Sarcouncil journal of Medical sciences

ISSN(Online): 2945-3526

Volume- 03| Issue- 12| 2024

Research Article

Received: 19-10-2024 | Accepted: 20-11-2024 | Published: 20-12-2024

Comparison of Chat GPT and Gemini in ENT Evaluation Questions

Yeşim Yüksel[1], Özer Erdem Gür[1], Cihan Bedel[2]* ,Fatih Selvi[2] , Ökkeş Zortuk[3], and Günay Yıldız[2]

[1] *Department of Otolaryngology Head and Neck Surgery, Health Science University, Antalya Training and Research Hospital, Antalya, Turkey*

[2] *Department of Emergency Medicine, Health Science University, Antalya Training and Research Hospital, Antalya, Turkey*

[3] *Department of Emergency Medicine,Hatay Defne State Hospital, Hatay, Antalya*

Abstract: Introduction: Artificial intelligence (AI) is a field of computer science that aims to simulate human intelligence through technological advancement. It has found numerous applications across various domains. ChatGPT and Gemini, two significant models developed by OpenAI and Google AI, are employed effectively in numerous domains. The objective of this study was to conduct a comparative analysis of the responses generated by ChatGPT and Gemini in response to questions pertaining to ear, nose, and throat diseases (ENT). **Methods:** In order to compare the success of artificial intelligence systems, multiple-choice questions were selected from the examinations used during endoscopy nursing assistant training. The answers provided by the artificial intelligence models Gemini and Chat GPT 4.0, which are subheadings of anatomy, neurology, and infection, were then compared. **Results** In the comparison made for Gemini, the mean infection score was found to be significantly higher than the mean scores for neurology and anatomy (F=7.66, p=0.002). Similarly, in the comparison made for Chat GPT, the mean anatomy score was observed to be significantly higher than the mean scores for neurology and infection (F=7.23, p=0.003). **Conclusion:** In the comparison conducted for Gemini, it was observed that the mean infection score was markedly higher than the mean scores for neurology and anatomy. In the comparison conducted for Chat GPT, it was noted that the mean anatomy score was significantly higher than the mean scores for neurology and infection.

Keywords: Gemini, ChatGPT, ENT, Education.

INTRODUCTION

Technological developments and advances have led to significant developments in almost every sector. The use of artificial intelligence (AI) is one such advancement, and it has been employed with increasing frequency in many sectors in recent years [Cosic, K. *et al*., 2024]. ChatGPT, developed by OpenAI and Google Gemini (the successor of Google Bard), are prominent examples of artificial intelligence-focused tools that have transformed sectors such as healthcare, finance, and education [Cheng, S. W. *et al*., 2023]. ChatGPT and Google Gemini, the two most frequently utilized artificial intelligence models in recent years, have the potential to be beneficial to numerous sectors [Dave, T. *et al*., 2023; Franco D'Souza, R. *et al*., 2023]. This is due to their ability to leverage advanced machine learning and deep learning techniques to enhance human-machine communication and automate complex tasks [Menz, B. D. *et al*., 2024]. Although it is known that ChatGPT and Google AI are frequently used in medical subjects, there is a paucity of evidence demonstrating their superiority. The efficacy of ChatGPT and Gemini, particularly in local and medical licensure examinations, has been evaluated in numerous publications, contributing to specialized domains with an accuracy rate of 70-80% [Kumah-Crystal, Y. *et al*., 2023; Uzunay, H. *et al*., 2021]. Both models have been tested in numerous examination contexts, yet they have demonstrated disparate levels of success and have been identified as promising candidates for further technological development [Yaïci, R. *et al*., 2024]. Despite recent studies indicating that ChatGPT and Gemini can facilitate medical history collection, symptom assessment, and decision support in ear, nose, and throat (ENT) clinic exams and education, thereby enhancing diagnostic accuracy and patient care, there is a paucity of literature on this subject [Frosolini, A. *et al*., 2023]. The evaluation of the adequacy of these technological products in the field of otolaryngology has yet to be elucidated. While their medical utility is established, the extent to which they can be beneficial remains unclear. In this study, we sought to compare the responses generated by ChatGPT and Gemini to otolaryngological queries.

METHODS

In order to compare the success rates of artificial intelligence systems, the study employed questions from the examinations utilized during the training of assistants in the field of Ear, Nose, and Throat Diseases. The anatomy section comprises 15 multiple-choice questions, which address the anatomical structures and the circulatory system in the head and neck region. The neurology section comprises 10 questions, which evaluate the nervous system in the head and neck region and the motor and sensory structures of this system.

***Corresponding Author:** Cihan Bedel

DOI-https://doi.org/10.5281/zenodo.14533624

The infection section consists of 10 multiple-choice questions on infections that occur in the head and neck region and the microorganisms that cause these infections. The questions were randomly selected from the training exam held in 2024.

Artificial Intelligence Application

In the present study, the free and open access versions of the Gemini multi-language application, which was launched by Google AI in 2023, and the Chat GPT 4.0 multi-language application, which was launched by OpenAI in 2023, were utilized. First, the scope was specified with an introductory prompt on the form of the exam. Subsequently, the exam questions were presented as a prompt, and the obtained answer options were recorded. To evaluate consistency, these processes were repeated ten times on different days, using browsers with zero cookie elements and changing the question locations. The obtained answers were recorded.

STATISTICAL ANALYSIS

After the data determined in the study was recorded, the relationship between them was analyzed using SPSS version 27 (IBM Co. USA). Graphpad Prism 9 was used to create the graphics. The data was defined and divided into groups. While percentage and frequency values were used to display the data defined as categorical, chi-square test and Monte Carlo correction were applied to evaluate the relationship between them. While mean ± Standard deviation was used to define the numerical data on which distribution analysis was performed, parametric tests and Tukey were used as posthoc tests for the relationships between them. The p value below 0.05 from the determined data were considered significant.

RESULTS

In regard to the question of anatomical accuracy, the mean success score for the Gemini program was found to be 58.67 ± 16.87, while the mean success score for the ChatGPT program was 66.67 ± 0.00. No statistically significant difference was observed between the two (p = 0.075, Table 1-3; Figure 1a). In the domain of neurology, the mean success score for Gemini was 50.00 ± 16.33, while that for ChatGPT was 56.00 ± 5.164. No significant difference was observed between the two (p = 0.141, Figure 1b).

Table 1: Anatomy questions compare

Anatomy	Gemini	GPT	p-Value
Q1	0	0	-
Q2	0	0	-
Q3	0	0	-
Q4	10	10	-
Q5	0	0	-
Q6	8	10	0,474
Q7	0	0	-
Q8	0	0	-
Q9	8	10	0,474
Q10	8	10	0,474
Q11	8	10	0,474
Q12	8	10	0,474
Q13	8	10	0,474
Q14	10	10	-
Q15	10	10	-

Publisher: SARC Publisher

Table 2: Neurology questions compare

Neurology	Gemini	GPT	p-Value
Q1	0	0	-
Q2	0	6	0,005
Q3	8	10	0,474
Q4	8	10	0,474
Q5	10	0	<0,001
Q6	0	0	-
Q7	10	10	-
Q8	6	10	0,087
Q9	8	10	0,474
Q10	0	0	-

Table 3: Infections questions compare

Infections	Gemini	GPT	p-Value
Q1	9	9	0,999
Q2	3	3	0,999
Q3	9	1	0,001
Q4	9	9	0,999
Q5	10	9	0,500
Q6	3	3	0,999
Q7	10	9	0,500
Q8	3	3	0,999
Q9	9	9	0,999
Q10	9	3	0,020

The mean success score for infection questions was 74.00 ± 5.164 for Gemini, while the mean success score for Chat GPT was 58.00 ± 10.33. The difference in success between Gemini and Chat GPT was found to be statistically significant ($p < 0.001$, Figure 1c).

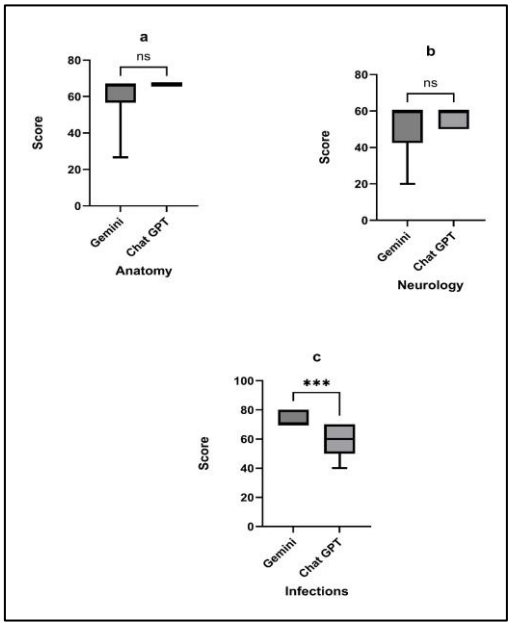

Figure 1: AI comparison by exam type

The comparison of the data for Gemini revealed a significant disparity in the mean infection score when compared to the mean scores for neurology and anatomy (F=7.66, p=0.002, Figure 2a). In the comparison made for ChatGPT, it was observed that the mean anatomy score was significantly higher than the mean neurology and infection scores (F=7.23, p=0.003, Figure 2b).

Publisher: SARC Publisher

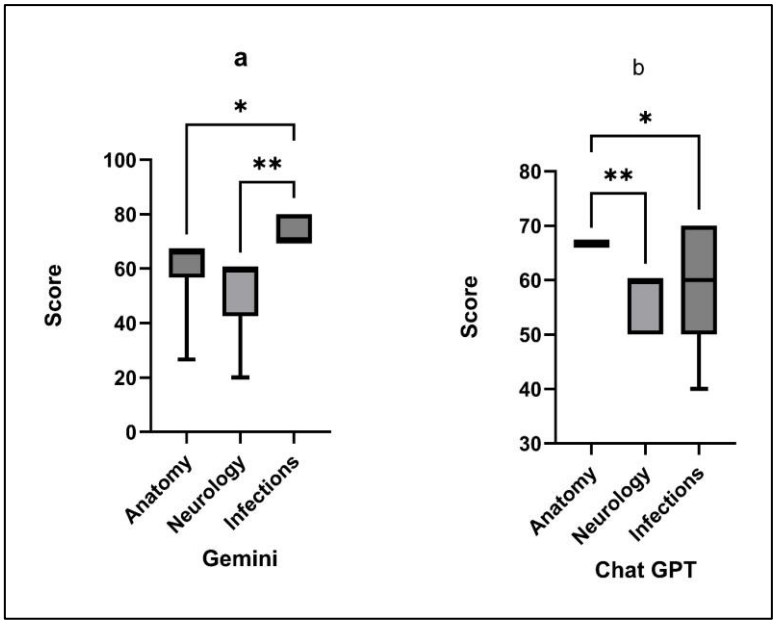

Figure 2: Comparison by exam sections for AI

DİSCUSSİON

In the present study, the responses provided by Gemini and ChatGPT to questions pertaining to the ear, nose, and throat (ENT) were compared. Upon examination of the responses classified under the anatomical, neurological, and infection subheadings, it was observed that the mean score for infection was markedly higher for Gemini compared to the mean scores for neurology and anatomy. Conversely, the mean score for anatomy was found to be significantly higher than the mean scores for neurology and infection in the case of ChatGPT.

The incorporation of AI models, including ChatGPT and Gemini, into the field of ENT science has yielded advantages in numerous domains, particularly in the realms of clinical decision-making and patient engagement [Lechien, J. R. *et al*., 2024]. The appropriate and efficacious deployment of these artificial intelligence models in suitable contexts has partially mitigated the shortcomings and emerged as a contemporary approach that can be efficacious in ENT [Özcan, İ. *et al*., 2006; Fiore, M. *et al*., 2024]. The judicious use of such technological advances confers benefits in numerous domains. With these models, which offer advantages in a multitude of medical domains, intricate medical scenarios can be resolved with the requisite guidance and can also provide clinical benefits to physicians in ENT practice [Temsah, M. H. *et al*., 2023].

In a study conducted by Lorenzi *et al*. that compared the performance of ChatGPT and Gemini in the analysis of ear, nose, and throat (ENT) examinations, ChatGPT demonstrated superior capabilities compared to other artificial intelligence (AI) methods [Lorenzi, A. *et al*., 2024]. Similarly, in a different study, it was reported that ChatGPT could produce more practical and applicable projects than Gemini in the context of research project development [Bedel, C. *et al*., 2021]. A comparison of the effectiveness of ChatGPT and Gemini in referencing articles on ENT subjects revealed that ChatGPT was more effective and accurate. A comparison of the effectiveness of AI technologies in diagnosis revealed that ChatGPT is useful in simple diagnoses but lacks the capacity to provide accurate information in complex technologies [Gill, G. S. *et al*., 2024]. In the context of oncological diseases within the domain of ENT practice, ChatGPT was observed to demonstrate superior efficacy in formulating guideline-based treatment recommendations relative to Gemini [Antaki, F. *et al*., 2023]. However, its performance in the domain of surgical decision-making was deemed to be inadequate. In evaluations of AI models' performance in medical education and examination settings, ChatGPT demonstrated superior outcomes compared to Gemini in ophthalmology examinations [Botross, M. *et al*., 2024]. Additionally, studies have reported its effectiveness in both straightforward and more challenging questions. However, neither model provided answers that were deemed to be particularly reliable from a scientific perspective [Mete, U. *et al*., 2024]. Similarly, in the context of sudden sensorineural hearing loss, Gemini was

Publisher: SARC Publisher

observed to demonstrate superior performance and proficiency in terms of medical accuracy [Gül, F. *et al*., 2024]. In a study conducted for the ophthalmology board exam, Gemini was found to answer more than half of the questions correctly [Gill, G. S. *et al*., 2024]. Both ChatGPT and Gemini exhibited significant deficiencies in their responses to cataract-related queries, with ChatGPT displaying lower readability compared to Gemini. In a recent study, ChatGPT-4 has been reported to be effective in educational applications in Virology [Sallam, M. *et al*., 2024]. The use of AI such as ChatGPT in exams poses challenges to academic integrity and the integrity of online assessments, requiring strong security measures and ethical considerations [Kochanek, K. *et al*., 2024]. The potential of AI to assist medical practice is acknowledged, but the variability in performance and the need for oversight are also clearly evident in studies. In comparison of technological advances within themselves, studies have shown that GPT-4 is more accurate and capable than its lower versions for the Apon Orthopedic Surgery Examination Board [Giorgino, R. *et al*., 2023]. In our study, in the comparison made for Gemini, it was seen that the mean infection score was significantly higher than the mean neurology and anatomy scores; in the comparison made for Chat GPT, the mean anatomy score was significantly higher than the mean neurology and infection scores.

It should be noted that the present study is subject to certain limitations. Primarily, it is a single-center study. Furthermore, although the questions were randomly selected, the inclusion of a greater number of questions and a broader range of question models in the comparison was necessary to ensure the effective comparison of subgroups. To ascertain the significance of the subject, it is essential to conduct studies utilising diverse models and a multitude of subgroups, particularly with a larger population, where disparate learning methodologies will be integrated.

CONCLUSION

In the comparison conducted for Gemini, it was observed that the mean infection score was markedly higher than the mean scores for neurology and anatomy. In the comparison conducted for Chat GPT, it was noted that the mean anatomy score was significantly higher than the mean scores for neurology and infection.

REFERENCES

1. Cosic, K., Kopilas, V. & Jovanovic, T. "War, emotions, mental health, and artificial intelligence." *Frontiers in Psychology*, 15 (2024): 1394045.
2. Cheng, S. W., Chang, C. W., Chang, W. J., Wang, H. W., Liang, C. S., Kishimoto, T., Chang, J. P., Kuo, J. S. & Su, K. P. "The now and future of ChatGPT and GPT in psychiatry." *Psychiatry and Clinical Neurosciences*, 77.11 (2023): 592-596.
3. Dave, T., Athaluri, S. A. & Singh, S. "ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations." *Frontiers in Artificial Intelligence*, 6 (2023): 1169595.
4. Franco D'Souza, R., Amanullah, S., Mathew, M. & Surapaneni, K. M. "Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes." *Asian Journal of Psychiatry*, 89 (2023): 103770.
5. Menz, B. D., Kuderer, N. M., Chin-Yee, B., Logan, J. M., Rowland, A., Sorich, M. J. & Hopkins, A. M. "Gender representation of health care professionals in large language model-generated stories." *JAMA Network Open*, 7.9 (2024): e2434997.
6. Kumah-Crystal, Y., Mankowitz, S., Embi, P. & Lehmann, C. U. "ChatGPT and the clinical informatics board examination: The end of unproctored maintenance of certification?" *Journal of the American Medical Informatics Association*, 30.9 (2023): 1558-1560.
7. Uzunay, H., Selvi, F., Bedel, C. & Karakoyun, O. F. "Comparison of ETCO2 value and blood gas PCO2 value of patients receiving non-invasive mechanical ventilation treatment in emergency department." *SN Comprehensive Clinical Medicine*, 3.8 (2021): 1717-1721.
8. Yaïci, R., Cieplucha, M., Bock, R., Moayed, F., Bechrakis, N. E., Berens, P., Feltgen, N., Friedburg, D., Gräf, M., Guthoff, R., Hoffmann, E. M., Hoerauf, H., Hintschich, C., Kohnen, T., Messmer, E. M., Nentwich, M. M., Pleyer, U., Schaudig, U., Seitz, B., Geerling, G. & Roth, M. "ChatGPT und die deutsche Facharztprüfung für Augenheilkunde: eine Evaluierung [ChatGPT and the German board examination for ophthalmology: an evaluation]." *Ophthalmologie*, 121.7 (2024): 554-564.
9. Frosolini, A., Franz, L., Benedetti, S., Vaira, L. A., de Filippis, C., Gennaro, P., Marioni, G. & Gabriele, G. "Assessing the accuracy of

Publisher: SARC Publisher

ChatGPT references in head and neck and ENT disciplines." *European Archives of Otorhinolaryngology*, 280.11 (2023): 5129-5133.
10. Lechien, J. R., Briganti, G. & Vaira, L. A. "Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology-head and neck surgery." *European Archives of Otorhinolaryngology*, 281.4 (2024): 2159-2165.
11. Özcan, İ., Gedikli, Y., Özcan, K. M., Akdoğan, Ö., Albayrak, L. & Dere, H. "Kikuchi Hastalığı." *Turkiye Klinikleri Journal of Medical Sciences*, 26.4 (2006): 457-460.
12. Fiore, M., Bianconi, A., Acuti Martellucci, C., Rosso, A., Zauli, E., Flacco, M. E. & Manzoli, L. "Vaccination hesitancy: Agreement between WHO and ChatGPT-4.0 or Gemini Advanced." *Annali di Igiene*, (2024).
13. Temsah, M. H., Aljamaan, F., Malki, K. H., Alhasan, K., Altamimi, I., Aljarbou, R., Bazuhair, F., *et al*. "ChatGPT and the future of digital health: A study on healthcare workers' perceptions and expectations." *Healthcare (Basel)*, 11.13 (2023): 1812.
14. Lorenzi, A., Pugliese, G., Maniaci, A., Lechien, J. R., Allevi, F., Boscolo-Rizzo, P., Vaira, L. A. & Saibene, A. M. "Reliability of large language models for advanced head and neck malignancies management: A comparison between ChatGPT 4 and Gemini Advanced." *European Archives of Otorhinolaryngology*, 281.9 (2024): 5001-5006.
15. Bedel, C., Korkut, M., Selvi, F. & Zortuk, Ö. "The immature granulocyte count is a new predictor of the 30-day mortality in intracerebral haemorrhage patients: Preliminary study." *Indian Journal of Neurosurgery*, 10.2 (2021): 114-120.
16. Gill, G. S., Tsai, J., Moxam, J., Sanghvi, H. A. & Gupta, S. "Comparison of Gemini Advanced and ChatGPT 4.0's performances on the ophthalmology resident ophthalmic knowledge assessment program (OKAP) examination review question banks." *Cureus*, 16.9 (2024): e69612.
17. Antaki, F., Touma, S., Milad, D., El-Khoury, J. & Duval, R. "Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings." *Ophthalmology Science*, 3.4 (2023): 100324.
18. Botross, M., Mohammadi, S. O., Montgomery, K. & Crawford, C. "Performance of Google's artificial intelligence chatbot 'Bard' (now 'Gemini') on ophthalmology board exam practice questions." *Cureus*, 16.3 (2024): e57348.
19. Mete, U. "Evaluating the performance of ChatGPT, Gemini, and Bing compared with resident surgeons in the otorhinolaryngology in-service training examination." *Turkish Archives of Otorhinolaryngology*, 62.2 (2024): 48-57.
20. Gül, F., Şerifler, S., Bulut, K. Ş. & Babademez, M. A. "May AI robots provide accurate information about SSHL? A comparative analysis of ChatGPT and Gemini." *Annals of Medical Research*, 31.9 (2024): 675-675.
21. Sallam, M., Al-Mahzoum, K., Almutawaa, R. A., Alhashash, J. A., Dashti, R. A., AlSafy, D. R., Almutairi, R. A. & Barakat, M. "The performance of OpenAI ChatGPT-4 and Google Gemini in virology multiple-choice questions: A comparative analysis of English and Arabic responses." *BMC Research Notes*, 17.1 (2024): 247.
22. Kochanek, K., Skarzynski, H. & Jedrzejczak, W. W. "Accuracy and repeatability of ChatGPT based on a set of multiple-choice questions on objective tests of hearing." *Cureus*, 16.5 (2024): e59857.
23. Giorgino, R., Alessandri-Bonetti, M., Luca, A., Migliorini, F., Rossi, N., Peretti, G. M. & Mangiavini, L. "ChatGPT in orthopedics: A narrative review exploring the potential of artificial intelligence in orthopedic practice." *Frontiers in Surgery*, 10 (2023): 1284015.

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Yüksel, Y., Gür, O.E., Bedel, C., Selvi, F., Zortuk, O. and Yıldız, G. "Comparison of Chat GPT and Gemini in ENT Evaluation Questions." *Sarcouncil journal of Medical sciences* 3.12 (2024): pp 17-22.

Publisher: SARC Publisher