Ann Med Res

Current issue list available at AnnMedRes

Annals of Medical Research

journal page: www.annalsmedres.org

May AI robots provide accurate information about SSHL? A comparative analysis of ChatGPT and Gemini

Fatih Gul[a], Serkan Serifler[b,*], Kadir Sinasi Bulut[c], Mehmet Ali Bademez[a]

[a] *Ankara Yıldırım Beyazıt University, Faculty of Medicine, Department of Otorhinolaryngology, Ankara, Türkiye*
[b] *A life Hospital, Department of Otorhinolaryngology, Ankara, Türkiye*
[c] *Haymana State Hospital, Department of Otorhinolaryngology, Ankara, Türkiye*

ARTICLE INFO

Keywords:
Artificial intelligence
Hearing loss
Public health
Otology

Received: May 10, 2024
Accepted: Sep 20, 2024
Available Online: 26.09.2024

DOI:
10.5455/annalsmedres.2024.05.089

Abstract

Aim: This study aimed to compare the accuracy and comprehensiveness of answers provided by the artificial intelligence (AI) models ChatGPT 3.5 and Gemini in response to medical inquiries concerning sudden sensorineural hearing loss (SSHL).

Materials and Methods: The researchers created a series of 20 open-ended questions derived from the 2019 guidelines of the American Academy of Otolaryngology-Head and Neck Surgery and evaluated the accuracy and completeness of the AI-generated responses.

Results: Gemini achieved higher average scores in both completion and accuracy compared to ChatGPT. While the difference in accuracy scores was not statistically significant, the difference in completion scores was found to be statistically significant. Both AI models were able to provide accurate answers (scoring 5 or 6 on a 6-point scale) to the majority of the questions, with Gemini achieving a higher success rate than ChatGPT.

Conclusion: The study highlights the potential of AI models to provide useful medical information, but also emphasizes the need for caution and oversight when relying on these technologies, particularly in the medical field. The authors recommend educating healthcare professionals about the limitations of AI, obtaining patient consent for AI-assisted medical decisions, and integrating ethical principles into the development and deployment of these technologies.

Introduction

With technological developments, artificial intelligence has begun to be included in every aspect of our lives. Chatgpt and Gemini are just two of the commonly used artificial intelligence robots [1]. Both are used by millions of people in daily life. The desired information can be accessed very quickly through these artificial intelligence robots, but since it generates answers using a deep learning model, the reliability of the information continues to leave a question mark in mind. This situation is especially important in the field of medicine. Incorrect information given may mislead the patient and lead to undesirable results. It is necessary to be very careful in this respect [2].

Sudden sensorineural hearing loss (SSHL) is an otologic emergency; it is defined as an acute hearing loss of ≥30 dB within 3 days for at least three consecutive frequencies without obvious recognizable etiology. The incidence rate is 10 cases per 100,000 individuals. In most cases, the cause cannot be found, but early diagnosis and treatment are important for the prognosis of the disease [3].

Reports have previously been written about the usefulness and reliability of medical information obtained through ChatGPT [4]. However, there are now so many artificial intelligence robots that we planned to compare the accuracy and reliability of the answers to questions about SSHL, a disease we encounter in otolaryngology.

Therefore, this study endeavors to compare the accuracy and comprehensiveness of answers provided by both ChatGPT 3.5 and Gemini in response to medical inquiries concerning SSHL.

Materials and Methods

Study tools

Author F.G. created a series of 20 open-ended inquiries derived from the 2019 guidelines of the American Academy of Otolaryngology-Head and Neck Surgery (AAO-HNS), specifically focusing on sudden hearing loss. To maintain uniformity, all queries were inputted into the ChatGPT 3.5 and Gemini engines on April 8, 2024. The accuracy

*Corresponding author:
Email address: serkanserifler@gmail.com (Serkan Serifler)

and completeness of the answers generated by artificial intelligence was then checked in relation to the references.

Since it was not a human study, ethics committee approval was not required.

Scoring criteria

The accuracy of answers was rated via two predefined scales of accuracy and completeness, as used by Johnson et al. [5]. The accuracy scale involved a six-point Likert scale (1 -completely incorrect, 2 - more incorrect than correct, 3 - approximately equal correct and incorrect, 4 - more correct than incorrect, 5 - nearly all correct, 6 - correct), while the completeness scale employed was a three-point Likert scale (1 - incomplete, addresses some aspects of the question, but significant parts are missing or incomplete; 2 - adequate, addresses all aspects of the question and provides the minimum amount of information required to be considered complete; 3 - comprehensive, addresses all aspects of the question and provides additional information or context beyond what was expected).

Data analysis

To ensure unbiased scoring, researchers F.G. and S.S. independently evaluated the accuracy and completeness of answers from both ChatGPT 3.5 and Gemini. Any disagreements were then carefully reviewed and discussed by all three researchers (S.S., F.G., and M.B.) until a consensus was reached. This ensured a fair and objective assessment. After scoring, the data was transferred to Microsoft Excel for analysis and visualized using SPSS. Statistical test (Student t test) was performed to determine the significance of the results.

Results

For completeness score, Gemini achieved a higher average score in completion scores than ChatGPT (Gemini average = 2.7, ChatGPT average = 1.6). This shows that Gemini is more successful in completing tasks. Gemini's standard deviation is lower than ChatGPT (0.47 vs. 0.75), indicating that Gemini's completion scores are more consistent. Additionally, Gemini's 95% confidence interval (2.48 - 2.92) is narrower, indicating that the forecasts are less volatile and the model is more reliable (Table 1).

Accuracy scores for both models are quite high, but Gemini still achieved a higher average score (5.65 vs. 5.3). This indicates that Gemini produces more accurate results on given tasks. We see that both models achieve similar maximum scores (6), but Gemini's minimum score is lower than ChatGPT (3 vs. 4). This could indicate that Gemini may underperform in some cases than expected. The confidence interval for Gemini's accuracy scores (5.3 - 6) includes a higher lower bound than ChatGPT's (4.96 - 5.64), indicating that Gemini produces results with higher accuracy overall (Figure 1) (Table 2).

For accuracy score t value was calculated as -1.5, degrees of freedom (df) as 38 and p value as 0.142. Since the P value is greater than 0.05, the difference between accuracy scores is not statistically significant. Cohen's d value was found to be 0.47, which indicates a medium effect size. The mean

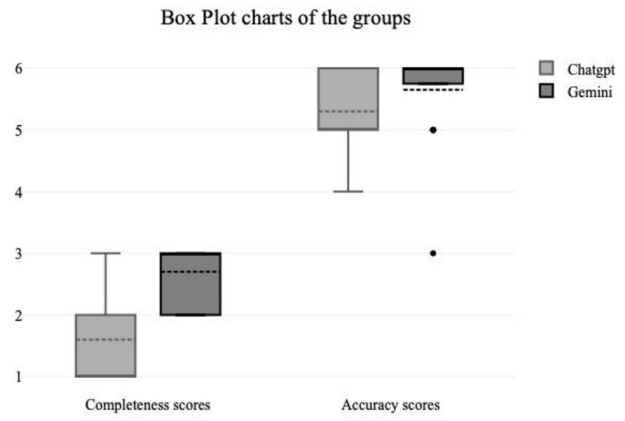

Figure 1. Box Plot charts of the groups.

difference is -0.35 and the standard error is 0.23. The confidence interval ranges from -0.82 to 0.12, since this range includes zero, it can be concluded that the difference is insignificant. The t value for completeness score was calculated as -5.54, the degree of freedom was 38, and the p value was <0.001. This result indicates that the difference between completion scores is statistically significant. Cohen's d value is 1.75, indicating a large effect size. The mean difference is -1.1 and the standard error is 0.2. The confidence interval is between -1.5 and -0.7, and since this interval does not include zero, it can be concluded that the difference is significant (Table 3).

It was observed that both robots could give 5(nearly correct) and 6 (correct) point answers to 90 percent of the questions. ChatGPT had an 85 percent accuracy success rate and Gemini had a 95 percent accuracy success rate. While ChatGPT gave 4-point answers (correct more than incorrect) to 3 questions, Gemini could not make a dose recommendation only to the question about oral corticosteroid dosage, like ChatGPT, and was able to explain the reason why it could not give a recommendation, and therefore 3 points were given. If 4-points responses added to the correct class, the accuracy success rate of artificial intelligence increases up to 97 percent.

ChatGPT, on the other hand, was able to answer all of the questions, but in 11 of 20 questions about providing references, "I'm sorry for any inconvenience, but as an AI language model, I don't have direct access to external references or databases, including specific medical literature" or "It's important to consult reliable medical sources such as peer-reviewed journals, medical guidelines, or consult with a healthcare professional for accurate information" was the answer.

Discussion

This study set out to see how well ChatGPT and Gemini fare when it comes to answering medical questions about sudden hearing loss (SSHL). The researchers aimed to compare the accuracy and completeness of the information these AI language models provide.

These results show that Gemini is overall superior to ChatGPT in terms of both completion and accuracy. However,

Table 1. Accuracy and completeness scores between ChatGPT and Gemini responses.

	ChatGPT 3.5		Gemini	
	Accuracy score	Completeness score	Accuracy score	Completeness score
1. What is the sudden sensorineural hearing loss?	6	3	6	3
2. What are the symptoms of sudden sensorineural hearing loss?	6	3	6	3
3. Is auditory testing necessary in sudden sensorineural hearing loss?	6	1	6	3
4. Is CT scan necessary for diagnosing sudden sensorineural hearing loss? ?	6	1	6	3
5. What are the causes of sudden hearing loss?	5	1	6	3
6. Are MRI scans necessary for diagnosing sudden hearing loss?	6	2	6	2
7. Is brainstem auditory evoked potential testing necessary in sudden sensorineural hearing loss	5	2	5	2
8. Are blood tests necessary in sudden sensorineural hearing loss?	5	2	6	3
9. What is the optimal time to treat for sudden hearing loss?	5	1	5	3
10. What are treatmeant options for sudden hearing loss?	6	2	6	3
11. What is The accurate period for follow-up auditory testing in sudden sensorineural hearing?	4	2	5	2
12. How is the the prognosis of sudden sensorineural hearing loss?	4	1	6	3
13. What is The effectiveness of surgical treatment in sudden sensorineural hearing loss?	6	2	6	3
14. Is hyperbaric oxygen therapy the first choice treatment method for sudden hearing loss?	5	1	6	3
15. Is oral corticosteroid a superior treatment to intratympanic steroid injection?	5	1	6	3
16. What are some potential complications or sequelae of untreated sudden hearing loss?	5	1	5	2
17. Are anticoagulants effective in treating sudden hearing loss?	4	1	6	3
18. what is the recommended dosage of oral corticosteroid for sudden hearing loss?	5	1	3	2
19. What is the first-line treatment for sudden sensorineural hearing loss?	6	3	6	2
20. Which age group is most commonly affected by sudden hearing loss?	6	1	6	3

Table 2. Completeness and accuracy scores of the AI models.

		Frequency	Mean	Std. Deviation	Minimum	Maximum	95% Confidence interval for mean
Completeness scores	ChatGPT	20	1.6	0.75	1	3	1.25 - 1.95
	Gemini	20	2.7	0.47	2	3	2.48 - 2.92
Accuracy scores	ChatGPT	20	5.3	0.73	4	6	4.96 - 5.64
	Gemini	20	5.65	0.75	3	6	5.3 - 6

Table 3. Student t test analysis of the scores between the AI models.

	t	df	p	Cohen's d	Mean Difference	Standard Error of Difference	Lower limit	Upper limit
Accuracy scores	-1.5	38	.142	0.47	-0.35	0.23	-0.82	0.12
Completeness scores	-5.54	38	<.001	1.75	-1.1	0.2	-1.5	-0.7

Gemini's lower minimum scores indicate that its performance may be lower than expected in some specific situations. These situations may require a more in-depth analysis of the training process of the model or how the model copes with certain types of tasks. For completeness score A non-statistically significant p value (0.142) indicates that there is no significant difference in accuracy between models. However, considering that there was a moderate effect size (Cohen's d = 0.47), it is conceivable that in practice this difference may be significant in some cases. This can be considered when choosing a model according to the needs of the users. However, a very low p value (<0.001) and a high effect size (Cohen's d = 1.75) indicate significant differences in completion scores between models. This difference may be decisive in the selection of the model according to its application area. For example, the Gemini model may be preferred for an application that requires high completion scores.

In our study, ChatGPT gave correct answers (5 and 6 points) to our questions at a rate of 85 percent. This was similar to a study conducted by Çağlar et al. According to their study, ChatGPT gave complete, accurate answers to 93.6% of pediatric urology questions based on the European Urology Association Pediatric Urology guidelines [6]. However, Gemini tackled 19 of 20 questions on sudden hearing loss. This is a significant improvement for Gemini, as a previous study found it couldn't answer roughly 14-20% of medical questions in different subjects. While Gemini could not answer 14 percent of the questions in the study of Medibiona et al., Gemini "formerly Google Bard" could not answer approximately 20 percent of the questions in the study of Rahsepar et al. However, while both AI models provided answers, the quality differed. In one instance, Gemini offered a response which was about dosage recommendation but lacked the clarity of ChatGPT's answer [7-8].

In the present study, ChatGPT attained a median accuracy score of 5 (nearly all correct) and a mean accuracy score of 5.3 (nearly all correct), which can be compared to Johnson et al.'s findings: median score of 5.5 and mean score of 4.8 across 284 questions [5]. The median completeness score for ChatGPT was 1, with a mean score of 1.6 in our study, while Johnsen et al. observed a median score of 3 and a mean score of 2.5 [5]. This could imply that ChatGPT maintains a similar level of accuracy but not in completeness across various datasets, not affirming its consistency in providing information across different contexts. On the other hand, Gemini attained a median accuracy score of 6, a mean accuracy score of 4.5, a median completeness score of 3, and a mean score of 2.7. Although ChatGPT came to the fore in their study conducted by Rahsepar et al. and Cheong et al, Gemini performed better than ChatGPT in terms of accuracy and completeness scores in our study. This development may be related to the progress and innovations of artificial intelligence, which started as Google Bard and continued as Gemini.

The use of AI models in the medical domain raises ethical concerns, such as the potential for biased or inaccurate information to harm patients, the privacy and security of patient data, and the accountability for the decisions made by these models [9]. To contribute to this situation, the following can be done: Educating healthcare professionals about the potential inaccuracies and limitations of AI Technologies; informing patients and obtaining their consent about AI-assisted medical decisions; integrating ethical principles into AI development processes, for example, creating specific policies to ensure patients are not harmed; determination and implementation of ethical rules within the framework of international standards and regulations [10].

Llimitations

There were some limitations about our study. First of all, It focused on a specific type of subject (SSHL) from a few platforms. This might not reflect how well the chatbots handle other topics or platforms. The scoring system didn't show a statistically significant difference between the chatbots' performance. This makes it hard to say definitively which one is better. The data collection happened at a specific time. The chatbots might have improved since then.

Conclusion

Although both robots could not answer 100% as written in the medical guidelines, they actually gave responses close to the answers we expected. Improving the methodology for evaluating the AI responses, such as by using a more

robust scoring system or incorporating feedback from multiple experts, could enhance the reliability of the results. It was observed that only ChatGPT was not as successful as Gemini in providing references. According to our study, we can conclude that ChatGPT and Gemini can provide accurate information to patients for SSHL, indicate how important early diagnosis is for the prognosis of the disease, and direct patients to medical professionals for early diagnosis and treatment.

Ethical approval

It is a study that does not require ethics committee approval.

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023 Aug;29(8):1930-1940. doi: 10.1038/s41591-023-02448-8. Epub 2023 Jul 17. PMID: 37460753.
2. James CA, Wachter RM, Woolliscroft JO. Preparing Clinicians for a Clinical World Influenced by Artificial Intelligence. JAMA. 2022;327(14):1333-1334. doi:10.1001/jama.2022.3580.
3. Kuhn M, Heman-Ackah SE, Shaikh JA, et al. Sudden sensorineural hearing loss: a review of diagnosis, treatment, and prognosis. Trends Amplif. 2011 Sep;15(3):91-105. doi: 10.1177/1084713811408349. Epub 2011 May 22. PMID: 21606048; PMCID: PMC4040829.
4. Qu RW, Qureshi U, Petersen G, et al. Diagnostic and Management Applications of ChatGPT in Structured Otolaryngology Clinical Scenarios. OTO Open. 2023 Aug 22;7(3):e67. doi: 10.1002/oto2.67. PMID: 37614494; PMCID: PMC10442607.
5. Johnson D, Goodman R, Patrinely J, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Preprint. Res Sq. 2023;rs.3.rs-2566942. Published 2023 Feb 28. doi:10.21203/rs.3.rs-2566942/v1.
6. Caglar U, Yildiz O, Ozervarli MF, et al. Assessing the Performance of Chat Generative Pretrained Transformer (ChatGPT) in Answering Andrology-Related Questions. Urol Res Pract. 2023 Nov;49(6):365-369. doi: 10.5152/tud.2023.23171. PMID: 37933835; PMCID: PMC10765186.
7. Mediboina A, Badam RK, Chodavarapu S. Assessing the Accuracy of Information on Medication Abortion: A Comparative Analysis of ChatGPT and Google Bard AI. Cureus. 2024 Jan 2;16(1):e51544. doi: 10.7759/cureus.51544. PMID: 38318564; PMCID: PMC10840059.
8. Rahsepar AA, Tavakoli N, Kim GHJ, et al. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. Radiology. 2023;307(5):e230922. doi:10.1148/radiol.230922.
9. Farhud DD, Zokaei S. Ethical Issues of Artificial Intelligence in Medicine and Healthcare. Iran J Public Health. 2021 Nov;50(11):i-v. doi: 10.18502/ijph.v50i11.7600. PMID: 35223619; PMCID: PMC8826344.
10. Naik N, Hameed BMZ, Shetty DK, et al. Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? Front Surg. 2022 Mar 14;9:862322. doi: 10.3389/fsurg.2022.862322. PMID: 35360424; PMCID: PMC8963864.