

Original Paper

ChatGPT Versus Consultants: Blinded Evaluation on Answering Otorhinolaryngology Case–Based Questions

Christoph Raphael Buhr^{1,2}, MSc, MD, Dr med; Harry Smith³, MSc; Tilman Huppertz¹, MD, Dr med; Katharina Bahr-Hamm¹, MD, Dr med; Christoph Matthias¹, MD, Prof Dr Med; Andrew Blaikie², MD; Tom Kelsey³, PhD, Prof Dr; Sebastian Kuhn⁴, MME, MD, Prof Dr Med; Jonas Eckrich¹, MD, Dr med

¹Department of Otorhinolaryngology, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany

²School of Medicine, University of St Andrews, St Andrews, United Kingdom

³School of Computer Science, University of St Andrews, St Andrews, United Kingdom

⁴Institute of Digital Medicine, Philipps-University Marburg and University Hospital of Giessen and Marburg, Marburg, Germany

Corresponding Author:

Christoph Raphael Buhr, MSc, MD, Dr med

Department of Otorhinolaryngology

University Medical Center of the Johannes Gutenberg-University Mainz

Langenbeckstraße 1

Mainz, 55131

Germany

Phone: 49 6131 17 7361

Email: buhrchri@uni-mainz.de

Abstract

Background: Large language models (LLMs), such as ChatGPT (Open AI), are increasingly used in medicine and supplement standard search engines as information sources. This leads to more “consultations” of LLMs about personal medical symptoms.

Objective: This study aims to evaluate ChatGPT’s performance in answering clinical case–based questions in otorhinolaryngology (ORL) in comparison to ORL consultants’ answers.

Methods: We used 41 case-based questions from established ORL study books and past German state examinations for doctors. The questions were answered by both ORL consultants and ChatGPT 3. ORL consultants rated all responses, except their own, on medical adequacy, conciseness, coherence, and comprehensibility using a 6-point Likert scale. They also identified (in a blinded setting) if the answer was created by an ORL consultant or ChatGPT. Additionally, the character count was compared. Due to the rapidly evolving pace of technology, a comparison between responses generated by ChatGPT 3 and ChatGPT 4 was included to give an insight into the evolving potential of LLMs.

Results: Ratings in all categories were significantly higher for ORL consultants ($P<.001$). Although inferior to the scores of the ORL consultants, ChatGPT’s scores were relatively higher in semantic categories (conciseness, coherence, and comprehensibility) compared to medical adequacy. ORL consultants identified ChatGPT as the source correctly in 98.4% (121/123) of cases. ChatGPT’s answers had a significantly higher character count compared to ORL consultants ($P<.001$). Comparison between responses generated by ChatGPT 3 and ChatGPT 4 showed a slight improvement in medical accuracy as well as a better coherence of the answers provided. Contrarily, neither the conciseness ($P=.06$) nor the comprehensibility ($P=.08$) improved significantly despite the significant increase in the mean amount of characters by 52.5% ($n=(1470-964)/964$; $P<.001$).

Conclusions: While ChatGPT provided longer answers to medical problems, medical adequacy and conciseness were significantly lower compared to ORL consultants’ answers. LLMs have potential as augmentative tools for medical care, but their “consultation” for medical problems carries a high risk of misinformation as their high semantic quality may mask contextual deficits.

(*JMIR Med Educ* 2023;9:e49183) doi: [10.2196/49183](https://doi.org/10.2196/49183)

KEYWORDS

large language models; LLMs; LLM; artificial intelligence; AI; ChatGPT; otorhinolaryngology; ORL; digital health; chatbots; global health; low- and middle-income countries; telemedicine; telehealth; language model; chatbot

Introduction

The use of large language models (LLMs) is becoming increasingly common. Open access services such as Bard, Bing, and ChatGPT (Open AI) [1] have proven to be useful for a multitude of everyday applications [2,3]. Some experts argue that LLM services will soon augment, supplement, or replace today's search engines, and their application will become more common in specific areas across a broad range of established software applications [4]. The growing relevance and interest in this technology are also evident by the recent acquisitions of various artificial intelligence (AI)-specialized companies by major software corporations [5-9].

Launched in November 2022, ChatGPT has become one of the most popular LLMs. It uses a so-called "deep neural network architecture" to analyze and generate human-like language responses based on the input it receives. "Deep neural network architecture" refers to a specific type of machine learning model designed to recognize patterns and relationships in data using multiple (hidden) layers of interconnected nodes or "neurons." These nodes are organized into multiple layers, with each layer performing a specific computation on the input data and passing the results to the next layer. The term "deep" indicates the use of multiple hidden layers, allowing the detection of more complex patterns and relationships in the data compared to a "shallow" network with only 1 or 2 layers. The architecture is also classified as "neural" due to its interconnections and communication structure that are inspired by the interconnections of the human brain.

The architecture of ChatGPT is based on a transformer model, enabling it to process and understand sequences of text and generate natural language responses. A transformer model is a type of digital neural network architecture designed for natural language processing tasks, such as language translation, question answering, and text summarization. Introduced by Vaswani et al [10] in 2017, it has since become one of the most widely used architectures in natural language processing. The transformer model uses a self-attention mechanism, allowing it to capture long-range dependencies between words in a sentence without requiring sequential processing. This makes transformer models more efficient than traditional recurrent neural network architectures, which process input sequentially and are, therefore, slower and more computationally expensive. Moreover, self-attention appears to be a more interpretable class of models, linking the semantic and syntactic structure of inputs [10].

ChatGPT 3 has been trained on a vast and diverse corpus of text data, including a data set of web pages and internet content, and the BooksCorpus, a data set comprising over 11,000 books in various genres. During the training process, the model was trained to identify patterns in language, understand syntax and grammar, and generate coherent and meaningful responses to a wide variety of input prompts in different languages.

The digital age, particularly the advent of powerful search engines, has led to increasing accessibility of medical information for lay people. Thus, consulting "Dr Google" is

now a common means for patients to understand their symptoms and decide how to manage their medical issues [11-15]. Considering the extensive source database and the natural language of the answers provided, LLMs will likely become a relevant "go-to" tool for initial medical consulting in the future. However, using chatbots such as ChatGPT for medical consultation is not without risk [12,16]. While search engines and LLM-based chatbots typically warn users that their generated answers do not substitute for a consultation with a specialist, many patients may trust the information and make their own diagnostic or therapeutic conclusions. Consequently, misinterpretation of their symptoms may lead to incorrect conclusions, resulting in false illness convictions, increased anxiety, and potentially dangerous self-treatment or nontreatment [17,18].

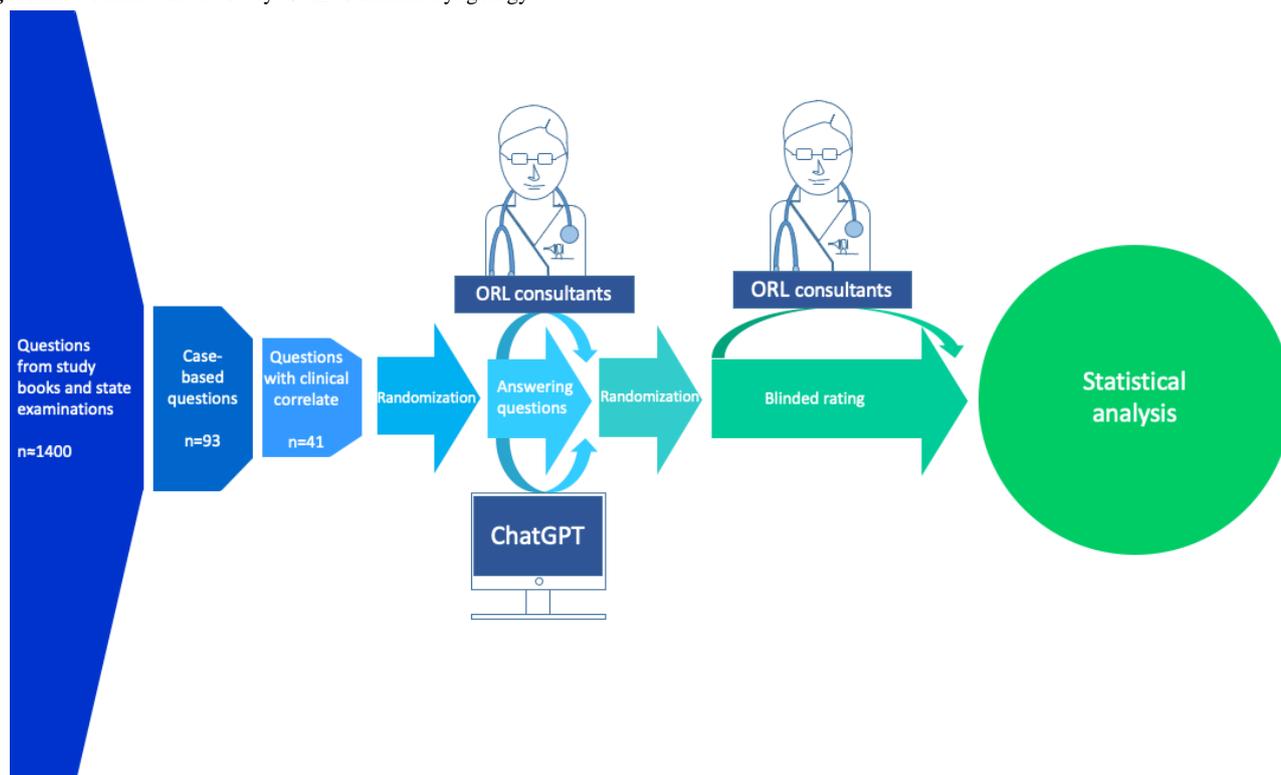
In addition, patients may harbor preconceptions that lead to conflict with their doctor, compromising the doctor-patient relationship [19,20]. The abilities and performance of LLMs, however, should not be underestimated. The fact that ChatGPT has been used as a tool to pass the United States Medical Licensing Examination (USMLE) demonstrates that LLMs can accurately answer medical questions [21,22]. Therefore, especially during times of specialist doctor shortages, long distances, and increased waiting periods, the availability of LLMs may further lower patients' threshold to consult an LLM-based chatbot such as ChatGPT rather than a trained professional. On the other hand, access to better medical information can also be considered beneficial for understanding specific symptoms, diagnoses, or treatments. However, unsupervised medical consultation of LLMs carries a high risk of misinformation without the guidance of an experienced specialist.

The ability of LLMs to pass a general medical examination has been proven, but the performance of LLM-derived answers to specific clinical case-based questions based on symptoms and clinical cases in otorhinolaryngology (ORL) has not yet been evaluated. Given these recent developments, this pilot study aims to assess the performance of ChatGPT when answering clinical case-based questions. ORL is one of the clinical disciplines with the highest consultation rate and encompasses a wide spectrum of conditions, ranging from relatively harmless to severe and potentially life-threatening diseases. Therefore, we analyzed the performance of ChatGPT in the field of ORL and compared it to the answers of ORL consultants.

Methods

Study Design

The workflow of this study is shown in [Figure 1](#). We studied established ORL textbooks and questions from previous German state examinations for doctors for case-based questions resembling realistic and authentic clinical scenarios [23,24]. Subsequently, clinical authenticity was verified by matching equivalent cases in the University Medical Center of Mainz. If cases did not have a homologous clinical correlation, they were exempt from the questionnaire. For an exemplary question, see Example S1 in [Multimedia Appendix 1](#).

Figure 1. Workflow of the study. ORL: otorhinolaryngology.

Answers to the 41 questions were recorded from 3 ORL consultants (coauthors of this paper) and the OpenAI chatbot ChatGPT 3 on March 11, 2023 [1]. Each ORL consultant (with at least 5 years of ORL-specific training) received the blinded answers of the other ORL consultants and those created by the ChatGPT LLM and was asked to rate them using a Likert scale (1=very poor and 6=excellent) for medical adequacy, conciseness, coherence, and comprehensibility. A 6-point Likert scale was chosen in order to avoid raters from taking the comfortable “neutral” position in the middle of the scale.

They then recorded whether they thought the answers were created by an ORL consultant or the ChatGPT LLM. After normality testing of the ratings using D’Agostino and Pearson test, the character count for every answer was recorded and compared using the Mann-Whitney *U* test with Prism for Windows (version 9.5.1; GraphPad Software).

Since this study aimed for maintaining a low barrier setup, simulating widespread availability, and considering a global health perspective, the experimental setup deliberately opted for the freely available versions of ChatGPT. However, challenged by the rapidly evolving pace of technology, a comparison between responses generated by ChatGPT 3 and ChatGPT 4 was included to give an insight into the evolving potential of LLMs. Ratings for answers provided by ChatGPT 3 and ChatGPT 4 were compared using the Mann-Whitney *U* test. As the amount of characters showed a Gaussian distribution, the 2-tailed *t* test was used.

Ethical Considerations

Written correspondence of March 3, 2023, with the ethics committee of the regional medical association

Rhineland-Palatinate determined that there is no need for any specific ethics approval due to the use of anonymous text-based questions.

Results

Cumulative results of ratings in every category were significantly higher for the answers given by the ORL consultants in comparison to the ChatGPT LLM ($P<.001$), with a similar range of ratings for medical adequacy and coherence and a broader range for conciseness and comprehensibility.

In detail, medical adequacy was rated with a median of 6 (IQR 5-6; range 1-6) for the ORL consultants compared to a 4 (IQR 4-5; range 1-6) for ChatGPT LLM ($P<.001$) when tested with the Mann-Whitney *U* test. Conciseness was rated with a median 6 (IQR 6-6; range 4-6) for the ORL consultants compared to a 4 (IQR 3-5; range 2-6) for ChatGPT LLM ($P<.001$) when tested with the Mann-Whitney *U* test. Furthermore, coherence was rated with a median of 6 (IQR 5-6; range 2-6) for the ORL consultants compared to a 5 (IQR 4-5; range 2-6) for ChatGPT LLM ($P<.001$) when tested with the Mann-Whitney *U* test, and comprehensibility was rated with a median of 6 (IQR 6-6; range 4-6) for the ORL consultants and 6 (IQR 5-6; range 2-6) for ChatGPT LLM ($P<.001$) when tested with the Mann-Whitney *U* test.

Comparative results of statistical testing are shown in [Table 1](#) and [Figure 2](#). Scores for all 3 ORL consultants were combined and compared to the ratings of answers by the ChatGPT LLM. Individual ratings of all ORL consultants in comparison to ratings for the answers provided by the ChatGPT LLM are shown in Figures S1-S5 in [Multimedia Appendix 1](#).

Table 1. Comparative results^a.

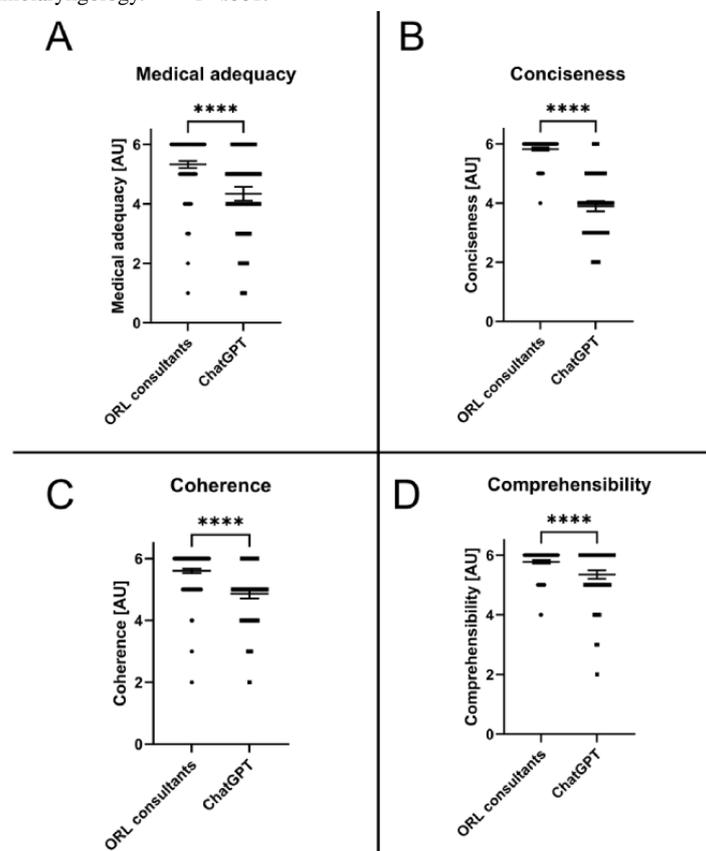
Result and source	Ratings, n	Values, mean (SD)	Rating, median (IQR)	Rating, 95% CI
Medical adequacy				
ORL ^b consultants	246	5.3 (0.9)	6 (5-6)	5-6
LLM ^c (ChatGPT)	123	4.3 (1.3)	4 (4-5)	4-5
Conciseness				
ORL consultants	246	5.8 (0.4)	6 (6-6)	6-6
LLM (ChatGPT)	123	3.9 (1.0)	4 (3-5)	4-4
Coherence				
ORL consultants	246	5.6 (0.6)	6 (5-6)	6-6
LLM (ChatGPT)	123	4.9 (0.8)	5 (4-5)	5-5
Comprehensibility				
ORL consultants	246	5.8 (0.5)	6 (6-6)	6-6
LLM (ChatGPT)	123	5.4 (0.8)	6 (5-6)	5-6

^a $P < .001$ when tested with the Mann-Whitney U test.

^bORL: otorhinolaryngology.

^cLLM: large language model.

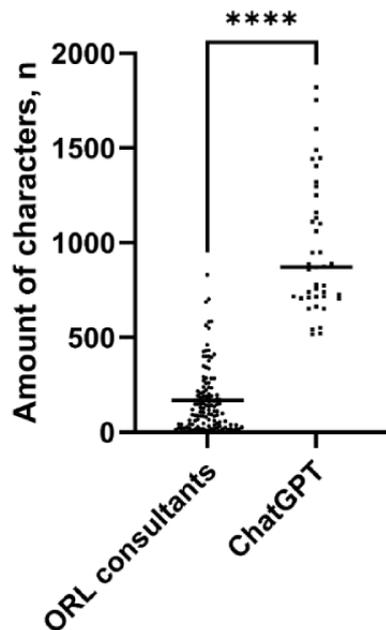
Figure 2. Comparison between ORL consultants and the LLM (ChatGPT) for all evaluated categories. Data shown as a scatter dot blot with each point resembling an absolute value (bar width resembling a high amount of individual values). Horizontal lines represent mean (95% CI). The nonparametric Mann-Whitney U test was used to compare the 2 groups. Cumulative results of ratings for (A) medical adequacy, (B) conciseness, (C) coherence, and (D) comprehensibility. ORL: otorhinolaryngology. **** $P < .001$.



The amount of characters of answers provided by the ORL consultants was significantly lower (median 119.0 (range 4-831; IQR 38.0-223.0) compared to (median 870.0 (IQR 712.5-1205.0) characters per answer for answers by ChatGPT LLM ($P < .001$).

when tested with the Mann-Whitney U test (Figure 3). For 98.4% (369/375) of the answers, the ORL consultants correctly identified the source of the answer.

Figure 3. The number of characters per answer used by ORL consultants and ChatGPT. Data shown as a scatter dot blot with each point resembling an absolute value. Horizontal lines represent the median. The nonparametric Mann-Whitney U test was used to compare the 2 groups. ORL: otorhinolaryngology. **** $P < .001$.



The supplemented comparison between responses generated by ChatGPT 3 and ChatGPT 4 showed a slight improvement in medical accuracy ($P = .03$). Additionally, ChatGPT 4 was rated with a better coherence of the answers provided ($P = .005$).

On the other hand, neither the conciseness ($P = .06$) nor the comprehensibility ($P = .08$) improved significantly (Figure 4), whereas the number of characters significantly increased by 52.5% ($n = (1470 - 964) / 964$; $P < .001$; Figure 5) when using the most recent version of ChatGPT.

Figure 4. Comparison between LLMs (ChatGPT 3 vs ChatGPT4) for all evaluated categories. Data shown as a scatter dot blot with each point resembling an absolute value (bar width resembling a high amount of individual values). Horizontal lines represent mean (95% CI). The nonparametric Mann-Whitney U test was used to compare the 2 groups. Cumulative results of ratings for (A) medical adequacy, (B) conciseness, (C) coherence, and (D) comprehensibility. ns: not significantly different. * $P < .05$; ** $P < .01$.

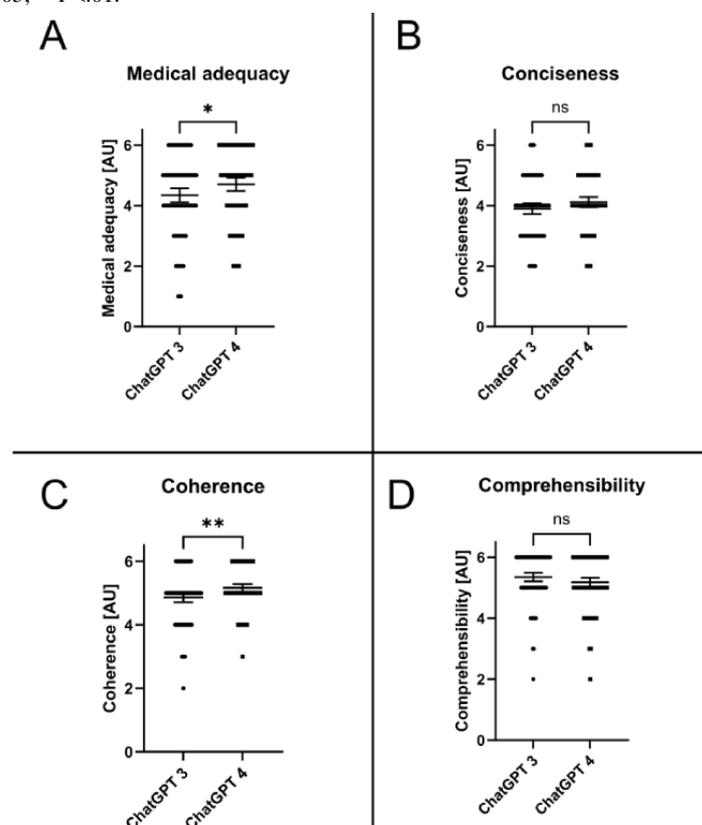
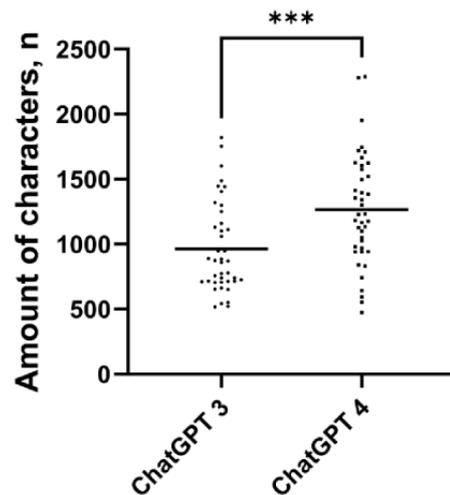


Figure 5. The number of characters used by ChatGPT 3 and ChatGPT 4. Data shown as a scatter dot blot with each point resembling an absolute value. Horizontal lines represent the mean. The Welch 2-tailed t test was used to compare the 2 groups. *** $P < .001$.



Discussion

Principal Findings

This pilot study assessed the performance of the ChatGPT LLM in answering clinical case-based questions in the field of ORL and compared it with certified ORL consultants. Overall, the ORL consultants significantly outperformed ChatGPT in medical adequacy, conciseness, comprehensibility, and coherence (Table 1 and Figure 2).

Comparison to Prior Work

Medical adequacy should be considered the most critical parameter, as even minor inaccuracies can lead to misdiagnosis or misinterpretation, resulting in increased anxiety, incorrect conclusions, and inadequate therapy or nontherapy [17,18]. Despite being an open access service without specific medical training, ChatGPT achieved relatively high ratings for medical adequacy. However, it still lagged behind the ORL consultants. ChatGPT's high-quality language output and coherent answers could potentially mislead users into believing they are receiving medically accurate information due to the halo effect [25]. This is concerning, especially since patients may struggle to interpret and apply the generated advice without physical examination, specialized tests, or clinical consultation. In this study, 10.6% (13/123) of ChatGPT's responses were rated "poor" or "very poor" in the category "medical adequacy" by at least 1 rater. Contrarily, only 1.2% (3/246) of answers by ORL consultants in "medical adequacy" were rated in the worst categories. This emphasizes the significance of a potential hazard caused by inadequate answers provided by ChatGPT. For instance, ChatGPT did conclude allergic symptoms in response to a case evolving around a potentially life-threatening cutaneous abscess, which was adequately recognized by all ORL consultants.

Moreover, ChatGPT's inability to recognize nonverbal cues or misunderstandings further highlights the limitations of LLMs in comparison to human physicians [26,27]. The high ratings for coherence and comprehensibility of ChatGPT's responses emphasize its semantic output quality but do not guarantee medical accuracy. In this study, ORL consultants easily distinguished between LLM-generated and human answers,

indicating that ChatGPT failed a simplified Turing test [28]. Although ORL consultants knew that 1 answer was generated by a machine, which represents a potential study bias, the high recognition rate is still relevant. The recognizability may be explained by an answering style consisting of long answers and a wording and semantic structure characteristic for ChatGPT (see Example S1 in Multimedia Appendix 1). Nevertheless, laypeople might be more susceptible to ChatGPT's eloquence.

Despite ChatGPT's inferior performance in all evaluated categories, the potential for future improvements cannot be ignored. LLM-based chatbots such as ChatGPT could revolutionize clinical care by increasing the availability of medical information, especially in low-resource settings. As new and improved LLMs are developed, their medical accuracy may improve, making them valuable augmentative tools for medical professionals. This could lead to more precise, time-efficient, and individualized medicine.

Strengths and Limitations

However, the current limitations of LLMs, such as data protection and legal issues, must be addressed before they can be integrated into clinical practice [29-31]. This study design has certain limitations. First, the use of case-based questions does not properly reflect the style or quality of laypeople questions. Furthermore, considering the accuracy of identification of the source of the answers provided may influence the rating and limit the characterization as a "single blinded study." Although the questions were specifically selected in concordance with equivalent cases in the ORL department, using text-based questions is an obvious limitation of this study design.

Furthermore, the evolution in the field of LLMs is progressing rapidly. Thus, all scientific data obtained in this field will ultimately only be able to depict a specific time point of LLMs evolving potential.

Future Directions

While early LLMs were trained on small data sets of text and code, they often generated rather inaccurate answers. Yet, a significant increase in the size and complexity of LLM data sets

resulted in improvements in the accuracy and reliability of LLM-generated medical answers [32]. To allow an insight into the current state of the art of LLMs, a comparison of findings obtained with the latest (fee-based) version of ChatGPT (ChatGPT 4) to ChatGPT 3 was added during the review process. As shown in Figure 4, the latest version showed only a slight improvement for medical adequacy ($P=.03$). Yet, ChatGPT 4 was rated with a better coherence of the answers provided when compared to ChatGPT 3 ($P=.005$). In contrast, the conciseness or comprehensibility did not improve significantly although the amount of characters increased by a highly significant 52.5% ($n=(1470-964)/964$; Figure 5). These findings are in concordance with recent published data highlighting that LLMs often generate inconcise answers due to the sheer amount of information provided [33]. Through years of training, clinicians also possess a large knowledge base concerning their specific field. Ultimately, soft skills such as the identification of nonverbal communication as personal

experience cannot be replicated by an LLM. In clinical practice, therapeutic decisions are rarely based on the anamnesis alone. Instead, clinicians can gather a multidimensional view of the patients' pathology combining findings of the anamnesis, examinations, and personal impressions. These are advantages an LLM currently simply cannot match. We believe that the rapid evolution of LLMs will soon provide better and more specialized advice for medical problems, making them more relevant as an augmentative option especially in areas with insufficient availability of medical care. Nevertheless, we are convinced that human consultation will remain the undisputed gold standard in medical care in the near future.

Therefore, this pilot study serves as a starting point for evaluating the performance of LLMs in the field of ORL. Further research should investigate the potential of LLMs on a larger scale and for different audiences, focusing on the development of specialized LLMs that could assist health care professionals without replacing their expertise.

Acknowledgments

Figure 1 was drawn by JE and CRB using Microsoft PowerPoint (Microsoft Corp). Figures 2-5 were assembled by JE and CRB using Prism for Windows (version 9.5.1; GraphPad Software). No third-party funding was used for the design of the study; the collection, analysis and interpretation of data; and in writing the paper.

Authors' Contributions

CRB made substantial contributions to the conception of the work. He drafted the paper and approved the submitted version. As the corresponding author, he claims responsible for ensuring that all listed authors have approved the paper before submission. He declares that all scientific writers and anyone else who assisted with the preparation of the paper content are properly acknowledged, along with their source of funding. He further states that no authors on earlier versions have been removed and no new authors were added. CRB, HS, and TH were involved in acquisition, analysis, and interpretation of data. HS and TH further participated in drafting and revising of the paper and approved the submitted version. KB-H was involved in analysis and interpretation of data. CM was substantially involved in acquisition and analysis of data. He further participated in revising the *Methods* section of the paper and approved the submitted version. AB was involved in acquisition and interpretation of data. TK and SK made substantial contributions to the design of the work. They further were involved in the interpretation of data. JE made substantial contributions to the conception and design of the work. He substantially participated in the acquisition, analysis, and interpretation of data and the drafting process of this paper. KB-H, AB, TK, and SK revised the paper and approved the submitted version. All authors agreed both to be personally accountable for the author's own contributions and for the accuracy and integrity of any part of the work.

Conflicts of Interest

SK is the founder and shareholder of MED.digital.

Multimedia Appendix 1

Example question and individual consultant comparisons.

[DOCX File, 536 KB-Multimedia Appendix 1]

References

1. ChatGPT. OpenAI. 2021. URL: <https://openai.com/chatgpt> [accessed 2023-11-17]
2. Surameery NMS, Shakor MY. Use Chat GPT to solve programming bugs. IJITC. 2023;3(1):17-22 [FREE Full text] [doi: 10.55529/ijitc.31.17.22]
3. Zielinski C, Winker MA, Aggarwal R, Ferris LE, Heinemann M, Lapeña JFJ, et al. Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. WAME. WAME; 2023. URL: <https://wame.org/page3.php?id=106> [accessed 2023-11-17]
4. Grant N, Metz C. A new chat bot is a 'code red' for Google's search business. The New York Times. 2023. URL: <https://www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html> [accessed 2023-11-17]

5. Google buys UK artificial intelligence start-up DeepMind. BBC. 2014. URL: <https://www.bbc.com/news/technology-25908379> [accessed 2023-11-17]
6. Bass D, Baker L. Microsoft to acquire Nuance for \$19.6 billion in health-care bet. Bloomberg. 2021. URL: <https://www.bloomberg.com/news/articles/2021-04-12/microsoft-buys-nuance-for-19-7-billion-in-bet-on-health-care#xj4y7vzkg> [accessed 2023-11-17]
7. Boyle A, Soper T, Bishop T. Exclusive: Apple acquires Xnor.ai, edge AI spin-out from Paul Allen's AI2, for price in \$200M range. GeekWire. 2020. URL: <https://www.geekwire.com/2020/exclusive-apple-acquires-xnor-ai-edge-ai-spin-paul-allens-ai2-price-200m-range/> [accessed 2023-11-17]
8. Cuthbertson A. Facebook buys mind-reading startup CTRL-Labs for \$1bn. Independent. 2019. URL: <https://www.independent.co.uk/tech/facebook-mind-reading-ctrl-labs-wristband-neural-interface-a9117801.html> [accessed 2023-11-17]
9. Weise K, Griffith E. Amazon to buy Zoox, in a move toward self-driving cars. The New York Times. 2020. URL: <https://www.nytimes.com/2020/06/26/business/amazon-zoox.html#:~:text=SEATTLE%20%E2%80%94%20Amazon%20said%20on%20Friday,person%20familiar%20with%20the%20deal> [accessed 2023-11-17]
10. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017; Long Beach, CA, USA
11. Tang H, Ng JHK. Googling for a diagnosis—use of Google as a diagnostic aid: internet based study. *BMJ*. 2006;333(7579):1143-1145 [FREE Full text] [doi: [10.1136/bmj.39003.640567.AE](https://doi.org/10.1136/bmj.39003.640567.AE)] [Medline: [17098763](https://pubmed.ncbi.nlm.nih.gov/17098763/)]
12. Jungmann SM, Brand S, Kolb J, Withhöft M. Do Dr. Google and health apps have (comparable) side effects? an experimental study. *Clin Psychol Sci*. 2020;8(2):306-317 [doi: [10.1177/2167702619894904](https://doi.org/10.1177/2167702619894904)]
13. Lee K, Hoti K, Hughes JD, Emmerton L. Dr Google and the consumer: a qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. *J Med Internet Res*. 2014;16(12):e262 [FREE Full text] [doi: [10.2196/jmir.3706](https://doi.org/10.2196/jmir.3706)] [Medline: [25470306](https://pubmed.ncbi.nlm.nih.gov/25470306/)]
14. Cocco AM, Zordan R, Taylor DM, Weiland TJ, Dilley SJ, Kant J, et al. Dr Google in the ED: searching for online health information by adult emergency department patients. *Med J Aust*. 2018;209(8):342-347 [doi: [10.5694/mja17.00889](https://doi.org/10.5694/mja17.00889)] [Medline: [30107763](https://pubmed.ncbi.nlm.nih.gov/30107763/)]
15. Jacobs W, Amuta AO, Jeon KC. Health information seeking in the digital age: an analysis of health information seeking behavior among US adults. *Cogent Soc Sci*. 2017;3(1):1302785 [FREE Full text] [doi: [10.1080/23311886.2017.1302785](https://doi.org/10.1080/23311886.2017.1302785)]
16. Wangler J, Jansky M. Internet-associated health anxieties in primary care—results of a survey among general practitioners and primary care internists in Hesse. *Dtsch Med Wochenschr*. 2019;144(16):e102-e108 [FREE Full text] [doi: [10.1055/a-0842-8285](https://doi.org/10.1055/a-0842-8285)] [Medline: [30822804](https://pubmed.ncbi.nlm.nih.gov/30822804/)]
17. Norr AM, Capron DW, Schmidt NB. Medical information seeking: impact on risk for anxiety psychopathology. *J Behav Ther Exp Psychiatry*. 2014;45(3):402-407 [doi: [10.1016/j.jbtep.2014.04.003](https://doi.org/10.1016/j.jbtep.2014.04.003)] [Medline: [24818986](https://pubmed.ncbi.nlm.nih.gov/24818986/)]
18. McMullan RD, Berle D, Arnáez S, Starcevic V. The relationships between health anxiety, online health information seeking, and cyberchondria: systematic review and meta-analysis. *J Affect Disord*. 2019;245:270-278 [doi: [10.1016/j.jad.2018.11.037](https://doi.org/10.1016/j.jad.2018.11.037)] [Medline: [30419526](https://pubmed.ncbi.nlm.nih.gov/30419526/)]
19. Caiata-Zufferey M, Abraham A, Sommerhalder K, Schulz PJ. Online health information seeking in the context of the medical consultation in Switzerland. *Qual Health Res*. 2010;20(8):1050-1061 [doi: [10.1177/1049732310368404](https://doi.org/10.1177/1049732310368404)] [Medline: [20442347](https://pubmed.ncbi.nlm.nih.gov/20442347/)]
20. Fox S. After Dr Google: peer-to-peer health care. *Pediatrics*. 2013;131(Suppl 4):S224-S225 [doi: [10.1542/peds.2012-3786K](https://doi.org/10.1542/peds.2012-3786K)] [Medline: [23729765](https://pubmed.ncbi.nlm.nih.gov/23729765/)]
21. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198 [FREE Full text] [doi: [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198)] [Medline: [36812645](https://pubmed.ncbi.nlm.nih.gov/36812645/)]
22. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? the implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312 [FREE Full text] [doi: [10.2196/45312](https://doi.org/10.2196/45312)] [Medline: [36753318](https://pubmed.ncbi.nlm.nih.gov/36753318/)]
23. Reineke U, Riemann R. *Facharztprüfung Hals-Nasen-Ohrenheilkunde: 1000 kommentierte Prüfungsfragen*. Stuttgart. Thieme; 2007.
24. Lenarz T, Boenninghaus HG. *Hals-Nasen-Ohren-Heilkunde*. Berlin, Heidelberg. Springer-Verlag; 2012.
25. Thorndike EL. A constant error in psychological ratings. *J Appl Psychol*. 1920;4(1):25-29 [doi: [10.1037/h0071663](https://doi.org/10.1037/h0071663)]
26. Mast MS. On the importance of nonverbal communication in the physician-patient interaction. *Patient Educ Couns*. 2007;67(3):315-318 [doi: [10.1016/j.pec.2007.03.005](https://doi.org/10.1016/j.pec.2007.03.005)] [Medline: [17478072](https://pubmed.ncbi.nlm.nih.gov/17478072/)]
27. Marcinowicz L, Konstantynowicz J, Godlewski C. Patients' perceptions of GP non-verbal communication: a qualitative study. *Br J Gen Pract*. 2010;60(571):83-87 [FREE Full text] [doi: [10.3399/bjgp10X483111](https://doi.org/10.3399/bjgp10X483111)] [Medline: [20132701](https://pubmed.ncbi.nlm.nih.gov/20132701/)]
28. Turing AM. I.—Computing machinery and intelligence. *Mind*. 1950;LIX(236):433-460 [FREE Full text] [doi: [10.1093/mind/LIX.236.433](https://doi.org/10.1093/mind/LIX.236.433)]

29. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;6(1):120 [FREE Full text] [doi: [10.1038/s41746-023-00873-0](https://doi.org/10.1038/s41746-023-00873-0)] [Medline: [37414860](https://pubmed.ncbi.nlm.nih.gov/37414860/)]
30. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. *JAMA*. 2023;330(4):315-316 [FREE Full text] [doi: [10.1001/jama.2023.9651](https://doi.org/10.1001/jama.2023.9651)] [Medline: [37410482](https://pubmed.ncbi.nlm.nih.gov/37410482/)]
31. Yaeger KA, Martini M, Yaniv G, Oermann EK, Costa AB. United States regulatory approval of medical devices and software applications enhanced by artificial intelligence. *Health Policy Technol*. 2019;8(2):192-197 [FREE Full text] [doi: [10.1016/j.hlpt.2019.05.006](https://doi.org/10.1016/j.hlpt.2019.05.006)]
32. de Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health*. 2023;11:1166120 [FREE Full text] [doi: [10.3389/fpubh.2023.1166120](https://doi.org/10.3389/fpubh.2023.1166120)] [Medline: [37181697](https://pubmed.ncbi.nlm.nih.gov/37181697/)]
33. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv. Preprint posted online on May 16, 2023. [FREE Full text] [doi: [10.48550/arXiv.2305.09617](https://doi.org/10.48550/arXiv.2305.09617)]

Abbreviations

- AI:** artificial intelligence
LMM: large language model
ORL: otorhinolaryngology
USMLE: United States Medical Licensing Examination

Edited by MN Kamel Boulos, K Venkatesh; submitted 20.05.23; peer-reviewed by C Zielinski, P Costa, N Domingues; comments to author 12.07.23; revised version received 20.07.23; accepted 20.10.23; published 05.12.23

Please cite as:

*Buhr CR, Smith H, Huppertz T, Bahr-Hamm K, Matthias C, Blaikie A, Kelsey T, Kuhn S, Eckrich J
ChatGPT Versus Consultants: Blinded Evaluation on Answering Otorhinolaryngology Case-Based Questions*

JMIR Med Educ 2023;9:e49183

URL: <https://mededu.jmir.org/2023/1/e49183>

doi: [10.2196/49183](https://doi.org/10.2196/49183)

PMID:

©Christoph Raphael Buhr, Harry Smith, Tilman Huppertz, Katharina Bahr-Hamm, Christoph Matthias, Andrew Blaikie, Tom Kelsey, Sebastian Kuhn, Jonas Eckrich. Originally published in *JMIR Medical Education* (<https://mededu.jmir.org>), 05.12.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Medical Education*, is properly cited. The complete bibliographic information, a link to the original publication on <https://mededu.jmir.org/>, as well as this copyright and license information must be included.