Sarcouncil Journal of Medicine and Surgery

ISSN(Online): 2945-3534

Volume- 03| Issue- 12| 2024

Research Article **Received:** 05-10-2024 | **Accepted:** 20-11-2024 | **Published:** 08-12-2024

Comparison of Chat GPT and Gemini in Neurosurgical Evaluation Questions

Umut Ogün Mutlucan[1],Ökkeş Zortuk[2], Cihan Bedel[3], Fatih Selvi[3], and Cezmi Çağrı Türk[1]

[1]Department of Neurosurgery Health Science University, Antalya Training and Research Hospital, Antalya, Turkey

[2]Department of Emergency Medicine, Defne State Hospital, Hatay, Turkey

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

Abstract: Introduction: Artificial intelligence (AI) represents a transformative technology that emulates human intelligence through computer systems. ChatGPT and Gemini, two prominent conversational AI models developed by OpenAI and Google AI. we sought to evaluate the appropriateness of GPT 4 and Geminin's response to neurosurgical evaluation questions. **Methods:** The questions used in this study consist of 40 randomly selected questions from the board exam prepared by an internationally recognized neurosurgical institution. The study used the free version of the Gemini artificial intelligence system developed by Google and the free version of the ChatGPT 4.o multilingual algorithm system developed by OpenAI. Questions in four different categories were asked to two artificial intelligence programs in different orders on 10 different days. The results were recorded and the correct and incorrect answers were displayed. **Results** ChatGPT 4.0 was significantly more successful than Gemini in the anatomic and radiologic evaluation of the models ($p<0.001$). When compared with the answers given by the Gemini artificial intelligence system, a significant difference was found between the test classes ($F=26.33$, $p<0.001$). **Conclusion:** In our study, it was seen that ChatGPT 4.0 was more successful than Gemini in anatomical and radiological evaluation of brain surgery questions.

Keywords: Gemini, ChatGPT, Brain Surgery, Education.

INTRODUCTION

Artificial intelligence (AI) represents a transformative technology that emulates human intelligence through computer systems, thereby enabling machines to perform tasks that typically require human cognitive functions. Artificial intelligence (AI) encompasses a range of methodologies, including machine learning and deep learning, which facilitate the acquisition of knowledge from data and enable systems to enhance their capabilities over time (Hunter, D. J. *et al*., 2023). The term "AI" is used to describe the simulation of human intelligence processes by machines, particularly computer systems. Such processes include learning, reasoning, and self-correction (Salvagno, M. *et al*., 2023). In the field of healthcare, the advent of AI has brought about a revolutionary change in the way medical research and care are conducted. By processing complex data, AI is able to enhance the accuracy of diagnostics and treatment plans, particularly in areas such as oncology. The capabilities of ChatGPT and Gemini, two prominent conversational AI models developed by OpenAI and Google AI, respectively, have been evaluated across a range of domains, including biological knowledge retrieval, pharmacometrics, usability, business management, and clinical queries (Meyer, A. *et al*., 2024). The GPT-3 model, which was initially released and constituted a significant advancement in the field, has undergone subsequent developments and is now known as GPT-4. A number of studies have conducted a comprehensive comparison of Gemini and ChatGPT, evaluating their respective performance values (Rane, N. *et al*., 2024).

The neurosurgical board examination, which also contributes to a critical certification process, is widely used in many countries, particularly in the United States. These examinations assess the candidate's comprehensive knowledge and skills in the field of neurosurgery and ensure that they meet the high standards required for practice. The examination is divided into written and oral components, each testing different aspects of neurosurgical expertise (Hopkins, B. S. *et al*., 2023). The application of machine learning has become an important area of investigation in the field of neurosurgery in recent years, largely due to the advancement of relevant technology. In addition, the use of ChatGPT among neurosurgical residents is on the rise. An examination of the capabilities of ChatGPT may provide insight into whether and how residents can use it for learning, as well as the areas in which it can be applied (Sahin, M. C. *et al*., 2024). Despite the abundance of literature on neurosurgical board evaluation questions, there is a paucity of studies on the use of Gemini and GPT, which have emerged with the advancement of technology. The extent to which these technological products are appropriate for use in neurosurgical board evaluations remains unclear. However, previous studies have shown that their responses may facilitate diagnosis and clinical utility in the medical setting. Therefore, in this study, we sought to evaluate the

***Corresponding Author:** Cihan Bedel

DOI- https://doi.org/10.5281/zenodo.14314657

appropriateness of GPT 4 and Geminin's response to neurosurgical evaluation questions.

MATERIALS AND METODS

This study was carried out by evaluating the exam questions prepared by the association to which the institutions providing specialized training in the field of brain and neurosurgery are affiliated and used with the permission of the association, and the answers given to the questions by artificial intelligence programs.

Question Selection

The questions used in this study consist of 40 randomly selected questions from the board exam prepared by an internationally recognized neurosurgical institution. A total of 40 questions were created by combining questions from these question sets with a stratified selection of 10 questions in each category. It was composed of 5 multiple-choice questions. The questions in the exam were grouped to measure clinical, anatomic, pathologic, and radiologic skills. While the correct answer to each question was accepted as 1 point, question sets were prepared for each section with a maximum score of 10 and a minimum score of 0.

Analysis with Artificial Intelligence

The study used the free version of the Gemini artificial intelligence system developed by Google and the free version of the ChatGPT 4.o multilingual algorithm system developed by OpenAI. Questions in four different categories were asked to two artificial intelligence programs in different orders on 10 different days. The results were recorded and the correct and incorrect answers were displayed.

STATISTICAL ANALYSIS

After recording the responses and total scores obtained from the questions, the data form created was analyzed using SPSS version 27 (IBM Co. USA). Graphpad Prism 9 was used to organize the graphs. While percentages and frequencies were used to define categorical data, chi-squared test was used to evaluate the relationship between them. Distribution analysis of numerical data was performed, and data conforming to normal distribution were reported as mean ± standard deviation, while the relationship between them was analyzed using t-test and ANOVA. Tukey's test was used as a post-hoc test for ANOVA. Results with p-value over 0.05 from the determined data were considered significant.

RESULTS

When we examined the answers given by the AI models to the questions, we found that there were differences in the types and shapes of the questions in both the Gemini and ChatGPT 4.0 models ($p<0.001$). There were differences compared to the AI model in all question groups and subgroups, and the analysis of the questions is shown in Table 1.

Table 1: Comparison of correct answers by question groups

	Clinical			Anatomycal			Pathology			Radiology		
Q. no.	Gemini	GPT	p-Value	Gemini	GPT	p-Value	Gemini	GPT	p-Value	Gemini	GPT	p-Value
1	0	0	-	0	0	-	8	10	0,237	10	10	-
2	0	0	-	0	5	0,016	2	0	0,237	0	0	-
3	9	10	0,50	10	0	<0,001	8	8	0,418	0	6	0,005
4	9	0	<0,001	8	10	0,237	8	10	0,237	10	10	-
5	0	10	<0,001	2	2	0,418	10	0	0,001	10	10	-
6	0	1	0,5	0	5	0,016	0	4	0,043	0	10	<0,001
7	9	10	0,50	0	0	-	8	2	0,011	0	0	-
8	0	10	<0,001	0	0	-	2	2	0,418	0	0	-
9	9	0	<0,001	0	0	-	0	10	<0,001	10	10	-
10	10	10	-	0	0	-	8	0	<0,001	0	8	<0,001

When the relationship between the test scores and the subgroups was examined, it was seen that Gemini and ChatGPT 4.0 did not show a significant difference in clinical and pathologic scores. However, ChatGPT 4.0 was significantly more successful than Gemini in the anatomic and radiologic evaluation of the models ($p<0.001$; Table 2).

Publisher: SARC Publisher

Table 2: Comparison of average scores of tests according to artificial intelligence programs

	Gemini	GPT	p-Value
Clinical	4,6±1,26	5,1±0,31	0,119
Anatomy	2,0±0	2,2±4,21	<0,001
Pathology	5,4±1,26	4,6±0,84	0,380
Radiology	4,0±0	6,4±0,51	<0,001

When compared with the answers given by the Gemini artificial intelligence system, a significant difference was found between the test classes ($F=26.33$, $p<0.001$).

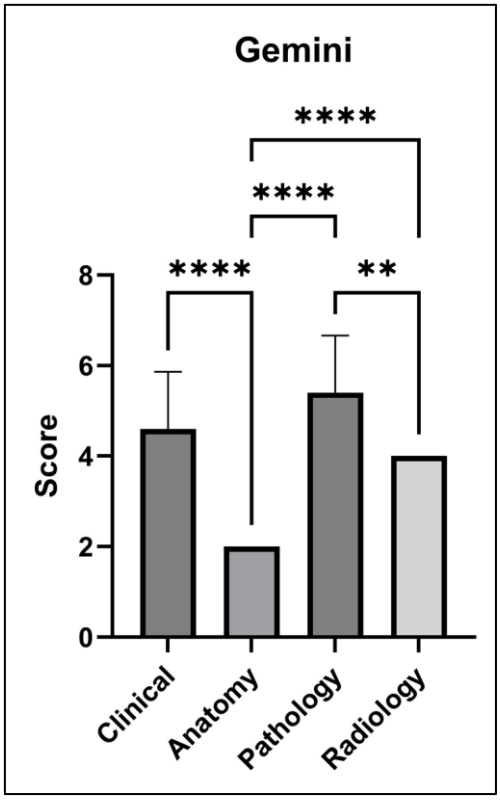

Figure 1. The comparison of the answers given by Gemini.

When comparing the answers given by Chat GPT and the scores obtained, it was found that there was a significant difference between the types of tests, with the highest score in the radiology field ($F=98.20$, $p<0.001$).

Publisher: SARC Publisher

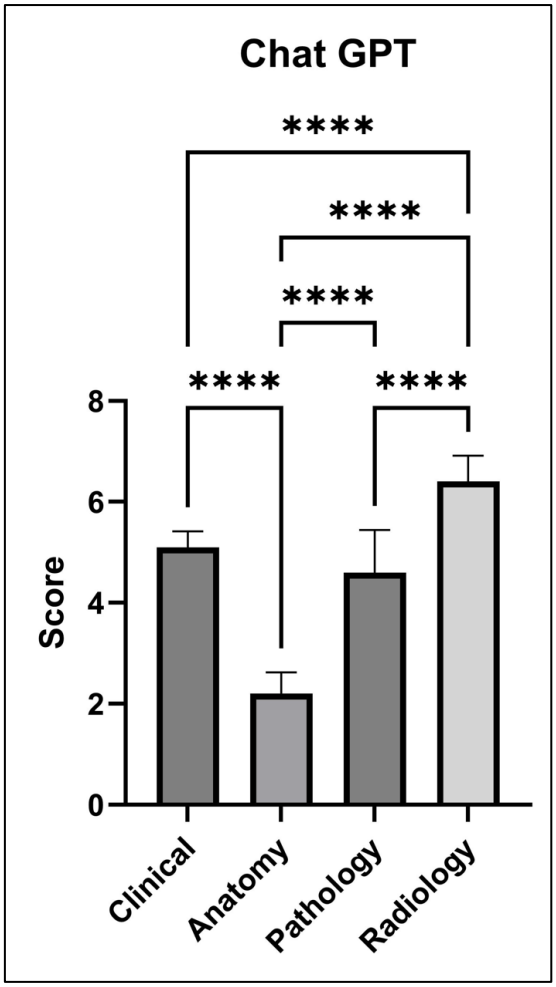

Figure 2 The comparison of the answers given by Chat GPT.

DISCUSSION

In our study, we compared the responses of Gemini and Chat GPT to neurosurgery exam questions and found that Chat GPT performed better, especially in assessing anatomical and radiological topics. Although GPT 4 and its previous versions, GPT-2 and GPT-3, have been shown to be suitable for use in many international examinations, several studies have shown that they are not sufficiently reliable. In our model, the Chat GPT has been shown to play a helpful role for neurosurgeons in the areas of anatomy and radiology.

The use of AI chatbots, such as ChatGPT and Gemini (formerly Bard), in exam contexts has been the subject of considerable research, particularly in the medical and educational domains. The performance of these AI models in answering exam questions revealed both their potential and their limitations. For example, Gemini demonstrated an accuracy rate of 62.4% in answering ophthalmology board exam questions, with notable differences in performance across different subject areas (Botross, M. *et al*., 2024). Similarly, ChatGPT has been evaluated in a variety of examination contexts, demonstrating proficiency with multiple-choice questions, but exhibiting challenges in responding to more complex, open-ended questions (GÖKTAŞ, L. S, 2023). In a comparative study, Claude outperformed both ChatGPT and Gemini on Polish medical exams, suggesting that different AI models exhibit varying degrees of proficiency depending on exam type and language (Wójcik, D. *et al*., 2024). In contrast, ChatGPT's performance in the European Board of Hand Surgery exam was less successful, with a correct response rate of only 54% over multiple attempts (Traore, S. Y. *et al*., 2023). In our study, the comparison between the answers given by ChatGPT and the scores obtained showed that there was a significant difference between the types of exams, with the highest score being in the field of radiology.

ChatGPT has been used in educational contexts to facilitate personalized learning experiences and to assist with problem-solving tasks. However, it is

Publisher: SARC Publisher

not without limitations, including a lack of comprehensive understanding and occasional inaccuracies (Suharmawan, W, 2023). In the context of tourism education, ChatGPT has demonstrated proficiency in answering multiple-choice questions, but has shown less reliability in answering open-ended tasks. This suggests that it may be particularly suited to specific examination formats (GÖKTAŞ, L. S. *et al*., 2023). ChatGPT demonstrated proficiency in answering questions that required lower-level thinking, but showed limitations in answering questions that required higher-level cognitive processes, particularly those involving picture descriptions and complex calculations. The accuracy rate was 69%, indicating the potential for ChatGPT to serve as a supplemental training tool (Krishna, S. *et al*., 2024). GPT-4 exceeded the established passing threshold and showed superior performance compared to ChatGPT (GPT-3.5) with an accuracy rate of 83.4%. It demonstrated superior performance on complex problem-solving tasks compared to its predecessor (Ali, R. *et al*., 2023). The recommendations provided by ChatGPT for glioma management were judged to be of poor quality in terms of diagnosis, but of value in terms of treatment and regimen recommendations. It has shown potential as a complementary tool, especially in resource-limited settings. However, it is not yet accurate enough to replace expert opinion (Haemmerli, J. *et al*., 2023). Treatment recommendations provided by ChatGPT were found to be inconsistent, highlighting the difficulties in replicating precise treatment plans. Performance varied between overall and case-specific responses, indicating the need for further improvements (Aghamaliyev, U. *et al*., 2024). In this study, when the relationship between test results and subgroups was examined, it was found that Gemini and ChatGPT 4.0 did not show any significant difference in clinical and pathologic evaluation, but ChatGPT 4.0 was found to be significantly more successful than Gemini in anatomic and radiologic evaluation.

There are some limitations of our study, the first of which is that since it is a study conducted with a single exam, the use of artificial intelligence in more comprehensive and numerous exams may increase the value of the study. In addition, many AI methods are known and studies with a high number of questions and comparing different methods are needed to determine the best and successful programme.

CONCLUSİON

In our study, it was seen that ChatGPT 4.0 was more successful than Gemini in anatomical and radiological evaluation of brain surgery questions.

REFERENCES

1. Hunter, D. J. & Holmes, C. "Where Medical Statistics Meets Artificial Intelligence." *New England Journal of Medicine* 389.13 (2023): 1211-1219.
2. Salvagno, M., Taccone, F. S. & Gerli, A. G. "Can Artificial Intelligence Help for Scientific Writing?" *Critical Care* 27.1 (2023): 75.
3. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H. & Wang, Y. "Artificial Intelligence in Healthcare: Past, Present, and Future." *Stroke and Vascular Neurology* 2.4 (2017): e000147.
4. Meyer, A., Soleman, A., Riese, J. & Streichert, T. "Comparison of ChatGPT, Gemini, and Le Chat with Physician Interpretations of Medical Laboratory Questions from an Online Health Forum." *Clinical Chemistry and Laboratory Medicine (CCLM)* (2024).
5. Rane, N., Choudhary, S. & Rane, J. "Gemini Versus ChatGPT: Applications, Performance, Architecture, Capabilities, and Implementation." *Performance, Architecture, Capabilities, and Implementation* (2024).
6. Hopkins, B. S., Nguyen, V. N., Dallas, J., Texakalidis, P., Yang, M., Renn, A., Guerra, G., Kashif, Z., Cheok, S., Zada, G. & Mack, W. J. "ChatGPT Versus the Neurosurgical Written Boards: A Comparative Analysis of Artificial Intelligence/Machine Learning Performance on Neurosurgical Board–Style Questions." *Journal of Neurosurgery* 139.3 (2023): 904-911.
7. Sahin, M. C., Sozer, A., Kuzucu, P., Turkmen, T., Sahin, M. B., Sozer, E., Tufek, O. Y., Nernekli, K., Emmez, H. & Celtikci, E. "Beyond Human in Neurosurgical Exams: ChatGPT's Success in the Turkish Neurosurgical Society Proficiency Board Exams." *Computers in Biology and Medicine* 169 (2024): 107807.
8. 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).
9. GÖKTAŞ, L. S. "ChatGPT Uzaktan Eğitim Sınavlarında Başarılı Olabilir Mi? Turizm Alanında Doğruluk ve Doğrulama Üzerine Bir

Publisher: SARC Publisher

Araştırma." *Journal of Tourism & Gastronomy Studies* 11.2 (2023): 892-905.
10. Wójcik, D., Adamiak, O., Czerepak, G., Tokarczuk, O. & Szalewski, L. "A Comparative Analysis of the Performance of ChatGPT4, Gemini, and Claude for the Polish Medical Final Diploma Exam and Medical-Dental Verification Exam." *medRxiv* (2024).
11. Traore, S. Y., Liverneaux, P. A., Goetsch, T., Muller, B. & Dabbagh, A. "ChatGPT Est-il en Mesure de Passer la Première Partie de l'Examen du Diplôme de l'European Board of Hand Surgery?" *Hand Surgery and Rehabilitation* 42.6 (2023): 595.
12. Suharmawan, W. "Pemanfaatan Chat GPT dalam Dunia Pendidikan." *Education Journal: Journal of Educational Research and Development* 7.2 (2023): 158-166.
13. Krishna, S., Bhambra, N., Bleakney, R. & Bhayana, R. "Evaluation of Reliability, Repeatability, Robustness, and Confidence of GPT-3.5 and GPT-4 on a Radiology Board–Style Examination." *Radiology* 311.2 (2024): e232715.
14. Ali, R., Tang, O. Y., Connolly, I. D., Zadnik Sullivan, P. L., Shin, J. H., Fridley, J. S., Asaad, W. F., Cielo, D., Oyelese, A. A., Doberstein, C. E. & Gokaslan, Z. L. "Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations." *Neurosurgery* 93.6 (2023): 1353-1365.
15. Haemmerli, J., Sveikata, L., Nouri, A., May, A., Egervari, K., Freyschlag, C., Lobrinus, J. A., Migliorini, D., Momjian, S., Sanda, N. & Schaller, K. "ChatGPT in Glioma Adjuvant Therapy Decision Making: Ready to Assume the Role of a Doctor in the Tumour Board?" *BMJ Health & Care Informatics* 30.1 (2023).
16. Aghamaliyev, U., Karimbayli, J., Giessen-Jung, C., Ilmer, M., Unger, K., Andrade, D., Hofmann, F. O., Weniger, M., Angele, M. K., Westphalen, C. B. & Werner, J. "ChatGPT's Gastrointestinal Tumor Board Tango: A Limping Dance Partner?" *European Journal of Cancer* 205 (2024): 114100.

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

Cite this article as:
Mutlucan,U.O., Zortuk, O., Bedel, C., Selvi, F. and Türk, C.C. "Comparison of Chat GPT and Gemini in Neurosurgical Evaluation Questions." *Sarcouncil Journal of Medicine and Surgery* 3.12 (2024): pp 10-15.

Publisher: SARC Publisher