory dysfunction in patients with CRSwNP. They are a first-line treatment for managing symptoms and preventing the need for surgery. 2. **After surgery**: Following sinus surgery, topical nasal corticosteroids are crucial for **preventing the recurrence** of nasal polyps and maintaining improvements in olfaction and nasal airflow. The surgery helps open the sinonasal passages, allowing better access of the medication to the target tissues, thereby enhancing its effectiveness in managing olfactory dysfunction. Intranasal corticosteroids are well-tolerated for long-term use, making them a key part of both pre- and post-surgical management strategies for CRSwNP. *Source:* Fokkens, W. J., Lund, V. J., Hopkins, C., et al. (2020). **European position paper on rhinosinusitis and nasal polyps 2020.** *Rhinology, 58*(S29), 1–464.	**Yes, topical nasal corticosteroids are an option for olfactory dysfunction in patients with chronic rhinosinusitis with nasal polyps, both before and after sinus surgery.** These medications can help reduce inflammation in the nasal passages and sinuses, which can improve nasal breathing and potentially enhance olfactory function. **Resource:** * **National Institutes of Health (NIH):** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7689254/ Intranasal corticosteroids can be used as a long-term maintenance therapy to prevent the recurrence of nasal polyps and improve olfactory function. They can also be used in combination with other treatments, such as sinus surgery, to optimize outcomes. It is important to consult with a healthcare provider to determine the most appropriate dosage and duration of treatment.

(Continues)

TABLE 1 | (Continued)

Policy	Question	ChatGPT Response	Google Gemini Response
Intervention: Laboratory tests, including serum autoimmune markers should be considered in individuals with OD and suspected underlying autoimmune disease	What laboratory tests are ordered with patients with Olfactory disorder/dysfunction and suspected underlying autoimmune disease? Cite 1 resource you used to answer this question.	For patients with olfactory dysfunction and suspected underlying **autoimmune disease**, the following laboratory tests are commonly ordered to evaluate for autoimmune conditions that may contribute to the olfactory disorder: 1. **Antinuclear Antibodies (ANA)**: This is a screening test for systemic autoimmune diseases, such as **systemic lupus erythematosus (SLE)**, **Sjogren's syndrome**, and **rheumatoid arthritis**. A positive ANA test may indicate the presence of an autoimmune disorder, though further specific testing is required for a definitive diagnosis. 2. **Rheumatoid Factor (RF)**: This test is often ordered if **rheumatoid arthritis** is suspected. High levels of RF can indicate autoimmune involvement in joints and potentially other systems. 3. **Anti-Double-Stranded DNA (anti-dsDNA) Antibodies**: This test is more specific for **systemic lupus erythematosus (SLE)** and may be ordered if lupus is suspected as the underlying cause of olfactory dysfunction. 4. **Anti-Ro/SSA and Anti-La/SSB Antibodies**: These are specific tests for **Sjogren's syndrome**, which can affect the mucous membranes and glands, leading to dryness and potential olfactory dysfunction. 5. **Thyroid Peroxidase Antibodies (TPO-Ab)**: This test is used to evaluate for **autoimmune thyroid disease**, such as **Hashimoto's thyroiditis**, which can also lead to olfactory dysfunction through its effects on thyroid function. 6. **Erythrocyte Sedimentation Rate (ESR) and C-Reactive Protein (CRP)**: These are non-specific markers of inflammation that may be elevated in autoimmune diseases. While not diagnostic, they are useful for assessing the degree of systemic inflammation. 7. **Anti-Neutrophil Cytoplasmic Antibodies (ANCA)**: This test is used to diagnose **granulomatosis with polyangiitis (GPA)**, an autoimmune disease that affects small blood vessels and can lead to sinusitis and olfactory dysfunction. These tests help identify autoimmune diseases that might affect the olfactory system by causing systemic inflammation, mucosal involvement, or direct neural damage. *Source:* Doty, R. L. (2012). **Olfactory dysfunction in systemic autoimmune disease.** *Handbook of Clinical Neurology, 117*, 349–362.	When olfactory disorder/dysfunction is suspected to be related to an underlying autoimmune disease, several laboratory tests may be ordered to help confirm the diagnosis: * **Antinuclear antibodies (ANA):** These antibodies are often found in autoimmune diseases, such as systemic lupus erythematosus (SLE). * **Rheumatoid factor (RF):** This antibody is associated with rheumatoid arthritis. * **Anti-neutrophil cytoplasmic antibodies (ANCA):** These antibodies are associated with vasculitis, a type of autoimmune disease that can affect the sinuses and nasal passages. * **C-reactive protein (CRP):** This is a marker of inflammation that can be elevated in autoimmune diseases. * **Erythrocyte sedimentation rate (ESR):** This is another marker of inflammation that can be elevated in autoimmune diseases. * **Specific antibodies for other autoimmune diseases:** Depending on the suspected diagnosis, tests for specific antibodies may be ordered, such as anti-dsDNA antibodies for SLE or anti-thyroid peroxidase antibodies for autoimmune thyroid disease. **Resource:** * **National Institutes of Health (NIH):** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7689254/ It is important to note that these are just a few examples of the laboratory tests that may be ordered in patients with olfactory disorder/dysfunction and suspected autoimmune disease. The specific tests ordered will depend on the patient's individual symptoms and medical history.

(Continues)

TABLE 1 | (Continued)

Policy	Question	ChatGPT Response	Google Gemini Response
Intervention: The potential benefit of topical vitaminA and the potential for local irritation can be discussed with the patient and if the shared decision-making process leads to choosing this option for treatment, it can be administered intranasally with the patient in the Kaiteki position at a dose of 10,000IU once daily for 8weeks.	Is vitamin A indicated in the treatment of Olfactory disorder/dysfunction? If so, what should the route of adminstration be, and at what dose for what duration? Cite 1 resource you used to answer this question.	Yes, **vitamin A** has been investigated as a potential treatment for certain types of **olfactory dysfunction**, particularly **post-viral olfactory loss**. Vitamin A is essential for epithelial cell health and regeneration, and it is thought to support the repair of the olfactory epithelium, where the sensory neurons responsible for smell are located. ### Route of Administration: Vitamin A is typically administered **intranasally** for olfactory dysfunction. This localized delivery allows the vitamin to directly target the olfactory mucosa, potentially enhancing the regeneration of olfactory receptor neurons. ### Dose and Duration: In clinical studies, **intranasal vitamin A** is usually administered at a dose of **10,000 IU once daily** for a period of **12weeks**. This dosing has shown potential benefits in improving olfactory function in patients with post-viral olfactory loss. However, it's important to note that the evidence for vitamin A in olfactory dysfunction is still evolving, and it may not work for all patients. Vitamin A therapy should be considered under medical supervision, especially given the potential for toxicity with excessive doses of vitamin A. *Source:* Hummel, T., Rissom, K., Reden, J., Hahner, A., Weidenbecher, M., & Hüttenbrink, K. B. (2009). **Effects of olfactory training in patients with olfactory loss.** *Laryngoscope, 119*(3), 496–499.	**Yes, vitamin A has been shown to be effective in the treatment of olfactory disorder/dysfunction, particularly in cases of post-infectious anosmia.** It is believed to work by promoting the regeneration of olfactory receptor neurons. **Route of administration:** Vitamin A is typically administered intranasally, as a nasal spray or drops. **Dosage:** The recommended dosage is 10,000IU per day for 2months. **Resource:** * **National Institutes of Health (NIH):** https://pubmed.ncbi.nlm.nih.gov/28434127/ It is important to consult with a healthcare professional before starting vitamin A supplementation, as it may not be appropriate for everyone.
Intervention: Oral PDEIs should not be recommended in patients with OD as the potential for benefit is inconclusive and there exists potential for harm. Providers should inform their patients that the evidence for intranasal theophylline is preliminary and inconclusive before considering its use.	Is theophylline or other phosphodiesterase Inhibitors indicated in the treatment of Olfactory disorder/dysfunction? Cite 1 resource you used to answer this question.	Yes, **theophylline**, a phosphodiesterase inhibitor, has been investigated and shown some potential benefit in the treatment of **olfactory dysfunction**, particularly in cases of **post-viral olfactory loss** or **post-traumatic anosmia**. Theophylline is thought to work by increasing levels of **cyclic AMP (cAMP)**, which can promote cellular signaling pathways involved in the regeneration and function of olfactory receptor neurons. Some studies suggest that both **oral** and **intranasal** formulations of theophylline may help improve olfactory function in patients with olfactory dysfunction, likely by enhancing the regeneration of damaged olfactory cells and reducing inflammation. Oral doses in studies have typically ranged from **200 mg to 800 mg daily**. However, theophylline is not widely used as a standard treatment for olfactory dysfunction, and its use is limited due to potential side effects such as nausea, tremors, and heart palpitations. Despite these concerns, it has shown promise as an adjunctive therapy for olfactory dysfunction, especially in cases where more conventional treatments have failed. *Source:* Henkin, R. I., & Schultz, M. (2009). **Theophylline improves smell and taste dysfunction in patients with olfactory and gustatory dysfunctions.** *Journal of Allergy and Clinical Immunology, 123*(2), 409–410.	**No, theophylline or other phosphodiesterase inhibitors are not indicated in the treatment of olfactory disorder/dysfunction.** These medications are primarily used for conditions like asthma and chronic obstructive pulmonary disease (COPD), and there is no evidence to support their use for olfactory dysfunction. **Resource:** * **National Institutes of Health (NIH):** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7689254/ If you have olfactory disorder/dysfunction, it is important to consult with a healthcare professional to determine the underlying cause and receive appropriate treatment.

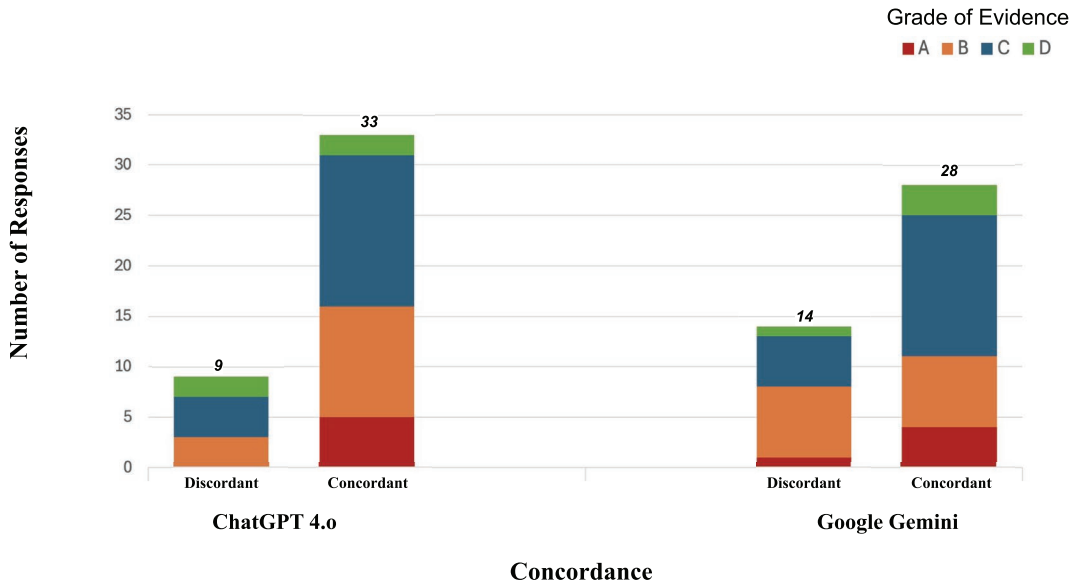

FIGURE 1 | Bar chart displaying concordance of responses provided by ChatGPT 4.o and Google Gemini.

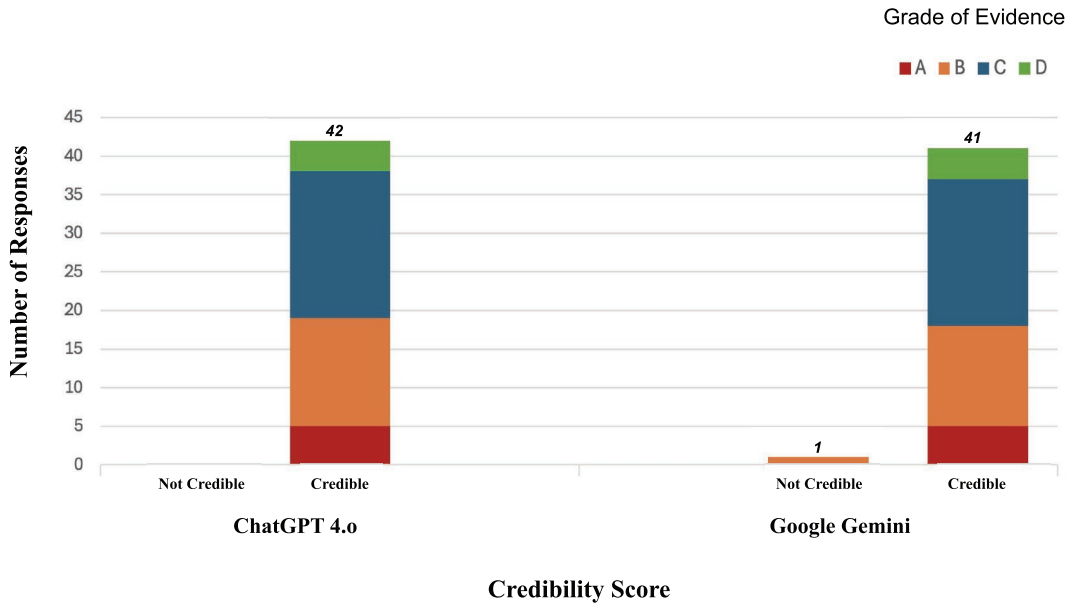

FIGURE 2 | Bar chart displaying the credibility of resources provided by ChatGPT 4.o and Google Gemini.

were credible. Of the 42 Gemini responses, 97.62% ($N = 41$) were credible. No significant differences in credibility were found ($p = 0.31$).

Figure 3 displays the accuracy scores for ChatGPT and Gemini. The mean accuracy score of the 42 ChatGPT responses was 1.85 (standard deviation: 0.36), correlating to generally a mix of accurate and inaccurate information. ChatGPT provided no completely inaccurate responses. The mean accuracy score of the 42 Gemini responses was 1.48 (standard deviation: 0.60). Gemini provided 2 completely inaccurate responses. ChatGPT responses were found to be significantly more accurate than Google responses ($p = 0.001$).

Six questions received discordant scores between ChatGPT and Google Gemini's responses. Of these, three pertained to recommendations with a C-level aggregate grade of evidence, two to B-level, and one to D-level evidence. Notably, all questions related to policies with A-level evidence elicited concordant responses from both models. Both ChatGPT and Google Gemini demonstrated difficulty in providing accurate responses to policies concerning alternative treatments and supplements for OD management. Examples of such treatments include omega-3, α-lipoic acid, Toki-shakuyaku-san, and minocycline. Both models struggled to deliver accurate information on these treatments.

4 | Discussion

In this study, we examined the ability of ChatGPT and Google Gemini to generate responses to clinical questions based on the ICAR guidelines for olfactory dysfunction/disorders. Studies

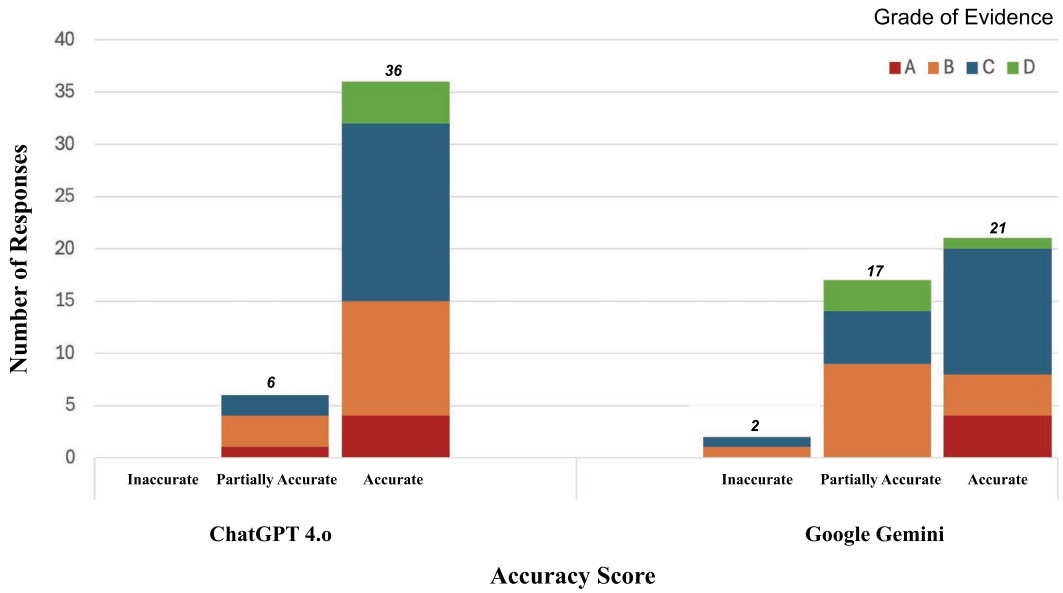

FIGURE 3 | Bar chart displaying accuracy of responses provided by ChatGPT 4.o and Google Gemini.

have investigated the ability of ChatGPT to adhere to guidelines on Rhinosinusitis, thyroid carcinoma, and various clinical questions from the American Academy of Otolaryngology—Head and Neck Surgery (AAO—HNS), yet this represents the first comparative analysis evaluating multiple LLMs, assessing resource credibility, accuracy, and concordance with established guidelines [23–25].

LLMs, particularly those in the "third epoch" of AI as characterized by Howel et al., can evaluate complex questions, generate text and images, and interact with users in a human-like manner [26–28]. Unlike older generations of AI, which must be re-trained on a new dataset for each new task and have a limited ability to generate new sentences or visual content, platforms in the third epoch require no retraining to perform new tasks; they can also condense information and interact with a user without being specifically trained to do so [6]. They are trained on significant amounts of data and evaluate the frequency of the sequence of words from a larger body of text [6, 29].

Our results revealed an overall mean concordance with the ICAR-O guidelines of 72.62%, with some instances of complete contradiction with the guidelines, echoing Tessler et al.'s findings of a 2.8% contradiction rate of ChatGPT to AAO-HNS guidelines. ChatGPT matched AAO-HNS guidelines on rhinoplasty in 60% of cases [30]. These findings have been corroborated in other areas of surgery, with studies finding ChatGPT 4.0 to be 67.9% concordant with spondylolisthesis guidelines [31]. We also found that both platforms displayed high rates of credible resources (97.62%–100%), aligning with studies on resource credibility of ChatGPT 4.0 [19, 30]. While these results are commendable, having low concordance with guidelines that the AI platforms should be able to access on the internet further contributes to concern over their use as clinical adjuncts.

We found significant differences in accuracy between ChatGPT and Gemini. ChatGPT yielded completely accurate information in 86% of responses, while Gemini generated completely accurate information in only 50% of responses, highlighting significant variability in the reliability of these platforms. Yoshiyasu et al. evaluated the accuracy of ChatGPT-4 responses against the ICAR Rhinosinusitis statement and revealed similar limitations, finding a 54% accuracy rate of all ChatGPT-responses [23]. Tessler et al. further described the limitations of ChatGPT within the rhinology domain, finding a greater accuracy rate of 66% in ChatGPT responses against AAO—HNS guidelines [25]. ChatGPT knowledge has been shown to be superior in other subspecialties such as head and neck, which is hypothesized to be due to the lack of access to pertinent information in the training data for each LLM [25]. However, it is important to consider other possible explanations for the lack of accuracy, especially considering ChatGPT is trained on data up through October 2023 and we studied its knowledge on the 2022 ICAR guidelines. For example, with LLMs being trained on such a vast amount of data and having access to inaccurate knowledge, accurate information has the potential to be flooded by unreliable information. The variability in accuracy and concordance of LLMs with clinical guidelines has significant implications for their applications in clinical practice. Inconsistent performance limits their reliability as educational resources and clinical support tools in the management of OD, emphasizing the need for regular validation against guidelines and cautious integration to avoid misinformation and suboptimal clinical outcomes. The ability to discern between what is reliable or not is arguably the most important skill for these LLMs to acquire. While overall these accuracy rates are noteworthy, any incorrect information has the potential to inflict patient harm [25]. The disparities in accuracy suggest further evaluation will be required before clinical implementation.

While LLMs represent a new age of innovation, the use of AI to enhance medical care is not a completely new advancement, as the US Food and Drug Administration authorized over 500 AI-based medical devices before the rise of ChatGPT and other LLMs [32]. These LLMs could serve as valuable tools for clinicians, especially in areas with limited access to specialized care, such as focused rhinology practices. By offering centralized

and readily accessible information, these models can democratize the availability of expert knowledge, thereby enhancing access to evidence-based guidelines across underserved regions [33]. Moreover, they can offer significant educational benefits, helping trainees and students familiarize themselves with standardized clinical recommendations. For instance, after further refinement, trainees could refer to ChatGPT for initial information questions regarding olfactory dysfunction and could confirm their findings with an attending physician. By rapidly processing vast amounts of medical literature and guidelines, AI models also provide an advantage for clinicians to gain insight into the most up-to-date literature in their respective fields. Additionally, fine-tuning existing LLMs with curated medical data, such as existing clinical guidelines or peer-reviewed consensus statements, could optimize their accuracy and reliability and allow them to be more suitable for clinical support and medical education. Further validation and oversight will be essential to guarantee that LLMs are clinically verified and complement evidence-based practice.

In the evolving field of olfaction, in which research on treatments such as sodium citrate for olfactory dysfunction is actively being studied, tools such as ChatGPT and Google Gemini can help clinicians stay updated on the latest recommendations. Research on olfactory dysfunction has gained significant importance due to its widespread impact on health and quality of life, particularly in the post-COVID-19 era [34]. Hura et al. discuss that while some treatments for olfactory disorders offer promising preliminary results, the lack of high-quality studies limits the ability of clinicians to make evidence-based recommendations for many therapies [35]. As research on olfaction continues to expand into current literature, LLMs show promise in assisting clinicians in gathering evidence to support future therapies. AI models may even aid rhinologists in gathering resources to update guidelines for olfactory disorders.

Despite their potential, LLMs have drawn several concerns about their use in clinical practice. A single inaccurate response from an LLM could have significant medical consequences, particularly in a field like rhinology, where emerging treatments for various pathologies, such as olfactory dysfunction, require careful consideration of evidence quality. Although the AI platforms provided accurate summaries for several ICAR-O recommendations, discrepancies in other responses raised concerns about their reliability, with potential errors arising from "hallucinations"—the tendency of AI models to generate plausible yet incorrect information [36]. For instance, ChatGPT provided partially accurate information regarding the use of minocycline in patients with olfactory dysfunction. If not checked by knowledgeable experts, discrepancies such as this have the potential to lead to unnecessary or even harmful therapies being recommended. Moreover, we found that both ChatGPT and Google Gemini provided responses that were discordant with some of the guidelines, specifically regarding the use of alternative treatments and supplements, such as omega-3, α-lipoic acid, and Toki-shakuyaku-san. These findings have significant implications for patients utilizing these tools, as they may receive outdated or incorrect medical information. This is particularly concerning in the context of evolving evidence for certain treatments or when patients seek guidance on supplements, where evidence-based guidance is critical. For otolaryngologists utilizing LLMs in clinical settings, this lack of alignment with current research could result in the dissemination of incomplete or inaccurate information to patients.

The lack of transparency regarding the datasets used to train these models introduces further uncertainty. Without clear knowledge of the sources of their training data, it becomes difficult to determine whether the information they provide reflects the most recent clinical standards or includes outdated practices. This risk is compounded by the inability to verify real-time updates to the models' training, which is critical in fields with rapidly changing evidence bases, such as rhinology [37]. The discrepancies observed in LLMs performance against established guidelines may in part stem from limitations inherent in their training data: LLMs are trained on large datasets that may have outdated information or may include inaccuracies, all of which can impact the reliability of their use in clinical practice. Further studies are required to evaluate the long-term impact of their integration into clinical practice, and strong regulatory oversight will be necessary to ensure patient safety and the ethical use of these technologies. Health professionals must be prepared to critically evaluate the information produced by LLMs as their potential impact on clinical medicine increases. Clinicians should be aware of the potential bias and limitations inherent in LLMs and should ensure that the information produced by LLMs is manually checked or verified before use in a clinical setting or as a clinical decision aid. Ultimately, the integration of AI as a support tool, not as an independent tool, incorporated with clinical expertise, would allow for effective and safe medicine to be practiced.

Several potential risks develop with increased reliance on information from LLMs when making clinical decisions. The potential for LLMs to rely on outdated information, deviate from established guidelines, or contain incorrect data could translate into harmful clinical decisions [38]. For example, as of February 2025, the latest version of ChatGPT is trained on data up through June 2024 [39]. Datasets must be frequently updated to ensure that their information contains the latest advancements. Current literature expresses concern about confidently composed but inaccurate LLM outputs, which could lead to decreased accountability and the proliferation of low-quality clinical information [40]. It is very difficult to validate the reliability of LLMs for new and unique clinical scenarios, as models can perform tasks for the first time rather than predefined tasks and may not contain or comprehend the full history behind a clinical scenario or a patient's history. Thus, anticipating failures in the model becomes a difficult feat [38]. Moreover, there are several ethical concerns with the utilization of LLMs in a clinical context. With the complexity of the data they are trained on, there is a potential for biases in the responses produced by LLMs that could even be dangerous or discriminatory [38]. One study showed racial and gender bias in the depiction of medical professionals by DALL-E2, an image-generating model by OpenAI, the creators of ChatGPT [41]. Increased reliance on LLMs without vetting out potentially biased algorithms could lead to biased care for some patient populations. Finally, there are many concerns around privacy and confidentiality, especially when utilizing identifiable patient data. If patient information were to be used in prompts by clinicians, there is a risk of models memorizing this data and even disclosing sensitive

information [42]. There is also the potential for breaches of patient data. Some solutions have been proposed to help prevent this, such as the use of pseudonyms or the training of LLMs to guard specific categories of information [42]. Whatever the solutions may be, there is a great need for regulatory guidelines and policies to be put in place before the implementation of LLMs in the clinical setting.

Our study has several limitations. We utilized two popular LLMs in this investigation; however, comparing other platforms may reveal other strengths and weaknesses of AI platforms. Additionally, two reviewers independently generated questions. Other reviewers may have generated alternate questions or have graded responses differently. Despite these limitations, our study is the first and largest comparative analysis of LLM-generated information on olfactory disorders against clinical guidelines. Future studies should evaluate the reliability of other LLMs to produce accurate information on olfactory disorders and other evolving subjects within otolaryngology for further validation. An additional area for future research is utilizing a more nuanced system to grade the accuracy of LLM responses. Thirdly, future studies should study the performance of LLMs across a wider range of clinical guidelines or within certain subspecialties. Lastly, future studies could assess the performance of LLMs for newly developed treatments for OD and evaluate their ability to process more subtle, context-dependent questions. While LLMs demonstrate potential as clinical adjuncts in helping practitioners quickly attain up-to-date knowledge on nuanced topics within otolaryngology and increase access to information, they are currently inadequate as independent resources for OD.

5 | Conclusion

ChatGPT version 4.o and Google Gemini generate information of variable accuracy in response to questions based on the ICAR guidelines on olfactory dysfunction/disorders. While these platforms are promising for their ability to aid clinicians and trainees, such LLMs cannot reliably be utilized as independent clinical supports or educational tools. We urge clinicians, researchers, and developers to collaborate on implementing standardized guidelines, improving algorithmic transparency, and conducting further interdisciplinary research to enhance the equity and accuracy of LLMs.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. P. Ajmera, N. Nischal, S. Ariyaratne, et al., "Validity of ChatGPT-Generated Musculoskeletal Images," *Skeletal Radiology* 53, no. 8 (2024): 1583–1593, https://doi.org/10.1007/s00256-024-04638-y.

2. D. Chen, R. S. Huang, J. Jomy, et al., "Performance of Multimodal Artificial Intelligence Chatbots Evaluated on Clinical Oncology Cases," *JAMA Network Open* 7, no. 10 (2024): e2437711, https://doi.org/10.1001/jamanetworkopen.2024.37711.

3. D. Menon and K. Shilpa, "'Chatting With ChatGPT': Analyzing the Factors Influencing Users' Intention to Use the Open AI's ChatGPT Using the UTAUT Model," *Heliyon* 9, no. 11 (2023): e20962, https://doi.org/10.1016/j.heliyon.2023.e20962.

4. "Neuro-Symbolic AI: Explainability, Challenges, and Future Trends," accessed November 18, 2024, https://arxiv.org/html/2411.04383v1.

5. A. Bohr and K. Memarzadeh, "The Rise of Artificial Intelligence in Healthcare Applications," *Artificial Intelligence in Healthcare* (2020): 25–60, https://doi.org/10.1016/B978-0-12-818438-7.00002-2.

6. N. H. Shah, D. Entwistle, and M. A. Pfeffer, "Creation and Adoption of Large Language Models in Medicine," *Journal of the American Medical Association* 330, no. 9 (2023): 866–869, https://doi.org/10.1001/jama.2023.14217.

7. M. Quinn, J. D. Milner, P. Schmitt, et al., "Artificial Intelligence Large Language Models Address Anterior Cruciate Ligament Reconstruction: Superior Clarity and Completeness by Gemini Compared With ChatGPT-4 in Response to American Academy of Orthopaedic Surgeons Clinical Practice Guidelines," *Arthroscopy* (2024), https://doi.org/10.1016/j.arthro.2024.09.020.

8. J. R. Bellinger, J. S. De La Chapa, M. W. Kwak, G. A. Ramos, D. Morrison, and B. W. Kesser, "BPPV Information on Google Versus AI (ChatGPT)," *Otolaryngology—Head and Neck Surgery* 170, no. 6 (2024): 1504–1511, https://doi.org/10.1002/ohn.506.

9. R. C. T. Cheong, S. Unadkat, V. Mcneillis, et al., "Artificial Intelligence Chatbots as Sources of Patient Education Material for Obstructive Sleep Apnoea: ChatGPT Versus Google Bard," *European Archives of Oto-Rhino-Laryngology* 281, no. 2 (2024): 985–993, https://doi.org/10.1007/s00405-023-08319-9.

10. A. L. Shaari, I. Herzog, E. Kokush, S. Zabielski, S. Langer, and A. Filimonov, "Exploring the Ability of ChatGPT to Act as a Research Aid in Otolaryngology," *Indian Journal of Otolaryngology and Head & Neck Surgery* 76, no. 6 (2024): 6100–6105, https://doi.org/10.1007/s12070-024-05101-z.

11. M. K. R. Meyer, C. K. Kandathil, S. J. Davis, et al., "Evaluation of Rhinoplasty Information From ChatGPT, Gemini, and Claude for Readability and Accuracy," *Aesthetic Plastic Surgery* (2024), https://doi.org/10.1007/s00266-024-04343-0.

12. A. Maniaci, C. M. Chiesa-Estomba, and J. R. Lechien, "ChatGPT-4 Consistency in Interpreting Laryngeal Clinical Images of Common Lesions and Disorders," *Otolaryngology—Head and Neck Surgery* 171, no. 4 (2024): 1106–1113, https://doi.org/10.1002/ohn.897.

13. V. Bochicchio, S. Mezzalira, N. M. Maldonato, E. Cantone, and C. Scandurra, "Olfactory-Related Quality of Life Impacts Psychological Distress in People With COVID-19: The Affective Implications of Olfactory Dysfunctions," *Journal of Affective Disorders* 323 (2022): 741, https://doi.org/10.1016/j.jad.2022.12.049.

14. S. E. Erskine and C. M. Philpott, "An Unmet Need: Patients With Smell and Taste Disorders," *Clinical Otolaryngology* 45, no. 2 (2020): 197–203, https://doi.org/10.1111/coa.13484.

15. A. Jafari and E. H. Holbrook, "Therapies for Olfactory Dysfunction—An Update," *Current Allergy and Asthma Reports* 22, no. 3 (2022): 21–28, https://doi.org/10.1007/s11882-022-01028-z.

16. E. M. Sina, D. J. Campbell, A. Duffy, et al., "Evaluating ChatGPT as a Patient Education Tool for COVID-19-Induced Olfactory Dysfunction," *OTO Open* 8, no. 3 (2024): e70011, https://doi.org/10.1002/oto2.70011.

17. J. R. Lechien, G. Briganti, and L. A. Vaira, "Accuracy of ChatGPT-3.5 and -4 in Providing Scientific References in Otolaryngology–Head and Neck Surgery," *European Archives of Oto-Rhino-Laryngology* 281, no. 4 (2024): 2159–2165, https://doi.org/10.1007/s00405-023-08441-8.

18. R. R. Orlandi, T. T. Kingdom, T. L. Smith, et al., "International Consensus Statement on Allergy and Rhinology: Rhinosinusitis 2021,"

International Forum of Allergy & Rhinology 11, no. 3 (2021): 213–739, https://doi.org/10.1002/alr.22741.

19. R. A. Ho, A. L. Shaari, P. T. Cowan, and K. Yan, "ChatGPT Responses to Frequently Asked Questions on Ménière's Disease: A Comparison to Clinical Practice Guideline Answers," *OTO Open* 8, no. 3 (2024): e163, https://doi.org/10.1002/oto2.163.

20. "How To Find Reliable Health Information Online," National Institute on Aging. 2023, https://www.nia.nih.gov/health/healthy-aging/how-find-reliable-health-information-online.

21. A. L. Shaari, A. N. Fano, O. Anakwenze, and C. Klifto, "Appraisal of ChatGPT's Responses to Common Patient Questions Regarding Tommy John Surgery," *Shoulder & Elbow* 16, no. 4 (2024): 429–435, https://doi.org/10.1177/17585732241259754.

22. C. Connors, K. Gupta, J. A. Khusid, et al., "Evaluation of the Current Status of Artificial Intelligence for Endourology Patient Education: A Blind Comparison of ChatGPT and Google Bard Against Traditional Information Resources," *Journal of Endourology* 38, no. 8 (2024): 843–851, https://doi.org/10.1089/end.2023.0696.

23. Y. Yoshiyasu, F. Wu, A. K. Dhanda, D. Gorelik, M. Takashima, and O. G. Ahmed, "GPT-4 Accuracy and Completeness Against International Consensus Statement on Allergy and Rhinology: Rhinosinusitis," *International Forum of Allergy & Rhinology* 13, no. 12 (2023): 2231–2234, https://doi.org/10.1002/alr.23201.

24. S. Pandya, T. E. Bresler, T. Wilson, Z. Htway, and M. Fujita, "Decoding the NCCN Guidelines With AI: A Comparative Evaluation of ChatGPT-4.0 and Llama 2 in the Management of Thyroid Carcinoma," *American Surgeon* 91, no. 1 (2025): 31348241269430, https://doi.org/10.1177/00031348241269430.

25. I. Tessler, A. Wolfovitz, E. E. Alon, et al., "ChatGPT's adherence to otolaryngology clinical practice guidelines," *European Archives of Oto-Rhino-Laryngology* 281, no. 7 (2024): 3829–3834, https://doi.org/10.1007/s00405-024-08634-9.

26. M. D. Howell, G. S. Corrado, and K. B. DeSalvo, "Three Epochs of Artificial Intelligence in Health Care," *Journal of the American Medical Association* 331, no. 3 (2024): 242–244, https://doi.org/10.1001/jama.2023.25057.

27. E. Strong, A. DiGiammarino, Y. Weng, et al., "Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations," *JAMA Internal Medicine* 183, no. 9 (2023): 1028–1030, https://doi.org/10.1001/jamainternmed.2023.2909.

28. J. Ge and J. C. Lai, "Artificial Intelligence-Based Text Generators in Hepatology: ChatGPT Is Just the Beginning," *Hepatology Communications* 7, no. 4 (2023): e0097, https://doi.org/10.1097/HC9.0000000000000097.

29. T. A. Rengers, C. A. Thiels, and H. Salehinejad, "Academic Surgery in the Era of Large Language Models: A Review," *JAMA Surgery* 159, no. 4 (2024): 445–450, https://doi.org/10.1001/jamasurg.2023.6496.

30. D. Y. Kwon, A. Wang, M. R. Mejia, et al., "Adherence of a Large Language Model to Clinical Guidelines for Craniofacial Plastic and Reconstructive Surgeries," *Annals of Plastic Surgery* 92, no. 3 (2024): 261–262, https://doi.org/10.1097/SAP.0000000000003757.

31. W. Ahmed, M. Saturno, R. Rajjoub, et al., "ChatGPT Versus NASS Clinical Guidelines for Degenerative Spondylolisthesis: A Comparative Analysis," *European Spine Journal* 33, no. 11 (2024): 4182–4203, https://doi.org/10.1007/s00586-024-08198-6.

32. S. Gottlieb and L. Silvis, "How to Safely Integrate Large Language Models Into Health Care," *JAMA Health Forum* 4, no. 9 (2023): e233909, https://doi.org/10.1001/jamahealthforum.2023.3909.

33. Z. Strika, K. Petkovic, R. Likic, and R. Batenburg, "Bridging Healthcare Gaps: A Scoping Review on the Role of Artificial Intelligence, Deep Learning, and Large Language Models in Alleviating Problems in Medical Deserts," *Postgraduate Medical Journal* 101, no. 1191 (2024): qgae122, https://doi.org/10.1093/postmj/qgae122.

34. N. Neuman, E. Lövestam, J. Karlén, and P. Sandvik, "Sensing Sociality: Disruptions of Social Life When Living With Chemosensory Dysfunctions After COVID-19," *Qualitative Health Research* (2024): 10497323241278551, https://doi.org/10.1177/10497323241278551.

35. N. Hura, D. X. Xie, G. W. Choby, et al., "Treatment of Post-Viral Olfactory Dysfunction: An Evidence-Based Review With Recommendations," *International Forum of Allergy & Rhinology* 10, no. 9 (2020): 1065–1086, https://doi.org/10.1002/alr.22624.

36. B. Meskó and E. J. Topol, "The Imperative for Regulatory Oversight of Large Language Models (Or Generative AI) in Healthcare," *NPJ Digital Medicine* 6, no. 1 (2023): 1–6, https://doi.org/10.1038/s41746-023-00873-0.

37. T. Grote, "Trustworthy Medical AI Systems Need to Know When They don't Know," *Journal of Medical Ethics* 47, no. 5 (2021): 337–338, https://doi.org/10.1136/medethics-2021-107463.

38. I. A. Scott and G. Zuccon, "The New Paradigm in Machine Learning—Foundation Models, Large Language Models and Beyond: A Primer for Physicians," *Internal Medicine Journal* 54, no. 5 (2024): 705–715, https://doi.org/10.1111/imj.16393.

39. Model Release Notes, "OpenAI Help Center," accessed February 23, 2025, https://help.openai.com/en/articles/9624314-model-release-notes.

40. H. S. Yun, I. J. Marshall, T. A. Trikalinos, and B. C. Wallace, "Appraising the Potential Uses and Harms of LLMs for Medical Systematic Reviews," 2023, https://doi.org/10.48550/arXiv.2305.11828.

41. H. S. Choudhry, U. Toor, A. J. Sanchez, and S. I. Mian, "Perception of Race and Sex Diversity in Ophthalmology by Artificial Intelligence: A DALL E-2 Study," *Clin Ophthalmolog (Auckland, NZ)* 17 (2023): 2889–2899, https://doi.org/10.2147/OPTH.S427296.

42. J. C. L. Ong, S. Y. H. Chang, W. William, et al., "Ethical and Regulatory Challenges of Large Language Models in Medicine," *Lancet Digit Health* 6, no. 6 (2024): e428–e432, https://doi.org/10.1016/S2589-7500(24)00061-X.