Published in final edited form as:
Eur Arch Otorhinolaryngol. 2024 June ; 281(6): 3219–3225. doi:10.1007/s00405-024-08524-0.

ChatGPT vs. Web Search for Patient Questions: What Does ChatGPT Do Better?

Sarek A Shen, MD, MS[1], **Carlos A Perez-Heydrich, BS**[2], **Deborah X. Xie, MD**[1], **Jason Nellis, MD**[1]

[1]Johns Hopkins School of Medicine, Department of Otolaryngology-Head and Neck Surgery. Baltimore, Maryland. USA

[2]Johns Hopkins School of Medicine. Baltimore, Maryland. USA

Abstract

Purpose—Chat Generative Pretrained Transformer (ChatGPT) has the potential to significantly impact how patients acquire medical information online. Here, we characterize the readability and appropriateness of ChatGPT responses to a range of patient questions compared to results from traditional web searches.

Methods—Patient questions related to the published Clinical Practice Guidelines by the American Academy of Otolaryngology-Head and Neck Surgery were sourced from existing online posts. Questions were categorized using a modified Rothwell classification system into 1)Fact, 2)Policy, and 3)Diagnosis and Recommendations. These were queried using ChatGPT and traditional web search. All results were evaluated on readability (Flesch Reading Ease and Flesch-Kinkaid Grade Level) and understandability (Patient Education Materials Assessment Tool). Accuracy was assessed by two blinded clinical evaluators using a three-point ordinal scale.

Results—54 questions were organized into Fact (37.0%), Policy (37.0%), and Diagnosis (25.8%). The average readability for ChatGPT responses was lower than traditional web search (FRE: 42.3±13.1 vs. 55.6±10.5, p<0.001), while the PEMAT understandability was equivalent (93.8% vs. 93.5%, p=0.17). ChatGPT scored higher than web search for questions the 'Diagnosis' category (p<0.01); there was no difference in questions categorized as 'Fact' (p=0.15) or 'Policy' (p=0.22). Additional prompting improved ChatGPT response readability (FRE 55.6±13.6, p<0.01).

Conclusion—ChatGPT outperforms web search in answering patient questions related to symptom-based diagnoses and is equivalent in providing medical facts and established policy. Appropriate prompting can further improve readability while maintaining accuracy. Further patient

Corresponding Author: Sarek Shen, 601 North Caroline Street, Baltimore, MD 21287, USA, sarek.shen@gmail.com.
Author contribution: Dr. Sarek Shen led study design, analysis and interpretation of the data, and composing the manuscript. Dr. Xie assisted with design and evaluation of ChatGPT and web search responses. Mr. Perez-Heydrich provided literature review and quantification of response readability and understandability. Dr. Nellis helped conceive the project and reviewed the manuscript.

Conflicts of interest: None

Ethics approval: This study does not include the use of human or animal subjects and was deemed exempt by the Johns Hopkins Institutional Review Board

Data: Questions used within this project are included in the supplementary data

education is needed to relay the benefits and limitations of this technology as a source of medial information.

Keywords

Large language model; ChatGPT; patient education; patient questions; accuracy; readability; accessibility

Introduction

The availability of health information online has expanded exponentially in the last decade. Patients have increasingly turned to the internet to answer health-related questions and facilitate decision-making processes. Surveys have demonstrated that between 42% and 71% of adult internet users have searched for medical information online and include topics ranging from pharmacological side-effects to disease pathology[1–3]. However, online resources obtained via web searches demonstrate significant variation in quality and understandability. The variability can lead to patient confusion, delays in care, and miscommunication with providers[4].

The release of a publicly available large language model (LLM), ChatGPT-3.5 (Chat Generated Pre-Trained Transformer), has sparked significant discussion within the healthcare sector. This chat-based interface, also referred to as conversational artificial intelligence (AI), responds to a range of natural language queries in a conversational and intuitive fashion. The tool has demonstrated a range of capabilities, including passing the USMLE Step 1 and creating high quality, fictious medical abstracts[5, 6]. The model has also shown the capability to generate patient recommendations for cardiovascular disease prevention[7], as well as post-operative instructions[8]. It can provide empathetic responses to patient questions [9] and answer queries within a range of surgical subspecialties[10, 11]. Given the robust nature of its input parameters and conventional responses, ChatGPT has the potential to be a valuable tool for both patients and providers.

With the growing ubiquity of these LLM, including ChatGPT-3.5, it is likely that some patients may turn to this technology to answer questions that were previously directed to traditional web searches. There have been numerous investigations within otolaryngology on the quality and understandability of online patient education materials. These studies have largely found that internet resources tend vary significantly in reliability and are often written at grade levels above the average reading level [12–14], which fail to meet the standard of 6th grade reading level as recommended by the American Medical Association (AMA), National Institutes of Health (NIH), and Agency of Healthcare Research and Quality (AHRQ)[15, 16]. Given the adaptable input criteria, a LLM has the potential to synthesize personalized responses appropriate for patients. The purpose of this study was to analyze the readability, understandability, and accuracy of ChatGPT-3.5 responses to a spectrum of user-generated patient queries and compare them to results from traditional web searches.

Methods

Data sources

This study was deemed exempt by the Johns Hopkins Institutional Review Board. The data for this study was collected in July 2023. Utilizing the 18 Clinical Practice Guidelines (CPG) published by the American Academy of Otolaryngology-Head and Neck Surgery (2013–2022), our group amassed 54 total questions, three for each CPG topic, encompassing common post-operative queries, symptomatic concerns, pharmacologic options, differential diagnoses. These questions were drawn from existing social media posts (Reddit.com/r/AskDocs, Yahoo! Answers, Facebook) as well as commonly asked questions included within medical institution websites. The questions were categorized into three groups using a modified Rothwell criteria [17] into 1) Fact: asks for objective and factual information (i.e: How is Meniere's disease diagnosed?) 2) Policy: asks about a specific course of action, including preventative measures, for known diagnoses or scenarios (i.e: What can I eat after my tonsillectomy?) and 3) Diagnosis and Recommendations: asks for recommendations or diagnoses given symptoms (i.e: I have a lump in my neck, what could it be and what should I do?). The list of questions is included in Supplemental Table 1. Each question was input into the ChatGPT-3.5 interface twice and results were recorded. The questions were also entered into Google search using the Google Chrome browser in an incognito window with the history cleared. The results from the first two links were collected. Scientific articles and restricted websites were omitted from the search, as they are not representative of commonly accessed health material. Figures, tables, and image captions were not included in our assessment. To further investigate the effect of additional prompting in ChatGPT readability, the phrase 'Please answer at a 6th grade level' was included at the end of each question.

Outcome measures

Content readability was assessed using both the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL). These tools evaluate text for readability using a formula that incorporates average sentence length and average syllable per sentence. FRE scores are given between 0–100, with scores above 80 indicating that the text is the level of conversational English. FKGL scores give the approximate US grade level education needed to understand the text.

The understandability of the language model and search results were measured using the Patient Education Materials Assessment Tool (PEMAT). This is a validated instrument designed to assess educational materials that are appropriate for all patients[16]. As described by the Agency for Healthcare Research and Quality, understandability refers to the ease at which the reader can process and explain key messages. Given the nature of the generated queries, the other component of the PEMAT, 'actionability', was not consistently applicable and therefore excluded from our analysis.

The accuracy and completeness of the responses was each graded by an blinded, independent clinical reviewer (SAS, DXX) based on the recommendations given in the clinical practice guidelines published by the American Academy of Otolaryngology-Head and Neck surgery. The scoring was completed using an ordinal three-point scale [18]. A score of 3 was given for that the response was accurate, relevant, and comprehensive, 2 for inaccuracies or missing information, and 1 for major errors or irrelevance.

Statistical analysis

Hypothesis testing was performed comparing readability, accuracy, and accuracy between ChatGPT & traditional web search. Results were analyzed using descriptive statistics. Reliability of the ChatGPT and web search output were assessed using paired student t-tests. Student-t testing was used to evaluate the difference in the two groups in readability, understandability, and accuracy. For response accuracy, inter-observer reliability was assessed using intraclass correlation. Statistical analysis was performed on R Studio version 2022.12.0 (Vienna, Austria) and a significance level of $p<0.05$ was use for all analyses.

Results

Fifty-four questions were included in this study. There were 20 questions (37.0%) in Category 1: Fact, 20 (37.0%) in Category 2: Policy, and 14 (25.9%) in Category 3: Diagnosis and Recommendations. Four responses were obtained for each question, two from ChatGPT and two from traditional web search. Paired t-testing between the two responses for each modality was not significant for any of the assessments, indicating that the readability and understandability remained consistent between repeat queries for both ChatGPT and traditional web searches (Supplemental Table 2). The FRE reading levels for the average ChatGPT response were significantly lower than that of the average web searches (42.3±14.2 vs. 56.2±17.4, $p<0.01$), indicating a higher level of difficulty. The average grade level (FKRL) needed to understand the ChatGPT answers was higher than that of web searches (12.1 ± 2.8 vs. 9.4 ± 3.3, $p<0.01$). Overall, both ChatGPT and web search responses were highly understandable based on PEMAT (ChatGPT: 93.8% (57.1%–100.0%), web search: 88.4% (42.9%–100.0%)). These data are summarized in Table 1.

Two blinded, independent reviewers determined the accuracy of each response on an ordinal scale from 1–3. The mean ChatGPT score was 2.87 ± 0.34, significantly higher than the score of the web search response (2.61 ± 0.63, mean difference: 0.26, 95% CI: 0.16–0.36). Interrater reliability was high for both ChatGPT (Cohen’s Kappa: 0.82, 95% CI: 0.72–0.88) and web search (0.79, 95%CI: 0.70–0.87). On subgroup analysis, the accuracy of the language model and web searches were equivalent in Fact (2.93 2.93 ± 0.22 vs.2.72 ± 0.54, $p = 0.15$) and Policy (2.69 ± 0.43 vs. 2.50 ± 0.51, $p = 0.21$) categories. However, ChatGPT had a statistically higher score in response for questions organized into Diagnosis and Recommendations (2.92 ± 0.25 vs. 2.55 ± 0.43, $p = 0.02$) (Figure 1).

The 54 questions were posed again to ChatGPT with explicit instructions for the response to be generated at a 6th grade reading level. The mean FRE increased to 55.6 ± 13.4, and the mean FKRL decreased to 9.3 ± 2.67, both indicating increased readability. A one-way ANOVA was conducted to test for differences in readability between these three groups: ChatGPT, ChatGPT – 6th grade, and Web Search. On Tukey multiple pairwise comparison, there was no difference in readability between ChatGPT-6th Grade and standard web searches, and both were significantly easier to read than ChatGPT without prompting. The addition of the reading level prompt did not result in a change in accuracy scores.

(ChatGPT: 2.87 ± 0.34; ChatGPT 6th Gr: 2.81 ± 0.36, p = 0.43). These data are shown in Figure 2.

Discussion

The emergence of publicly available large language artificial intelligence has provoked significant discussion within the healthcare sphere. ChatGPT has the potential to improve patient engagement, broaden access to medical information, and minimize the cost of care. In this study, we analyzed the responses of this popular language model to a range of input that encompasses common patient concerns. Our study showed that this language model was able to provide consistent and readable responses to a range of patient questions as compared to traditional web search. Interestingly, we found that ChatGPT did a better job with queries asking for possible diagnoses and recommendations based on given symptoms, while providing equivalent responses to questions related to disease information or post-operative policies.

A significant concern with utilizing chat based AI in patient care is verifying the validity of its output. Despite its convincing text responses, there is little data in the field of otolaryngology on the accuracy and applicability of ChatGPT's results. Using the AAO-HNS Clinical Practice Guidelines as reference, our group found that the accuracy for the language model was equivalent to that of traditional web searches for certain question types. Notably, ChatGPT outperformed traditional web searches for queries asking for possible diagnoses recommendations based on symptoms (i.e. 'My face isn't moving, what could it be and what should I do?'). However, there were no differences in responses to questions regarding medical facts, such as disease definitions or diagnostic criteria (i.e. What is obstructed sleep disordered breathing), or policy related to established diagnoses (i.e. How much oxycodone should I take after my rhinoplasty?). In a recent study, Ayoub et. al similarly found that ChatGPT performed equivalently to Google Search in questions related to patient education[19]. However, they noted that the platform did worse when providing medical recommendations, which is partially discordant with our findings. These discrepancies may be explained in part by the differences in question sources; our study included questions taken verbatim from social media sources, which may include input errors in grammar or syntax, or vague medical terminology. The advanced language processing utilized by ChatGPT allows for better identification of user intent and relevant information which can improve flexibility of input for the LLM. This generalizability was also found by Gilson et al. in their analysis of ChatGPT's performance in answering medical questions [6]. Combined with the dialogic nature of its output, the model could represent an alternative for patients seeking medical information online.

Prior studies evaluating the most accessed online resources for patient information have shown that there is variable readability and accessibility [20–22]. We found that the average readability of search engine results to be at the ninth-grade reading level. ChatGPT responses were presented at an even higher reading level, with 56% of the responses at college-level or above. Unsurprisingly, the ChatGPT generated responses that cited scientific articles and clinical practice guidelines tended to require a higher reading level than those based on patient-directed resources. This occurs more frequently when questions included more technical terms, such as ‘acute bacterial rhinosinusitis’. However, when specific instructions were given to the model to answer questions at a sixth-grade reading level, we found that ChatGPT was able to provide responses closer to the current web search standard [22, 23] [24]. This functionality allows ChatGPT to provide answers at a wide range of education levels, which may have implications in increasing accessibility to medical information and reducing health care disparities[25].

For patients, these large language models represent an avenue for accessible, focused, and understandable education. In our investigation, we noted that ChatGPT was able to find appropriate answers to otolaryngology questions even if they lacked certain descriptors (i.e. ‘fluid’ instead of ‘ear fluid’), demonstrating adaptable input criteria not typically seen in traditional web searches. ChatGPT also does well answering queries with keywords that may be present in other medical fields; it correctly responded to ‘Do I need imaging for my allergies’, while the web search results listed links to contrast allergies. Similar advantages in other AI conversational agents have previously been reported[26, 27]; however, ChatGPT represents a significant advancement over prior iterations. Additionally, the LLM can tailor responses to subsequent questions based on prior queries, which may be more helpful to patients than the FAQ or bullet-point style formatting of current online resources.

Given these exploratory findings, it is evident that conversational AI has the potential to play a large role in the healthcare field; understanding the benefits and limitations of this technology is paramount to educating patients in the appropriate medical use of the platform. Instructing patients how to optimize search criteria, interpret ChatGPT responses, and ask follow-up questions is necessary to fully and safely utilize these LLMs. This has become even more important as traditional search engines have begun incorporating artificial intelligence in their search tools, such as Google Bard and Microsoft Copilot. Additionally, providers must also be aware of possible demographic bias arising from unsupervised training data, potential complications in medico-legal matters, and compromise of patient privacy due to AI-associated transparency requirements [28, 29]. As new iterations of these LLM continue to evolve, providers must endeavor to keep abreast of the potential hazards and restrictions of these technologies.

There are several limitations to this study. First, the questions that our group generated do not fully capture the range of possible queries that patients may have. We limited our study to topics with published guidelines by the AAO-HNS, which only represents a small fraction of the field of otolaryngology and medicine as a whole. Second, the three-point scale utilized by our team to assess the accuracy and completeness of the responses may not provide the ideal resolution. Accuracy, particularly within medicine, is highly dependent on clinical context; follow-up questions that would help clarify certain nuances are not routinely asked by the LLM. Third, results from only the first two links on Google were recorded, which does not fully approximate the overall information available via web search. Although including additional links may further improve the readability and accuracy of this approach, any discordance between results may introduce unwanted confusion, and further highlights the utility of ChatGPT as a central repository of information.

From a technological standpoint, there are notable caveats for utilizing this platform. As a language model, ChatGPT is inherently built to create plausible sounding, human-like responses, some of which may not be factually correct[30]. Many of its responses in our study drew from reliable sources, such as the Mayo Clinic, which underlies the high level of accuracy that we found. However, certain queries may result in 'hallucinations', a term describing AI generated responses that sound plausible but aren't so. Identification of these replies by trained providers is crucial to patient safety. Like all machine learning platforms, ChatGPT is susceptible to biases and limitations of training data and may omit recent developments outside of the training timeline[31]. Finally, the current language model is constrained to text responses. Figures and diagrams are essential to patient education, particularly in a surgical field, which unfortunately are not included in this iteration of the model.

Conclusion

ChatGPT can provide text responses to a range of patient questions with high readability and accuracy. The platform outperforms traditional web search in answering patient questions related to symptom-based diagnoses and is equivalent in providing medical information. Appropriate prompting within ChatGPT can tailor its responses to a range of reading levels. It is evident that similar artificial intelligence systems have the potential to improve health information accessibility. However, the potential for misinformation and confusion must also be addressed. It will be important for medical providers to be involved in the development of medical-focused large language models. Diligent provider oversight and curated training data will be needed as we explore the utility of similar LLMs within the field of otolaryngology.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding Support:

This work was supported in part by the National Institute of Deafness and Other Communication Disorders (NIDCD) Grant No. 5T32DC000027-33

References

1. Finney Rutten LJ, et al. , Online Health Information Seeking Among US Adults: Measuring Progress Toward a Healthy People 2020 Objective. Public Health Reports, 2019. 134(6): p. 617–625. [PubMed: 31513756]
2. Bergmo TS, et al. , Internet Use for Obtaining Medicine Information: Cross-sectional Survey. JMIR Form Res, 2023. 7: p. e40466. [PubMed: 36729577]
3. Amante DJ, et al. , Access to care and use of the Internet to search for health information: results from the US National Health Interview Survey. J Med Internet Res, 2015. 17(4): p. e106. [PubMed: 25925943]
4. O'Mathúna DP, How Should Clinicians Engage With Online Health Information? AMA J Ethics, 2018. 20(11): p. E1059–1066. [PubMed: 30499435]
5. Else H, Abstracts written by ChatGPT fool scientists. Nature, 2023. 613(7944): p. 423. [PubMed: 36635510]

6. Gilson A, 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: p. e45312. [PubMed: 36753318]
7. Sarraju A, et al. , Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. Jama, 2023.
8. Ayoub NF, et al. , Comparison Between ChatGPT and Google Search as Sources of Postoperative Patient Instructions. JAMA Otolaryngol Head Neck Surg, 2023.
9. Ayers JW, et al. , Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med, 2023. 183(6): p. 589–596. [PubMed: 37115527]
10. Gabriel J, et al. , The utility of the ChatGPT artificial intelligence tool for patient education and enquiry in robotic radical prostatectomy. Int Urol Nephrol, 2023.
11. Samaan JS, et al. , Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. Obes Surg, 2023. 33(6): p. 1790–1796. [PubMed: 37106269]
12. Shneyderman M, et al. , Readability of Online Materials Related to Vocal Cord Leukoplakia. OTO Open, 2021. 5(3): p. 2473974x211032644.
13. Hannabass K and Lee J, Readability Analysis of Otolaryngology Consent Documents on the iMed Consent Platform. Mil Med, 2022.
14. Kim JH, et al. , Readability of the American, Canadian, and British Otolaryngology-Head and Neck Surgery Societies’ Patient Materials. Otolaryngol Head Neck Surg, 2022. 166(5): p. 862–868. [PubMed: 34372717]
15. Weis BD, Health Literacy: A Manual for Clinicians. American Medical Association, American Medical Foundation, 2003.
16. Shoemaker SJ, Wolf MS, and Brach C, Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns, 2014. 96(3): p. 395–403. [PubMed: 24973195]
17. Rothwell JD, In Mixed Company 11e: Communicating in Small Groups and Teams. 2021: Oxford University Press, Incorporated.
18. Johnson D, et al. , Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Res Sq, 2023.
19. Ayoub NF, et al. , Head-to-Head Comparison of ChatGPT Versus Google Search for Medical Knowledge Acquisition. Otolaryngol Head Neck Surg, 2023.
20. Patel MJ, et al. , Analysis of Online Patient Education Materials on Rhinoplasty. Facial Plast Surg Aesthet Med, 2022. 24(4): p. 276–281. [PubMed: 33844930]
21. Kasabwala K, et al. , Readability assessment of patient education materials from the American Academy of Otolaryngology--Head and Neck Surgery Foundation. Otolaryngol Head Neck Surg, 2012. 147(3): p. 466–71. [PubMed: 22473833]
22. Chen LW, et al. , Search Trends and Quality of Online Resources Regarding Thyroidectomy. Otolaryngol Head Neck Surg, 2021. 165(1): p. 50–58. [PubMed: 33138718]
23. Misra P, et al. , Readability analysis of internet-based patient information regarding skull base tumors. J Neurooncol, 2012. 109(3): p. 573–80. [PubMed: 22810759]
24. Yang S, Lee CJ, and Beak J, Social Disparities in Online Health-Related Activities and Social Support: Findings from Health Information National Trends Survey. Health Commun, 2021: p. 1–12.
25. Eysenbach G, The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Medical Education, 2023. 9(1): p. e46885. [PubMed: 36863937]
26. Xu L, et al. , Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review. JMIR Cancer, 2021. 7(4): p. e27850. [PubMed: 34847056]
27. Pham KT, Nabizadeh A, and Selek S, Artificial Intelligence and Chatbots in Psychiatry. Psychiatr Q, 2022. 93(1): p. 249–253. [PubMed: 35212940]

28. Chakraborty C, et al. , Overview of Chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. Front Artif Intell, 2023. 6: p. 1237704. [PubMed: 38028668]

29. Liu J, Wang C, and Liu S, Utility of ChatGPT in Clinical Practice. J Med Internet Res, 2023. 25: p. e48568. [PubMed: 37379067]

30. van Dis EAM, et al. , ChatGPT: five priorities for research. Nature, 2023. 614(7947): p. 224–226. [PubMed: 36737653]

31. Rich AS and Gureckis TM, Lessons for artificial intelligence from the study of natural stupidity. Nature Machine Intelligence, 2019. 1(4): p. 174–180.

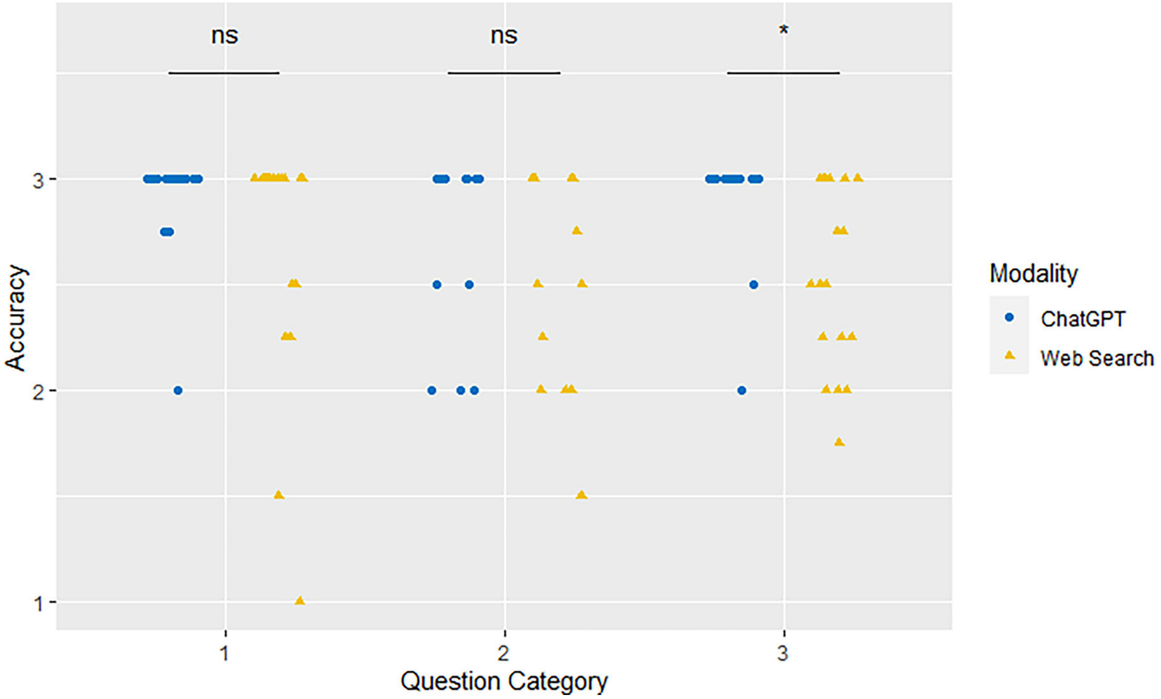

Figure 1.

Accuracy of ChatGPT and traditional web search responses grouped by question category. The scores were equivalent for questions in Category 1: Fact and Category 2: Policy. ChatGPT scored higher in Category 3: Diagnosis and Recommendations, compared to web search. ns - not significant, * - significant at p<0.05

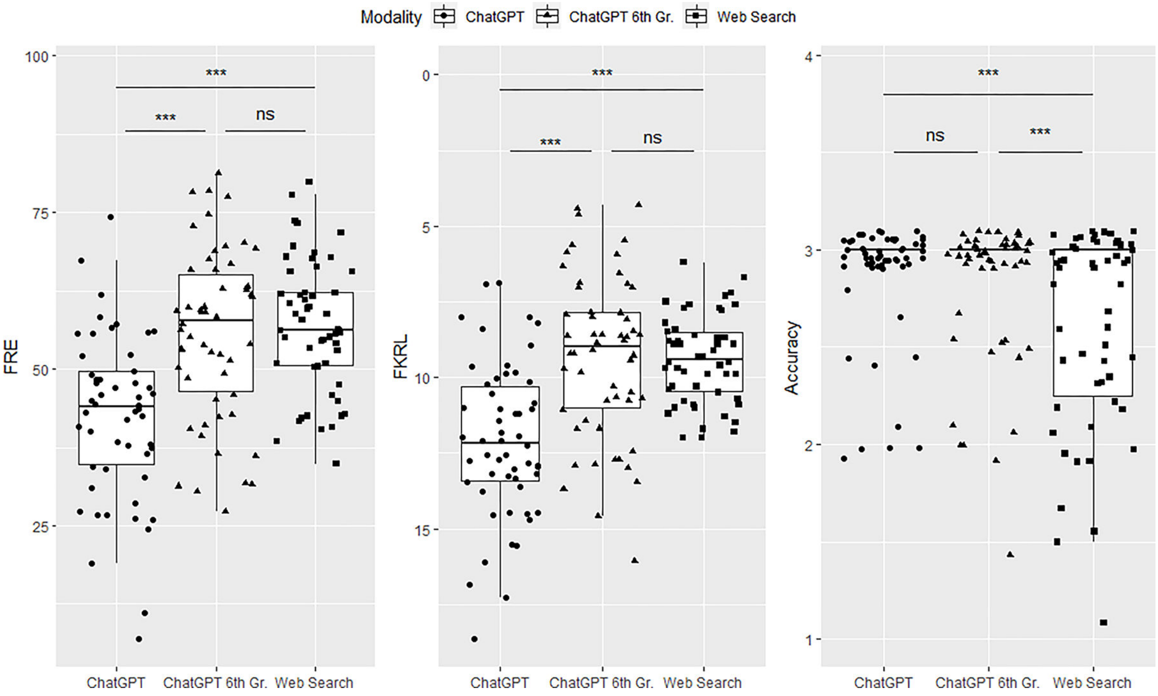

Figure 2.

Boxplots comparing readability and accuracy across the three search modalities. ns - not significant, *** - significant at p<0.01

Table 1.

Average readability and understandability scores of Chat GPT and web search responses to generated patient questions.

	ChatGPT	Web search	Mean difference, 95% CI
FRE	42.3 ± 14.2	56.2 ± 17.4	13.9 (10.6 – 15.3)
FKRL	12.1 ± 2.8	9.4 ±3.3	−2.7 (−3.6 – −1.7)
PEMAT Understandability	93.8 (57.1–100.0)	88.4 (42.9 – 100.0)	−5.3 (−1.2 – 9.6)

FRE: Flesch Reading Ease, FKGL: Flesch-Kincaid Grade Level