Evaluating the Potential of AI Chatbots in Treatment Decision-making for Acquired Bilateral Vocal Fold Paralysis in Adults

[*]Emilie A.C. Dronkers, [†]Ahmed Geneid, [*]Chadwan al Yaghchi, and [‡,§,¶]Jerome R. Lechien, *London, UK, †Helsinki, Finland, ‡Mons, ¶Brussels, Belgium, and §Paris, France

Summary: Objectives. The development of artificial intelligence-powered language models, such as Chatbot Generative Pre-trained Transformer (ChatGPT) or Large Language Model Meta AI (Llama), is emerging in medicine. Patients and practitioners have full access to chatbots that may provide medical information. The aim of this study was to explore the performance and accuracy of ChatGPT and Llama in treatment decision-making for bilateral vocal fold paralysis (BVFP).

Methods. Data of 20 clinical cases, treated between 2018 and 2023, were retrospectively collected from four tertiary laryngology centers in Europe. The cases were defined as the most common or most challenging scenarios regarding BVFP treatment. The treatment proposals were discussed in their local multidisciplinary teams (MDT). Each case was presented to ChatGPT-4.0 and Llama Chat-2.0, and potential treatment strategies were requested. The Artificial Intelligence Performance Instrument (AIPI) treatment subscore was used to compare both Chatbots' performances to MDT treatment proposal.

Results. Most common etiology of BVFP was thyroid surgery. A form of partial arytenoidectomy with or without posterior transverse cordotomy was the MDT proposal for most cases. The accuracy of both Chatbots was very low regarding their treatment proposals, with a maximum AIPI treatment score in 5% of the cases. In most cases even harmful assertions were made, including the suggestion of vocal fold medialisation to treat patients with stridor and dyspnea. ChatGPT-4.0 performed significantly better in suggesting the correct treatment as part of the treatment proposal (50%) compared to Llama Chat-2.0 (15%).

Conclusion. ChatGPT and Llama are judged as inaccurate in proposing correct treatment for BVFP. ChatGPT significantly outperformed Llama. Treatment decision-making for a complex condition such as BVFP is clearly beyond the Chatbot's knowledge expertise. This study highlights the complexity and heterogeneity of BVFP treatment, and the need for further guidelines dedicated to the management of BVFP.

Key Words: ChatGPT–Llama–Artificial intelligence–Laryngology–Bilateral vocal fold paralysis–Decision-making.

INTRODUCTION AND OBJECTIVE

Since the introduction of ChatGPT (Chatbot Generative Pre-trained Transformer) to the public in November 2022 by Open AI (San Francisco) large language models (LLM) powered by artificial intelligence (AI) algorithms are changing the way patients and healthcare professionals access and analyze medical information. Chatbots can respond to simple-to-complicated questions and they have the potential to support clinical decision-making utilizing the rapid access to information for both patients and healthcare professionals that is continuously updated.[1] There might be enhanced accuracy as well as chatbots analyze relevant data at once.[2]

ChatGPT is currently one of the most popular conversational chatbots with more than 175 billion parameters in the freely accessible GPT-3.5 version. Llama (Large Language Model Meta AI) is a small but powerful open-source LLM, launched by Meta (Meta AI, San Francisco). The Llama 2.0 model was trained on 2 trillion tokens (part of a text). Both are auto-regressive common-domain LLM that are pretrained on web-scale natural language by predicting the next token, and then fine-tuned to follow large-scale human instructions. Common-domain models were not trained to capture the medical-domain knowledge specifically or in detail, resulting in models that often provide incorrect medical responses.[3] The learning approach for both Llama and ChatGPT is unsupervised, meaning they do not rely on human-labeled data to learn. They train on extensive text data sourced from the internet or other resources, creating new text based on recognized patterns. Llama is based on a wide spectrum of texts, from scientific articles to news stories, while ChatGPT's training primarily comprises internet text like web pages and social media content. This suggests Llama might be a better fit for generating specialized language, such as in the medical field, whereas ChatGPT may perform better in creating informal or conversational language.

Accepted for publication February 20, 2024.

No funding was procured for this work.

From the *National Centre for Airway Reconstruction, Imperial College Healthcare NHS Trust, London, UK; †Department of Otolaryngology and Phoniatrics-Head and Neck Surgery, Helsinki University Hospital and University of Helsinki, Helsinki, Finland; ‡Division of Laryngology and Broncho-esophagology, Department of Otolaryngology-Head Neck Surgery, EpiCURA Hospital, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium; §Department of Otorhinolaryngology and Head and Neck Surgery, Foch Hospital, School of Medicine, UFR Simone Veil, Université Versailles Saint-Quentin-en-Yvelines (Paris Saclay University), Paris, France; and the ¶Department of Otorhinolaryngology and Head and Neck Surgery, CHU Saint-Pierre, Brussels, Belgium.

Address correspondence and reprint requests to Ahmed Geneid, Department of Otolaryngology and Phoniatrics-Head and Neck Surgery, Helsinki University Hospital and University of Helsinki, Helsinki, Finland. E-mail: Ahmed.Geneid@hus.fi

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

0892-1997

https://doi.org/10.1016/j.jvoice.2024.02.020

Bilateral vocal fold paralysis (BVFP) is a complete immobility of both vocal folds caused by damage or dysfunction of both recurrent laryngeal or vagus nerves. It often results in inspiratory dyspnea, due to the paramedian position of the vocal folds and the narrowing of the airway at the glottic level. It is a rare disease with an approximate incidence of 0.95 cases per 100,000 persons.[4] The most frequent etiology in adults is neurogenic injury with resultant bilateral recurrent laryngeal nerve paralysis caused by iatrogenic injury, cancer or neurological disease. The management of acquired BVFP in adults remains controversial and often unsatisfactory as most current surgical approaches, including both reversible and irreversible techniques, lead to a compromise between voice, airway, and swallowing.[5,6] It is important to distinguish between mechanical restriction of cricoarytenoid joint mobility causing immobility of the bilateral vocal folds, and true paralysis following neurogenic injury.

Two recent systematic reviews on the management of BVFP have shown significant heterogeneity among studies and a lack of large cohort-controlled randomized studies.[7,8] To date only one large multicenter registry is known in Germany and Austria, showing highly variable treatment decisions for BVFP.[6] This limits the ability to draw reliable conclusions about the superiority of one technique over others in terms of surgical, voice and respiratory outcomes, complications, and revision surgery. Given the complexity of clinical decision-making in BVFP, AI Chatbots could assist physicians in making informed decisions and improve the overall quality and efficiency of health care.[9]

The objective of the present study was to explore the potential and accuracy of ChatGPT 4.0 and Llama Chat 2.0, in treatment decision-making for BVFP, which may be considered as a complicated laryngological condition.

METHODS

Setting and patients

Twenty recent retrospective cases, treated for BVFP between 2018 and 2023, from four different expert laryngology centers in Europe (Charing Cross Hospital, Imperial Health College, London, United Kingdom; Helsinki University Central Hospital, Helsinki, Finland; Foch Hospital, Paris Saclay University, Paris, France; CHU Saint Pierre, Brussels, Belgium) were anonymously included. The main authors defined the cases as either the most common or the most challenging scenarios of long-term BVFP treatment. Patients suffered with neurogenic BVFP affecting breathing for at least 8 months. The treatment proposals for all cases were discussed in their local multidisciplinary teams (MDT). At least two senior laryngologists and a speech and language pathologist attended the MDT in all four centers, and scope videos were presented as well to illustrate the cases being discussed. The (surgical) management of each case was registered as "Human expert treatment proposal," and the outcome after treatment was also registered for each case. The retrospective use of anonymized patient data for this study has been approved or exempted by the local Institutional Board Review (IRB). Helsinki University Hospital Head and Neck Centre IRB number is 45/HUS/53/2023, University Hospital of Brussels (CHU Saint-Pierre) IRB number is CHU-SP-B0762023230708 and NHS Research Ethics Committee provided proof of exemption for the London cases.

Chatbot interrogation and performance assessment

The data of the 20 cases with complete information including past medical history, symptoms, comorbidities, and clinical examinations were presented to ChatGPT and Llama through standardized sentences. The information on clinical examination in these standardized sentences included a narrative description of scope videos that were presented during the MDT. Both Chatbots were systematically interrogated for primary and secondary (alternative) medical or surgical management proposals, with descriptive open-ended questions, based on the data provided.

An example of one of the standardized sentences is:

> "I have a 59 year old male patient with the following history: Thyroid cancer treated with total thyroidectomy. Nowadays, the symptoms are: dyspnoea during minimal exercise. And the clinical examination shows: Bilateral vocal fold paralysis in adducted position, minimal glottic opening, anteromedially prolapsed right arytenoid cartilage. I have 2 questions for you: 1) What is the primary medical or surgical management that you propose. 2) What is the secondary (alternative) medical or surgical management that you propose?"

Each clinical case was presented to ChatGPT and Llama, and potential treatment strategies were requested, with each answer being collected. To ensure consistency, all questions were entered into ChatGPT by one researcher (JL) and into Llama Chat by another researcher (ED). The same input was used for both Chatbots, and unconditional prompts (new chats) were used for each question to minimize bias. Appendices A and B in Supplementary Information show the input that was used for both Chatbots. ChatGPT version 4.0 was used on an automatic setting. Llama Chat version 2.0 70b was used on an automatic setting with the standard system prompt: "you are a helpful assistant and incorporate any applicable medical guidelines."

The descriptions of the treatment proposals of both Chatbots were compared with the Human expert proposal. An exact match was determined when the Chatbot proposal included all the MDT recommendations, but may have included extra, and potentially harmful, procedures as part of their treatment proposal that were not considered in the Human expert proposal. The treatment subscore of the Artificial Intelligence Performance Instrument (AIPI) was used to assess both Chatbots' performance, compared to Human expert treatment proposal.[10] AIPI is a validated

TABLE 1.
Artificial Intelligence Performance Instrument (AIPI) Treatment Score[10]

0	No adequate therapeutic approach
1	An association of pertinent, necessary, and inadequate therapeutic findings
2	All pertinent but incomplete therapeutic findings
3	All pertinent and necessary therapeutic findings

instrument assessing the accuracy and performance of chatbot in the management of real clinical cases. A maximum score of 3 relates to good clinical performance, a score of 0 relates to inadequate clinical performance (Table 1). The AIPI treatment score was calculated for each case and each Chatbots' response by one researcher (ED) and was based on the Human expert treatment proposal. As the Chatbots were asked for their primary and secondary medical or surgical management proposals, and the Human expert treatment proposal only showed the final MDT outcome, the order of treatment proposals suggested by the Chatbots were not taken into account during calculation of AIPI treatment score and scoring for exact match on treatment proposal.

Statistical analysis

Statistical analyses were performed through Prism Graphpad 10.1.0. Chi-square test was used to analyze differences between exact matches on treatment proposals by Chatbots and Human proposals. To quantify agreement between respectively ChatGPT 4.0 and Llama Chat 2.0 with Human expert treatment proposal, Cohen's kappa was calculated. Spearman correlation analysis was used to test correlation between ChatGPT 4.0 and Llama Chat 2.0 AIPI treatment scores. A 2-sided $P < 0.05$ was considered statistically significant.

RESULTS

The clinical characteristics of the 20 cases with long-standing BVFP recruited from the international expert laryngology centers were reported in Table 2. Among them, 7, 8, and 5 cases were recruited from London, Helsinki, Paris or Brussels, respectively. The complete findings including narrative context provided by the Chatbots and AIPI scores are available in Appendices A and B in Supplementary Information. The mean age of patients was 64 (SD 14) years. There were 11 males and 9 females. Most common etiology of BVFP was thyroid surgery. A form of partial arytenoidectomy with or without posterior transverse cordotomy or supraglottoplasty was the Human expert treatment proposal for most cases.

The results showed a majority of inadequate treatment proposals by both Chatbots for the treatment of real BVFP cases, including vocal fold injection medialisation, arytenoid adduction, type 1 thyroplasty, acupuncture, tracheoesophageal puncture, cricoarytenoid joint fixation and cricopharyngeal myotomy. However, for many cases both Chatbots provided at least the suggestion of a tracheotomy as management for BVFP, which was not the choice of treatment by the Human experts in all cases, but is deemed a valid treatment option for BVFP. However, tracheotomy was also suggested as primary treatment by Llama in cases that already were tracheostomised. ChatGPT was consistent in suggesting a nonspecified form of arytenoidectomy, the main Human treatment proposal, in all 20 cases, whereas Llama suggested this form of management in only four cases. Only in the nonsurgically managed case (case 3) both Chatbots provided the correct treatment proposal without any inadequate suggestions.

ChatGPT significantly performed better in mentioning the exact Human proposed treatment as part of their treatment proposal than Llama (chi-square test, $P < 0.0001$, Figure 1). ChatGPT mentioned the correct form of treatment as based on the Human expert treatment proposal in 50% (n = 10) of the cases (cases 2, 3, 4; cases 7, 8, 9; cases 16, 17, 18; and case 20). Llama proposed the correct form of treatment in 15% (n = 3) of the cases (cases 3, 8, and 18). The AIPI treatment score for each case and each Chatbots' response are shown in Figure 2. Mean AIPI treatment score for ChatGPT was 1.1 (SD 0.45), and 0.85 (SD 0.67) for Llama. Cohen's kappa showed a slight agreement for both ChatGPT (0.110, SE 0.026) and Llama (0.124, SE 0.031) versus Human expert treatment proposal. Spearman's correlation between ChatGPT and Llama AIPI treatment outcomes was 0.47 (95% CI 0.022–0.76).

DISCUSSION

The results support that the tested AI Chatbots, ChatGPT and Llama, are inadequate in proposing the correct treatment for a complex condition such as BVFP. Chat-GPT was already being used for clinical decision making and interesting results have been reported for common otolaryngological conditions.[10,12,13] In this work, we have tested the use of two commonly used Chatbots, Chat-GPT and Llama, for a complex otolaryngological condition. The accuracy of overall treatment proposals by both Chatbots is very low, with an AIPI treatment score of 3 in only 5% of the cases. In most cases, even harmful assertions (hallucinations[11]) were made, including the suggestion of vocal fold medialization techniques to treat patients with BVFP with stridor and dyspnea, and surgical treatment to unrelated anatomical area's (such as cricopharyngeus myotomy). Of the two Chatbots, ChatGPT significantly performed better than Llama in suggesting the correct treatment as part of the treatment proposal in 50% of the cases, versus 15% of the cases. Other studies conducted in otolaryngology practice showed a better accuracy rate of ChatGPT regarding treatment options.[10,12,13] A recent clinical study with general otolaryngology practice cases, including two cases of unilateral and BVFP, showed that treatments proposed by ChatGPT were judged as pertinent in 22% of the cases.[10] The level of difficulty of the clinical cases was not predictive for the performance of ChatGPT.

TABLE 2.
Clinical Cases and Summarized Chatbot Responses

Center	Case number	Gender	Age	Symptoms	Relevant past medical history	Clinical examination	Human expert clinical treatment proposal	Human outcome of case after treatment	ChatGPT, primary treatment*	ChatGPT, secondary treatment*	AIPI score and exact match ChatGPT	Llama, primary treatment*	Llama, secondary treatment*	AIPI score and exact match Llama
London	1	M	59	Dyspnea during minimal exercise	Thyroid cancer treated with total thyroidectomy	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, anteromedially prolapsed right arytenoid cartilage	Unilateral medial posterior CO_2 laser arytenoidectomy and supraglottoplasty	No revision surgery needed	Tracheotomy, vocal fold lateralization, arytenoidectomy or arytenoidopexy, medication	Injection medialization, speech therapy, CPAP, reinnervation procedure	AIPI: 1 Exact match: No	Tracheotomy, laryngectomy	Speech therapy, botox injection	AIPI: 1 Exact match: No
London	2	M	85	Stridor, dyspnea during minimal exercise, dysphagia	Medullary CVA, PEG fed	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, normal mobility cricoarytenoid joints	Unilateral medial posterior CO_2 laser arytenoidetomy	No revision surgery needed	Tracheotomy, vocal fold lateralization, corticosteroids, pulmonary supportive care	Injection medialization, arytenoidectomy or arytenoidopexy, speech therapy, CPAP	AIPI: 1 Exact match: Yes	Speech therapy and dietician input, tracheotomy	PEG placement, laryngectomy, pulmonary rehabilitation, palliative care	AIPI: 1 Exact match: No
London	3	M	56	Dyspnea during heavy exercise	Congenital bilateral vocal cord palsy, status after 4× unilateral medial posterior CO_2 laser arytenoidectomy	Bilateral vocal fold paralysis in adducted position, reasonable glottic opening, normal palpation cricoarytenoid joints	No surgical treatment	No surgical treatment needed	Monitoring, revision arytenoidectomy, injection laryngoplasty, tracheotomy	Lateralization procedure, reinnervation procedure, CPAP, pulmonary medication	AIPI: 3 Exact match: Yes	Monitoring, speech therapy, pulmonary rehabilitation	Botox injection, tracheotomy	AIPI: 3 Exact match: Yes
London	4	F	78	Stridor, tracheostomised	Multisystem atrophy (MSA)	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, normal palpation cricoarytenoid joints	Bilateral medial posterior CO_2 laser arytenoidectomy	Patient decannulated, no revision surgery needed	Optimization tracheostomy care, pulmonary rehabilitation, speech therapy, management of MSA	Arytenoidectomy or arytenoidopexy, injection medialization, speech therapy, palliative care	AIPI: 1 Exact match: Yes	Monitoring	Tracheotomy, laryngeal laser cordectomy	AIPI: 1 Exact match: No
London	5	F	69	Stridor, dyspnea during moderate exercise, dysphonia	Thyroid cancer treated with total thyroidectomy, status after 3× unilateral medial posterior CO_2 laser arytenoidectomy	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, normal palpation cricoarytenoid joints, anteromedially prolapsed left arytenoid cartilage	Unilateral medial posterior CO_2 laser arytenoidetomy and supraglottoplasty	Revision surgery needed, same procedure	Further arytenoidectomy, tracheotomy, revision of previous surgical site, pulmonary rehabilitation	Injection medialization, lateralization procedure, speech therapy, CPAP	AIPI: 1 Exact match: No	Monitoring, speech therapy, pulmonary rehabilitation, laryngotomy	Botox injection, tracheotomy, psychological counseling, acupuncture	AIPI: 1 Exact match: No

TABLE 2. (*Continued*)

Center	Case number	Gender	Age	Symptoms	Relevant past medical history	Clinical examination	Human expert clinical treatment proposal	Human outcome of case after treatment	ChatGPT, primary treatment*	ChatGPT, secondary treatment*	AIPI score and exact match ChatGPT	Llama, primary treatment*	Llama, secondary treatment*	AIPI score and exact match Llama
London	6	F	56	Dyspnea during moderate exercise	Thyroid goiter treated with total thyroidectomy, status after 1× unilateral medial posterior CO_2 laser arytenoidectomy	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, fixed right cricoarytenoid joint, anteromedially prolapsed right arytenoid cartilage	Unilateral medial posterior CO_2 laser arytenoidectomy and supraglottoplasty	Revision surgery needed, same procedure	Tracheotomy, revision arytenoidectomy or arytenoidopexy, injection medialization, pulmonary rehabilitation	Lateralization procedure, speech therapy, CPAP, reinnervation procedure	AIPI: 1 Exact match: No	Speech therapy, botox injection, thyroplasty with implant	Laryngeal physiotherapy, arytenoid adduction, thyroplasty with implant	AIPI: 0 Exact match: No
London	7	F	50	Stridor, dyspnea during moderate exercise	Charcot Marie Tooth disease	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, normal palpation cricoarytenoid joints	Bilateral medial posterior CO_2 laser arytenoidectomy	No revision surgery needed	Tracheotomy, arytenoidectomy or arytenoidopexy, injection medialization, pulmonary rehabilitation	Lateralization procedure, speech therapy, CPAP, reinnervation procedure	AIPI: 1 Exact match: Yes	Pulmonary rehabilitation, speech therapy, vocal cord medialization, tracheotomy	Botox injection, laryngeal mask airway, home oxygen therapy, psychological support, monitoring	AIPI: 1 Exact match: No
Helsinki	8	F	72	Dyspnea during minimal exercise	Thyroid goiter treated with total thyroidectomy, morbid obesity, OSA on CPAP	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, normal palpation cricoarytenoid joints	Unilateral anteromedial CO_2 laser arytenoidectomy	No revision surgery needed	Tracheotomy, pulmonary rehabilitation, optimization CPAP, injection medialization	Lateralization procedure, weight management, speech therapy, arytenoidectomy or arytenoidopexy	AIPI: 1 Exact match: Yes	Speech therapy, pulmonary rehabilitation, dietician input, exercise program	Thyroplasty with implant, cricopharyngeal myotomy, arytenoidectomy, palliative care	AIPI: 1 Exact match: Yes
Helsinki	9	M	70	Dyspnea during minimal exercise	Fascia augmentation bilateral vocal cords for aphonia	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, atrophic vocal folds, normal palpation cricoarytenoid joints	Unilateral anteromedial CO_2 laser arytenoidectomy	Improved breathing, but aphonia, patient had unilateral vocal fold injection augmentation with fascia	Tracheotomy, reversal of fascia augmentation, pulmonary rehabilitation, injection medialization	Lateralization procedure, speech therapy, arytenoidectomy or arytenoidopexy, reinnervation procedure	AIPI: 1 Exact match: Yes	Pulmonary rehabilitation, noninvasive ventilation, tracheotomy	Vocal fold injection, thyroplasty with implant, cricoarytenoid joint stabilization	AIPI: 1 Exact match: No

TABLE 2. *(Continued)*

Center	Case number	Gender	Age	Symptoms	Relevant past medical history	Clinical examination	Human expert clinical treatment proposal	Human outcome of case after treatment	ChatGPT, primary treatment*	ChatGPT, secondary treatment*	AIPI score and exact match ChatGPT	Llama, primary treatment*	Llama, secondary treatment*	AIPI score and exact match Llama
Helsinki	10	M	49	Stridor, tracheosto-mised	Thyroid goiter treated with total thyroi-dectomy	Bilateral vocal fold paralysis in adducted position, no glottic opening, false cords adducted as well, normal palpation cricoarytenoid joints	Unilateral anterome-dial CO_2 laser arytenoi-dectomy and posterior transverse cor-dotomy	No revision surgery needed	Optimization tracheostomy care, arytenoi-dectomy or arytenoido-pexy, pulmonary rehabilitation, revision surgery	Lateralization procedure, injection medialization, speech therapy, reinnervation procedure	AIPI: 1 Exact match: No	Tracheostomy tube exchange, monitoring, speech therapy	Vocal fold injection medialization, thyroplasty with implant, laryngectomy, tracheotomy decannulation	AIPI: 1 Exact match: No
Helsinki	11	M	77	Stridor	Childhood polio infection	Bilateral vocal fold paralysis in adducted position, minimal glottic opening	Unilateral anterome-dial CO_2 laser arytenoi-dectomy and posterior transverse cor-dotomy	No revision surgery needed	Tracheotomy, pulmonary rehabilitation, speech therapy, optimization general health	Injection medialization, arytenoi-dectomy or arytenoido-pexy, lateralization procedure, CPAP	AIPI: 1 Exact match: No	Monitoring, speech therapy, botox injection, thyroplasty with implant	Vocal fold injection medialization, arytenoid adduction, arytenoi-dectomy, tracheoeso-phageal puncture, laryngotomy, electrocardio-gram-accelerometer device, stem cell therapy, gene therapy	AIPI: 1 Exact match: No
Helsinki	12	F	81	Stridor	Thyroid goiter treated with total thyroi-dectomy, morbid obesity, OSA on CPAP	Bilateral vocal fold paralysis, right fold adducted position, left fold abducted position, insufficient glottic opening, normal palpation cricoarytenoid joints	Unilateral anterome-dial CO_2 laser arytenoi-dectomy and posterior transverse cor-dotomy	No revision surgery needed	Tracheotomy, optimization of CPAP, pulmonary rehabilitation and weight management, injection medialization right vocal fold	Arytenoidect-omy or arytenoido-pexy, lateralization of adducted vocal fold, speech therapy, bariatric consultation	AIPI: 1 Exact match: No	Optimize CPAP therapy	Botox injection, thyroplasty with implant, cricothyroi-dotomy, tracheotomy	AIPI: 1 Exact match: No

TABLE 2. (*Continued*)

Center	Case number	Gender	Age	Symptoms	Relevant past medical history	Clinical examination	Human expert clinical treatment proposal	Human outcome of case after treatment	ChatGPT, primary treatment*	ChatGPT, secondary treatment*	AIPI score and exact match ChatGPT	Llama, primary treatment*	Llama, secondary treatment*	AIPI score and exact match Llama
Helsinki	13	M	41	Stridor, tracheostomised	Posthanging injury and carotid dissection right side	Bilateral vocal fold paralysis in adducted position, no glottic opening, normal palpation cricoarytenoid joints	Unilateral anteromedial CO_2 laser arytenoidectomy and posterior transverse cordotomy	No revision surgery needed	Optimization tracheostomy care, arytenoidectomy or arytenoidopexy, pulmonary rehabilitation, neurological and vascular evaluation	Lateralization procedure, injection medialization, speech therapy, CPAP	AIPI: 1 Exact match: No	Tracheotomy, monitoring, decompression of recurrent laryngeal nerves	Tracheotomy, laryngotomy, speech therapy	AIPI: 0 Exact match: No
Helsinki	14	F	75	Stridor	Charcot Marie Tooth disease	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, normal palpation cricoarytenoid joints	Unilateral anteromedial CO_2 laser arytenoidectomy and posterior transverse cordotomy	Patient decannulated, no revision surgery needed	Tracheotomy, pulmonary rehabilitation, speech therapy, optimization general health	Injection medialization, arytenoidectomy or arytenoidopexy, lateralization procedure, CPAP	AIPI: 1 Exact match: No	Monitoring	Thyroplasty with implant, laryngectomy	AIPI: 0 Exact match: No
Helsinki	15	F	43	Stridor, tracheostomised	Thyroid goiter treated with total thyroidectomy	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, normal palpation cricoarytenoid joints	Unilateral anteromedial CO_2 laser arytenoidectomy and posterior transverse cordotomy	No revision surgery needed	Optimization tracheostomy care, arytenoidectomy or arytenoidopexy, pulmonary rehabilitation, revision surgery	Lateralization procedure, injection medialization, speech therapy, CPAP	AIPI: 1 Exact match: No	Monitoring, maintaining tracheotomy tube, noninvasive ventilation	Speech therapy, vocal fold injection medialization with Teflon or collagen, tracheoesophageal puncture, laryngotomy, cricothyroidotomy	AIPI: 1 Exact match: No
Paris/ Brussels	16	M	72	Stridor, tracheostomised	Thyroid cancer treated with total thyroidectomy	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, fixed cricoarytenoid joints	Bilateral anteromedial arytenoidectomy	Revision surgery needed, same procedure	Optimization tracheostomy care, pulmonary rehabilitation, tracheostomy revision, speech therapy	Lateralization procedure, injection medialization, arytenoidectomy or arytenoidopexy, CPAP	AIPI: 1 Exact match: Yes	Tracheotomy	Monitoring, speech therapy, arytenoid adduction, thyroplasty with implant, cricoarytenoid joint fixation, tracheoesophageal puncture	AIPI: 0 Exact match: No

TABLE 2. *(Continued)*

Center	Case number	Gender	Age	Symptoms	Relevant past medical history	Clinical examination	Human expert clinical treatment proposal	Human outcome of case after treatment	ChatGPT, primary treatment*	ChatGPT, secondary treatment*	AIPI score and exact match ChatGPT	Llama, primary treatment*	Llama, secondary treatment*	AIPI score and exact match Llama
Paris/ Brussels	17	M	68	Stridor, tracheosto-mised	Thyroid goiter treated with total thyroi-dectomy	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, fixed cricoarytenoid joints	Bilateral anterome-dial arytenoi-dectomy	Revision surgery needed, bilateral cor-dotomy, no decannu-lation	Optimization tracheostomy care, pulmonary rehabilitation, tracheostomy revision, speech therapy	Lateralization procedure, injection medialization, arytenoi-dectomy or arytenoido-pexy, CPAP	AIPI: 1 Exact match: Yes	Securing existing tracheotomy tube, speech therapy, botox injection, laryngeal framework surgery	Vocal fold injection, tracheotomy tube exchange, laryngotomy, temporary pacemaker implantation, speech therapy	AIPI: 1 Exact match: No
Paris/ Brussels	18	M	80	Stridor, tracheosto-mised	Thyroid goiter treated with total thyroi-dectomy	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, fixed cricoarytenoid joints	Bilateral anterome-dial arytenoi-dectomy	Patient decannu-lated, no revision surgery needed	Optimization tracheostomy care, pulmonary rehabilitation, speech therapy, monitoring	Lateralization procedure, injection medialization, arytenoi-dectomy or arytenoido-pexy, CPAP	AIPI: 1 Exact match: Yes	Tracheotomy tube change, corticosteroid treatment, temporary lateral neck stabilization device	Laser cordectomy, arytenoi-dectomy, speech therapy, dietician input, psychological support	AIPI: 1 Exact match: Yes
Paris/ Brussels	19	M	56	Dyspnea during moderate exercise	Idiopathic postviral (COVID-19) bilateral vocal fold paralysis	Bilateral vocal fold paralysis in adducted position, minimal glottic opening, fixed right cricoarytenoid joint	Unilateral posterior transverse cor-dotomy	No revision surgery needed	Tracheotomy, pulmonary rehabilitation, speech therapy, monitoring	Injection medialization, arytenoi-dectomy or arytenoido-pexy, CPAP	AIPI: 1 Exact match: No	Monitoring, speech therapy, botox injection	Thyroplasty with implant, arytenoi-dectomy, tracheotomy	AIPI: 1 Exact match: No
Paris/ Brussels	20	F	46	Dyspnea during moderate exercise, tracheosto-mised	Thyroid goiter treated with total thyroi-dectomy	Bilateral vocal fold paralysis in adducted position, no glottic opening, fixed cricoarytenoid joints	Bilateral anterome-dial arytenoi-dectomy	Patient decannu-lated, no revision surgery needed	Optimization tracheostomy care, pulmonary rehabilitation, speech therapy, monitoring	Injection medialization, arytenoi-dectomy or arytenoido-pexy, lateralization procedure, CPAP	AIPI: 1 Exact match: Yes	Speech therapy, pulmonary rehabilitation	Botox injection, tracheoeso-phageal puncture, thyroplasty with implant, cricothyroi-dotomy	AIPI: 0 Exact match: No

* The last columns of this table show the summarized answers provided by ChatGPT 4.0 and Llama Chat 2.0.

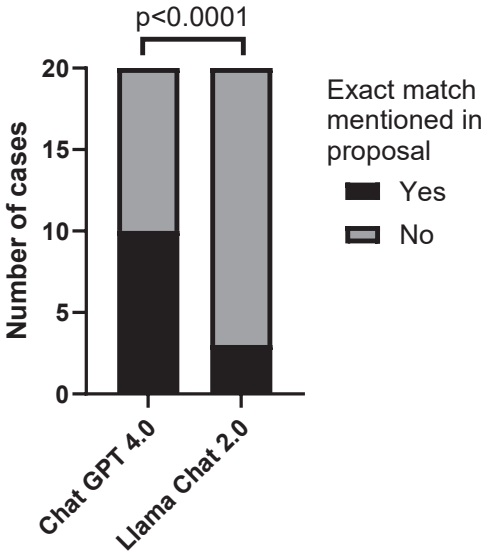

FIGURE 1. Exact match treatment mentioned in management proposal by ChatGPT 4.0 and Llama 2.0. Chi-square test 19.216 with 1 degree of freedom, 2-tailed *P* value < 0.0001.

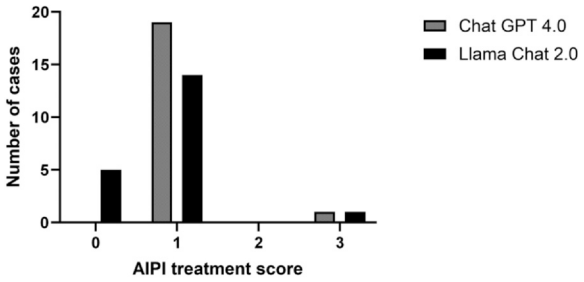

FIGURE 2. AIPI treatment scores of ChatGPT 4.0 and Llama Chat 2.0. AIPI treatment score[10]: (0) No adequate therapeutic approach, (1) An association of pertinent, necessary, and inadequate therapeutic findings, (2) All pertinent but incomplete therapeutic findings, (3) All pertinent and necessary therapeutic findings. AIPI, Artificial Intelligence Performance Instrument.

Another recent study, on accuracy of ChatGPT in head and neck oncological board decisions, showed an adequate judgment on therapeutic management in even 65% of cases, using the AIPI score.[12] However, in this study ChatGPT had some difficulties with proposing adequate therapeutic options for complicated laryngeal cancer cases. The findings of these studies and our present study suggest that the performance of ChatGPT is very unpredictable and varies widely from one clinical case to another, likely based on the complexity of the case and the fact that BVFP is rare in comparison to other more common otorhinolaryngologic practice cases.

The primary limitation of the present study is the low number of clinical cases. Also, in this study, only one subscore of the AIPI was used, which was not validated for use on its own.[10] The results of the present study suggest that treatment decision-making for BVFP is beyond the Chatbot's knowledge and expertise. Their knowledge is trained on information available on the Internet, and two recent systematic reviews on management of BVFP have shown significant heterogeneity among studies, highly variable forms of treatment, and a lack of standardized treatment protocols.[7,8] Based on the available literature, no reliable conclusions can be drawn about the superiority of one technique over others in terms of surgical, voice and respiratory outcomes, complications, and revision surgery. As the condition is rare, and surgical treatment is highly dependent on the experience and skills of the surgeon, and available equipment, it is very difficult to design a randomized study to answer these questions. The lack of guidelines dedicated to the management of BVFP is a shortfall in literature, and this study highlights the complexity and heterogeneity of BVFP management. Guidelines may also improve the body of evidence used to assess the performance and accuracy of the chatbots. There is a need for expert consensus on treatment approach for BVFP. The primary strength of this study were the comparison of ChatGPT and Llama and the inclusion of real clinical cases. To the best of our knowledge, there is no study assessing the accuracy of Llama in otolaryngology. Moreover, to date, only few studies included real clinical cases.[10,12–14] As our study shows that the performance of Llama is even worse than ChatGPT in terms of accuracy of treatment proposal for a complex condition as BVFP, there would be a need for more clinical studies comparing the accuracy of multiple chatbots in clinical decision making.

In the present study, the two most used Chatbots, ChatGPT and Llama, were evaluated. However, these are not specifically designed for the medical domain, although in practice they might be counseled by doctors and patients for medical advice. Glass AI (San Francisco), and independent start-up, combines LLMs with clinical guidelines created and peer-reviewed by an academic physician team employed by Glass AI, to create differential diagnosis and evidence-based clinical plan outputs.[15] It also collects user data to continually enhance the underlying LLMs, which might cause ethical issues with professional healthcare users entering confidential patient details without permission. The platform was picked up early across social media platforms, particularly on X (formerly Twitter), garnering significant interest from physicians, nurses, and medical trainees, and has signed up more than 59,000 users. Google is developing and testing Med-PaLM, a LLM designed to provide high-quality answers to medical questions. It is not accessible for the wide public yet, but was recently published in Nature, showing their AI system to be the first to surpass the pass mark on US Medical License Exam (USMLE) style questions.[16] However, several other AI healthcare startups, including Babylon Health (which was supported by the UK National Health Service) and Cass, have faced scrutiny for assertions regarding the superior capabilities of their AI systems and the potential dissemination of harmful advice.[17]

All healthcare professionals need to understand the capabilities of advanced AI tools, although clinical

judgment and human expertise cannot be replaced by a chatbot. As shown in the present study, chatbots can provide harmful assertions (hallucinations) about areas beyond the Chatbot's knowledge expertise or even fabric sources, potentially causing medical malpractice.[11] Assessing the safety and efficacy of these chatbots in answering clinical questions is critical to determine their suitability in facilitating complex decision-making.

CONCLUSION

The tested AI Chatbots, ChatGPT 4.0 and Llama 2.0, are inadequate in proposing the appropriate treatment for BVFP. Of the two Chatbots, ChatGPT significantly outperformed Llama in proposing a correct form of treatment for BVFP. Both chatbots mentioned many inadequate, and even harmful forms of treatment alongside the correct one for the same cases. Treatment decision-making for a complex condition as BVFP is clearly beyond the Chatbot's knowledge expertise and the lack of guidelines dedicated to the management of BVFP partly explains their difficulties with suggesting the correct form of treatment for individual real patient cases. There is a need for expert consensus on treatment approach for BVFP. An international consensus study using a Delphi method and focus groups has been designed by our research group to develop a guideline protocol regarding diagnosis, treatment options and patient selection for treatment of acquired BVFP in adults.

Informed consent

Not applicable.

Author contributions

Emilie A.C. Dronkers: design, acquisition of data, data analysis & interpretation, drafting, final approval, and accountability for the work; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Ahmed Geneid: design, acquisition of data; 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. Chadwan Al Yaghchi: design, data analysis & 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, acquisition of data, data analysis & 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.

Declaration of Competing Interest

The authors have no conflict of interest.

Acknowledgments

None.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jvoice.2024.02.020.

References

1. Hill-Yardin EL, Hutchinson MR, Laycock R, et al. A Chat(GPT) about the future of scientific publishing. *Brain Behav Immun.* 2023;110:152–154. https://doi.org/10.1016/j.bbi.2023.02.022.
2. Goodman RS, Patrinely JR, Stone Jr CA, et al. Accuracy and reliability of chatbot responses to physician questions. *JAMA Netw Open.* 2023;6:e2336483. https://doi.org/10.1001/jamanetworkopen.2023.36483.
3. Li Y, Li Z, Zhang K, et al. ChatDoctor: a medical chat model fine-tuned on a Large Language Model Meta-AI (LLaMA) using medical domain knowledge. *Cureus.* 2023;15:e40895. https://doi.org/10.7759/cureus.40895.
4. Djugai S, Boeger D, Buentzel J, et al. Chronic vocal cord palsy in Thuringia, Germany: a population-based study on epidemiology and outcome. *Eur Arch Otorhinolaryngol.* 2014;271:329–335. https://doi.org/10.1007/s00405-013-2655-1.
5. Sapundzhiev N, Lichtenberger G, Eckel HE, et al. Surgery of adult bilateral vocal fold paralysis in adduction: history and trends. *Eur Arch Otorhinolaryngol.* 2008;265:1501–1514. https://doi.org/10.1007/s00405-008-0665-1.
6. Nawka T, Gugatschka M, Kolmel JC, et al. Therapy of bilateral vocal fold paralysis: real world data of an international multi-center registry. *PLoS One.* 2019;14:e0216096. https://doi.org/10.1371/journal.pone.0216096.
7. de Almeida RBS, Costa CC, Silva Duarte PLE, et al. Surgical treatment applied to bilateral vocal fold paralysis in adults: systematic review. *J Voice.* 2023;37:289.e1–289.e13. https://doi.org/10.1016/j.jvoice.2020.11.018.
8. Titulaer K, Schlattmann P, Guntinas-Lichius O. Surgery for bilateral vocal fold paralysis: systematic review and meta-analysis. *Front Surg.* 2022;22:956338. https://doi.org/10.3389/fsurg.2022.956338.
9. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Inform Assoc.* 2023;30:1237–1245. https://doi.org/10.1093/jamia/ocad072.
10. Lechien JR, Maniaci A, Gengler I, et al. Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the Artificial Intelligence Performance Instrument (AIPI). *Eur Arch Otorhinolaryngol.* 2024;281:2063–2079. https://doi.org/10.1007/s00405-023-08219-y.
11. Vaishya R, Misra A, Vaish A. ChatGPT: is this version good for healthcare and research? *Diabetes Metab Syndr.* 2023;17:102744. https://doi.org/10.1016/j.dsx.2023.102744.
12. Lechien JR, Chiesa-Estomba CM, Baudouin R, et al. Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. *Eur Arch Otorhinolaryngol.* 2024;281:2105–2114. https://doi.org/10.1007/s00405-023-08326-w.
13. Lechien JR, Georgescu BM, Hans S, et al. ChatGPT performance in laryngology and head and neck surgery: a clinical case-series. *Eur*

Arch Otorhinolaryngol. 2024;281:319–333. https://doi.org/10.1007/s00405-023-08282-5.

14. Vaira LA, Lechien JR, Abbate V, et al. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. *Otolaryngol Head Neck Surg.* 2023. https://doi.org/10.1002/ohn.489. Online ahead of print.
15. Wiggers K. Glass health is building an AI for suggesting medical diagnoses; 2023. Available at: ⟨https://techcrunch.com/2023/09/08/glass-health-is-building-an-ai-for-suggesting-medical-diagnoses/⟩. Accessed December 12, 2023.
16. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature.* 2023;620:172–180. https://doi.org/10.1038/s41586-023-06291-2.
17. Kobie N. Babylon Disrupted the UK's Health System. Then It Left; 2023. Available at: ⟨https://www.wired.co.uk/article/babylon-disrupted-uk-health-system-then-left⟩. Accessed December 12, 2023.