AMERICAN ACADEMY OF OTOLARYNGOLOGY–HEAD AND NECK SURGERY FOUNDATION

Otolaryngology–Head and Neck Surgery
2023, Vol. 00(00) 1–4

DOI: 10.1002/ohn.526
http://otojournal.org
WILEY

Is ChatGPT-4 Accurate in Proofread a Manuscript in Otolaryngology–Head and Neck Surgery?

Jerome R. Lechien, MD, PhD, MS[1,2,3,4]*, Amy Gorton, MS[5]*, Jean Robertson, PhD[5†], and Luigi A. Vaira, MD[6,7†]

Abstract

ChatGPT is a new artificial intelligence-powered language model of chatbot able to help otolaryngologists in clinical practice and research. We investigated the ability of ChatGPT-4 in the editing of a manuscript in otolaryngology. Four papers were written by a nonnative English otolaryngologist and edited by a professional editing service. ChatGPT-4 was used to detect and correct errors in manuscripts. From the 171 errors in the manuscripts, ChatGPT-4 detected 86 errors (50.3%) including vocabulary (N = 36), determiner (N = 27), preposition (N = 24), capitalization (N = 20), and number (N = 11). ChatGPT-4 proposed appropriate corrections for 72 (83.7%) errors, while some errors were poorly detected (eg, capitalization [5%] and vocabulary [44.4%] errors. ChatGPT-4 claimed to change something that was already there in 82 cases. ChatGPT demonstrated usefulness in identifying some types of errors but not all. Nonnative English researchers should be aware of the current limits of ChatGPT-4 in the proofreading of manuscripts.

Keywords

artificial, ChatGPT-4, correction, GPT-4, head neck, intelligence, otolaryngology, proofread, scientific, surgery, writing

Received May 4, 2023; accepted August 18, 2023.

A chatbot is defined as an electronic system that simulates conversations by responding to keywords or phrases.[1] The Chatbot Generative Pre-trained Transformer (ChatGPT) is a new artificial intelligence-powered language model that was developed by OpenAI to use algorithms to respond to simple-to-complicated questions.[2] The version 4.0, ChatGPT-4, was able to pass exams from medical schools,[3] and could help the physician in consultation, scientific, and administrative tasks.[4-6] To date, there are no publications about the usefulness of ChatGPT-4 in the editing of scientific manuscripts written by nonnative English researchers. The objective of this study was to investigate the ability of ChatGPT-4 to proofread a manuscript in otolaryngology–head and neck surgery.

Methods

The authors selected 4 manuscripts that were written by one native-speaker of French (J.R.L.) with a B2 level in English according to the Common European Framework of Reference for Languages.[7] These 4 drafts were edited by a professional editing Service from Nature Journals (American Journal Expert). The American journal expert proposes professional editing for publications in all medical fields by native English speakers in the field, in this instance, otolaryngology.

The manuscripts included 2 reviews,[8,9] and 2 prospective studies,[10,11] which were published in the past few years. The authors randomly selected 2 of the following paper parts: abstract; introduction; methods; results; discussion and conclusion and asked ChatGPT-4 (March 2023) to proofread the paper section for several types of errors. In practice, the first author of the publication (J.R.L.) used the ChatGPT-4 interface

[1]Department of Otolaryngology–Head Neck Surgery, Division of Laryngology and Broncho–esophagology, EpiCURA Hospital, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium
[2]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
[3]Department of Otorhinolaryngology and Head and Neck Surgery, CHU de Bruxelles, CHU Saint-Pierre, School of Medicine, Université Libre de Bruxelles, Brussels, Belgium
[4]Polyclinique Elsan de Poitiers, Poitiers, France
[5]Faculty of Translation and Interpretation (FTI-EII), University of Mons, Mons, Belgium
[6]Maxillofacial Surgery Operative Unit, Department of Medicine, Surgery and Pharmacy, University of Sassari, Sassari, Italy
[7]PhD School of Biomedical Sciences, Department of Biomedical Sciences, University of Sassari, Sassari, Italy

*These authors contributed equally to this article and may be joined as cofirst authors.
†These authors contributed equally to the paper and may be joined as cosenior authors.

Corresponding Author:
Jerome R. Lechien, MD, PhD, MS, Department of Human Anatomy and Experimental Oncology, Faculty of Medicine, UMONS Research Institute for Health Sciences and Technology, Avenue du Champ de mars, 6, B7000 Mons, Belgium.
Email: Jerome.Lechien@umons.ac.be

accessible via the API (https://chat.openai.com). For each text, the following instruction was input in the chat: "Could you correct as a native US speaker the following text for grammar, spelling, and all types of errors? The text must be perfect from an orthographic standpoint." The corrected text was analyzed by experts. The following instruction was additionally input to verify the text changes performed by GPT-4: "please, list the correction that you have made." The same texts were submitted to 2 native English speakers (**Figure 1**), both language professionals, for editing. They are employed by the Faculty of Translation and Interpretation of UMONS (Belgium).

The institutional review board of CHU Saint-Pierre was not required for this study (ref.CHUST23).

Language Outcomes

They assessed the ChatGPT-4's detection and proposed corrections for the types of mistakes available in Supplemental Appendix 1, available online. A ratio between the number of initial draft mistakes and the number of mistakes detected by ChatGPT-4 was evaluated. Authors assessed the quality of the proposed corrections. The final versions provided by the professional Editing service were used as a control in case of doubt or inconsistencies between the 2 native English speaker experts. The test-retest reliability of ChatGPT-4's analysis was assessed through a request of correction of the same text a few days apart.

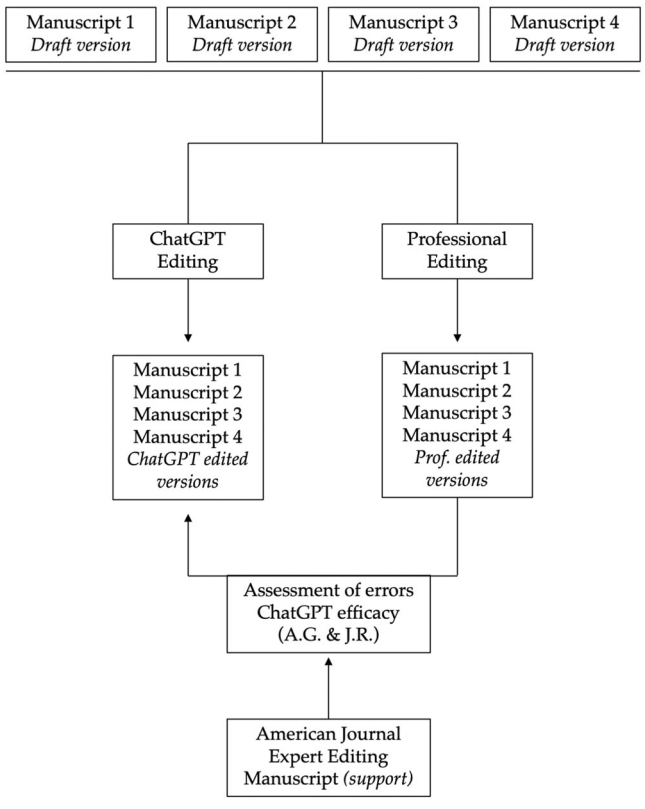

Figure 1. Chartflow. ChatGPT, Chatbot Generative Pre-trained Transformer.

Results

Several parts of 4 papers were analyzed by ChatGPT-4 (**Figure 1**). The nonnative English author committed 171 errors in the manuscripts. The five most common types of errors were vocabulary (N = 36), determiner (N = 27), preposition (N = 24), capitalization (N = 20), and number (N = 11). The proportion and types of errors detected or undetected by ChatGPT-4 for each manuscript are available in **Table 1**. Of the 171 errors, ChatGPT-4 detected 86 errors (50.3%) and proposed an appropriate correction for 72 (83.7%). Among the five most common errors, the most common detected errors included number (81.8%), determiner (66.7%), and preposition (62.5%) errors. Possessive structure and vocabulary errors were detected in 38.9% and 44.4% of cases, while capitalization errors were not detected in 95% of cases. Among the others, ChatGPT-4 detected more easily the following errors: subject/verb agreement (N = 2/2); spelling (N = 3/3); punctuation (N = 4/4), and negation (N = 1/1) but these errors were uncommon in the manuscripts. ChatGPT-4 claimed to change something that was already there in 32, 7, 39, and 4 cases of manuscripts 1, 2, 3, and 4, respectively. These changes included the addition of comma, hyphen, or capital which were already there. The test-retest reliability of ChatGPT-4 showed that it similarly corrected the first manuscript at Day 1 and day 7. The detection rate of errors was similar at these times. An example of interface is available in Supplemental Appendix 2, available online.

Discussion

Most highly ranked academic journals in otolaryngology head and neck surgery use English as the means of communication, which means that the otolaryngologist who wishes to have research internationally recognized needs to publish in English. Depending on the background of the writer, this task may be challenging. Thus, manuscripts from non-English speaking countries have a much lower acceptance rate than those from English-speaking countries (29.1% vs 40.3%).[12] Language may not be the only reason for rejection but an Editage Global Author Survey showed that 76% of non-English native speaker scientific authors find it difficult to prepare a manuscript in English.[13] ChatGPT-4 was recently suggested as an artificial means of helping to write a scientific paper[5,14] but, to date, there has been no study investigating the ability of ChatGPT-4 to proofread a scientific paper.

The primary findings of this preliminary paper suggest that ChatGPT-4 version 4.0 may detect and improve a drafted manuscript but is unable to detect all errors, particularly possessive structures, vocabulary errors, and unclear meaning. Moreover, our preliminary data support

Table 1. Ability of ChatGPT-4 in the Proofreading

	Manuscript 1			Manuscript 2			Manuscript 3			Manuscript 4					
		ChatGPT-4 errors			ChatGPT-4 errors			ChatGPT-4 errors			ChatGPT-4 errors			ChatGPT-4 errors	
Error types	N	De	Unde	N	De	Unde	N	De	Unde	N	De	Unde	N tot	De	Unde	%
Subject/verb agreement	1	1	0	0	0	0	1	1	0	0	0	0	2	2	0	100
Conjugation (verb-form)	0	0	0	1	0	1	4	4	0	1	1	0	6	5	1	83.3
Preposition	2	0	2	3	0	3	12	9	3 (1)	7	6	1	24	15	9	62.5
Determiner	1	0	1	5	0	5	16	13	3	5	5	0	27	18	9	66.7
Vocabulary	1	0	1	5	0	5	27	13	14 (5)	3	1	2	36	14	22	38.9
Connector	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1	0
Spelling	0	0	0	2	2	0	1	1	0	0	0	0	3	3	0	100
Capitalization	7	0	7	0	0	0	5	1	4	8	0	8	20	1	19	5.0
Number	1	1	0	0	0	0	6	5	1	4	3	1	11	9	2	81.8
Tense	0	0	0	1	0	1	1	0	1	2	2	0	4	2	2	50.0
Word order	0	0	0	1	0	1	2	1	1	0	0	0	3	1	2	33.3
Possessive structure	0	0	0	1	0	1	6	3	3	2	1	1	9	4	5	44.4
Relative pronoun	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1	0
Collocation	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1	0
Comparison (comparative structures)	0	0	0	1	0	1	1	1	0	0	0	0	2	1	1	50.0
Punctuation	0	0	0	0	0	0	3	3	0	1	1	0	4	4	0	100
Unclear meaning	0	0	0	0	0	0	8	2	6	0	0	0	8	2	6	25.0
Extra word deletion	0	0	0	0	0	0	0	0	0	2	1	1	2	1	1	50.0
Negation error	0	0	0	0	0	0	1	1	0	0	0	0	1	1	0	100
Word form	0	0	0	0	0	0	1	1	0	1	1	0	2	2	0	100
Missing word	0	0	0	0	0	0	2	0	2	0	0	0	2	0	2	0

Abbreviations: De, detected; N, number; Unde, undetected.

that ChatGPT-4 may propose inappropriate "corrections" where there is no error, especially for preposition and vocabulary errors. The findings of this study support that ChatGPT-4 is not yet able to proofread a paper in the field of otolaryngology. Moreover, ChatGPT-4 claimed to have changed something that was already there and deleted certain sections/sentences from the original manuscript.

As for the clinical applications, several ethical issues arise about using ChatGPT-4 in paper-writing and correction, such as the risk of plagiarism and inaccuracies. Another ethical issue that requires further study is the potential imbalance in its accessibility between high- and low-income countries, if the software should cease to be free. An international consensus on how to regulate the use of artificial intelligence-powered language model of chatbots in scientific writing will soon be imperative.

It is important to keep in mind that large language models like GPT-4 are nondeterministic, which implies that their outputs may vary with each run. This inherent unpredictability is curtailed to some degree by fine-tuning specific hyperparameters that are available via the GPT-4 API. The hyperparameter tuning may influence the ChatGPT findings of task through the preventing overfitting, the improvement of speed of response, performance, or balancing resources. However, in the case of ChatGPT-4, these advanced adjustments may not be readily accessible. We partly addressed this issue through the test-retest reliability approach but the lack of knowledge about the hyperparameters of ChatGPT-4 may limit the understanding of the system's responses and suggested edits.

Moreover, the hyperparameter fine-tuning may have influenced the performance of ChatGPT-4 in the proofread task, for example, the percentage of ChatGPT-4 appropriate corrections, which makes the current results not necessarily transposable to other similar studies.

The main limitation of the present study is the focus on some parts of only 4 manuscripts in the field of laryngology or head and neck surgery. Depending on the field, ChatGPT-4's ability to proofread manuscripts may vary; the complexity of the paper should, theoretically, influence ChatGPT-4's performance.

Conclusion

ChatGPT demonstrated usefulness in identifying various types of errors but not all, which supports that non-native English researchers should be aware of the current limits of ChatGPT-4 in the proofreading of manuscripts. Future studies using future versions of ChatGPT-4 are needed to assess its ability to improve manuscripts by non-native English researchers in otolaryngology.

Author Contributions

Jerome R. Lechien, design, acquisition of data, 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; **Amy Gorton**, Data analysis & interpretation, and proofread of the paper, 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; **Jean Robertson**, Data analysis & interpretation, and proofread of the paper, 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; **Luigi A. Vaira**, design, acquisition of data, drafting, final approval, and accountability for the work; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Disclosures

Competing interests: None.

Funding source: None.

Supplemental Material

Additional supporting information is available in the online version of the article.

References

1. Pernencar C, Saboia I, Dias JC. How far can conversational agents contribute to IBD patient health care—a review of the literature. *Front Public Health*. 2022;10:862432. doi:10.3389/fpubh.2022.862432
2. Hill-Yardin EL, Hutchinson MR, Laycock R, Spencer SJ. A Chat(GPT) about the future of scientific publishing. *Brain Behav Immun*. 2023;110:152-154. doi:10.1016/j.bbi.2023.02.022
3. Choi JH, Hickman KE, Monahan A, Schwarcz D. ChatGPT-4 goes to law school? Minnesota Legal Studies Research Paper No. 23-03; 2023.
4. Gupta R, Park JB, Bisht C, et al. Expanding cosmetic plastic surgery research using ChatGPT-4. *Aesthet Surg J*. 2023;43(8):930-937. doi:10.1093/asj/sjad069
5. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care*. 2023;27(1):75. doi:10.1186/s13054-023-04380-2
6. Zumsteg JM, Junn C. Will ChatGPT-4 Match to Your Program? *Am J Phys Med Rehabil*. 2023;102(6):545-547. doi:10.1097/PHM.0000000000002238
7. Common European Framework of reference for languages: learning, teaching, assessment. Companion volume, Council of Europe; 2020.
8. Lechien JR, Seminerio I, Descamps G, et al. Impact of HPV infection on the immune system in oropharyngeal and non-oropharyngeal squamous cell carcinoma: a systematic review. *Cells*. 2019;8(9):1061. doi:10.3390/cells8091061
9. Lechien JR, Bobin F, Muls V, et al. Validity and reliability of the reflux symptom score. *Laryngoscope*. 2020;130(3):E98-E107. doi:10.1002/lary.28017
10. Lechien JR, Blecic S, Huet K, et al. Voice quality outcomes of idiopathic Parkinson's disease medical treatment: a systematic review. *Clin Otolaryngol*. 2018;43(3):882-903. doi:10.1111/coa.13082
11. Lechien JR, Khalife M, Huet K, et al. Impact of chemoradiation after supra- or infrahyoid cancer on aerodynamic, subjective, and objective voice assessments: a multicenter prospective study. *J Voice*. 2018;32(2):257. doi:10.1016/j.jvoice.2017.04.009
12. Ehara S, Takahashi K. Reasons for rejection of manuscripts submitted to AJR by international authors. *Am J Roentgenol*. 2007;188(2):W113-6.
13. Author Perspectives on Academic Publishing: Global Survey Report. Editage; 2018. Accessed March 29, 2023. https://campaign.editage.com/global_survey_report_2018/
14. King MR. The future of AI in medicine: a perspective from a Chatbot. *Ann Biomed Eng*. 2022;51:291-295.