Jurnal Ilmiah
Teknologi Sistem
Informasi

Vol. 5 (2024) No. 4

https://jurnal-itsi.org/index.php/jitsi; E-ISSN: 2722-4600; ISSN: 2722-4619
DOI: 10.62527/jitsi.5.4.317

Performance of Generative Pre-Trained Transformers (GPT) in Answering Questions on Anatomy in The Turkish Dentistry Specialization Exam

Arif Keskin [#], Tayfun Aygün [#]

[#] Department of Anatomy, Giresun University Faculty of Medicine, Giresun, 28200, Türkiye
E-mail: arkeskin[at]gmail.com,tayfun.aygun[at]giresun.edu.tr

ABSTRACTS

Artificial intelligence based programs are used in various fields of daily life, often without our awareness. With the increasing integration of artificial intelligence applications into the educational system, accessing information has become faster. As a result, chat programs that produce text-based answers similar to those of humans are being used as educational tools. The accuracy of the content generated by these programs has always been a topic of interest. In our study, we evaluated the success of ChatGPT, Gemini, and Copilot applications in answering dental specialty exam anatomy questions from 2012-2021. In the computer environment, free versions of ChatGpt-4, Google Gemini and Microsoft Copilot were accessed. The responses were recorded as either correct or incorrect. Out of 74 anatomy questions ChatGPT, Gemini and Copilot gave 2, 10, and 1 incorrect answers, respectively. Although the evaluated programs showed sufficient success in answering anatomy questions, their use was deemed limited due to errors in the supplementary information they provided.

Manuscript received Nov 17, 2024; revised Nov 25, 2024. accepted Nov 27, 2024 Date of publication Dec 30, 2024. International Journal, JITSI : Jurnal Ilmiah Teknologi Sistem Informasi licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Keywords / Kata Kunci — ***Anatomy; Artificial intelligence; Chatgpt; Copilot; Dentistry; Gemini***

CORRESPONDING AUTHOR

Arif Keskin
Department of Anatomy, Giresun University Faculty of Medicine, Giresun, 28200, Türkiye
Email: arkeskin[at]gmail.com

1. INTRODUCTION

The concept of artificial intelligence (AI) was first introduced in 1956 by mathematician John McCarthy. AI applications, which have evolved alongside computer programming, are now utilized in diverse fields ranging from education to commerce. The quest for alternative methods of information access in education has been ongoing. Following the advent of libraries, the internet, and electronic resources, AI applications that generate human-like text have gained preference [1,2].

ChatGPT, developed by OpenAI in 2022, is an advanced AI-based chat program capable of answering questions with high accuracy and generating human-like textual responses using its 175 billion parameters. Its quick and intelligent responses have made it one of the most widely used educational tools [3,4[.

Similarly, Google's AI-based language model, Google Gemini, and Microsoft's Copilot application are prominent programs that generate human-like textual responses. These programs are preferred as large language model (LLM)-based chat programs [5,6].

By evaluating the contribution of rapidly advancing technology to education, challenges faced in the educational process can be addressed more swiftly [1]. Existing literature has assessed the success of AI-based programs in solving questions across various medical fields [5,7,8]. However, there is a lack of studies evaluating their success in answering anatomy questions from central examinations. Therefore, this study aims to evaluate the effectiveness of ChatGPT (OpenAI), Gemini (Google), and Copilot (Microsoft) in solving anatomy questions from the Dentistry Specialty Entrance Exam (DUS), thereby assessing the suitability of using AI-based chat programs in medical education

2. Research Methodology

As this study did not involve experiments or surveys on humans requiring ethics committee approval, no ethical clearance was obtained. The anatomy questions from the Dentistry Specialty Entrance Exam (DUS), which are publicly available on the official website of the Assessment, Selection, and Placement Center (ÖSYM) (https://www.osym.gov.tr/TR15070/dus-cikmis-sorular.html), were utilized. A total of 78 multiple-choice anatomy questions from 13 DUS exams held between 2012 and 2021 were compiled in Word format. Access to the free versions of ChatGPT-4, Google Gemini, and Microsoft Copilot was secured online. Before posing the questions, the following text was entered into the text field of the programs: "I am a faculty member in the Department of Anatomy at a medical school. I will ask you some questions, and I would like you to provide the correct answer for each." The responses were recorded as either correct or incorrect. The answers provided by the programs were compared with the answer key available on the ÖSYM website. For the questions that were answered incorrectly, the following prompt was sent: "Please answer this question again. I don't think your answer is correct." This process was repeated until the correct answer was provided. The number of times incorrect answers were corrected was recorded.

The four questions containing images or diagrams were excluded from the study. To prevent the programs from retaining memory between sessions, a new session was initiated for each question. The questions were posed to the programs on September 30, 2024.

Statistical Analysis

For statistical analysis, IBM SPSS 23 software was utilized. Descriptive statistics were presented as mean ± standard deviation for continuous variables and frequency (percentage) for categorical variables. Chi-square and Fisher's exact tests were employed to compare the proportions of two independent groups. A value of $p<0.05$ was considered statistically significant

3. Results And Discussion

A total of 74 anatomy questions were used to assess the performance of the AI-based chat programs. The number of correct and incorrect answers, along with the correct answer percentages for each program, is presented in Table 1.

Table 1. Number and percentage of programs giving

	ChatGpt	Gemini	Copilot	
Correct (%)	72 (97,3)	64 (86,49)	73 (98,65)	0,003*
Incorrect (%)	2 (2,7)	10 (13,51)	1 (1,35)	

ChatGPT answered two questions related to the nervous system incorrectly. Upon reconsideration, it corrected these errors. Gemini provided incorrect answers to 10 questions on muscle, bone, peripheral nerve, autonomic nerve, and cranial nerve topics. When re-prompted, Gemini corrected 4 of these but failed to provide the correct answers for the remaining 5. For one peripheral nerve question, it indicated an error but failed to offer a response. Copilot answered one question on muscle anatomy incorrectly but corrected its response when asked to reconsider.

In addition to providing the correct answer options, ChatGPT, Gemini, and Copilot also offered explanatory information about the anatomical structures mentioned in the questions. Both ChatGPT and Copilot provided satisfactory supplementary details to support their answers, whereas Gemini stood out as the program offering the most visual and reference-based information.

For one anatomy question, both Gemini and Copilot provided the same incorrect answer. Similarly, for another question, both ChatGPT and Gemini made the same error. Figure 1 shows the number of incorrect responses from each program for the exam years.

The Dental Specialty Examination (DUS) is a central exam that dental professionals must pass to begin their specialty training after graduation. This exam contains questions aimed at measuring medical knowledge. Due to the limited number of specialty positions at universities, the number of candidates taking the exam increases each year. Candidates often use a variety of educational tools, both printed and electronic, to prepare for the exam. Among these tools, AI-based programs serve as supplementary educational resources [9-12].

Anatomy is always at the forefront of health education. Anatomy, by its nature, is a discipline that is easily forgotten and requires regular review [13]. It contains extensive information due to its complex terminology and anatomical variations. Therefore, candidates studying exam-oriented anatomy need fast-response tools to access information. Technological advances have accelerated the development of LLMs. As a result, AI-based chat programs have emerged, producing human-like text to deliver the desired information quickly and accurately. Verifying the accuracy of the information provided by these programs increases trust [10].

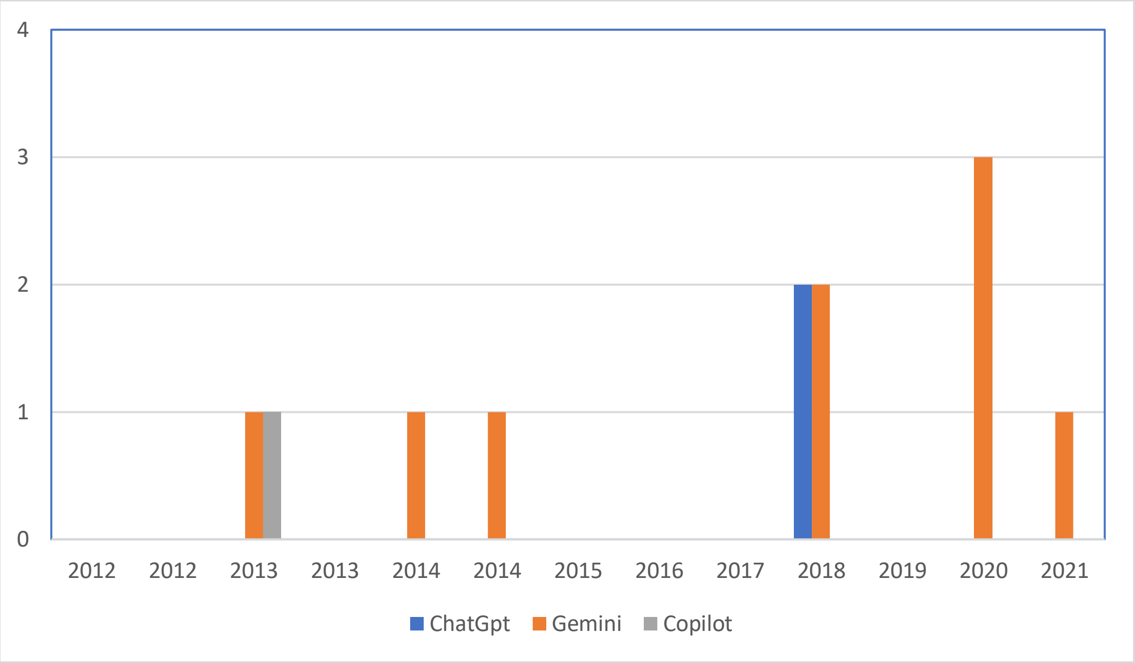

FIG 1. Number of wrong answers given by programs according to exam year

Initially not designed for the medical field, AI models have increasingly contributed to both healthcare systems and health education [14]. A study that used an AI-based chat program demonstrated that it scored higher than the candidates who took the 2021 Medical Specialty Examination (TUS), showing that it could be reliably used in medical education [5]. However, a study evaluating the responses of chat programs to anatomy questions showed that while they answered questions correctly at a high rate, the supplementary information they provided contained errors [15]. In another study, AI-based applications were evaluated for their use in anatomy education, and the results showed that the programs gave incorrect information to students due to errors in the answers to anatomy-related questions [16].In our study, we asked ChatGPT, Gemini, and Copilot to answer 74 anatomy questions from the DUS exams held between 2012 and 2021. ChatGPT answered 72 questions correctly, Gemini 64, and Copilot 73.

ChatGPT made two errors on questions from the 2018 DUS exam. One of these questions, a clinical question about anatomical structures in the neck, was initially answered incorrectly. After being prompted to reconsider its response, ChatGPT provided the correct answer. The second question, related to the neural innervation of the salivary glands, was answered incorrectly twice before ChatGPT provided the correct response. While the supplementary information provided by ChatGPT was accurate, it was unexpected that the program initially gave incorrect answers to these questions. Given that ChatGPT correctly answered 97.3% of the questions, we consider it successful in solving anatomy questions.

Gemini did not perform as well as ChatGPT in answering the same questions, with 13.51% of its responses being incorrect. Gemini answered half of the questions from the 2020 DUS exam incorrectly. However, the supplementary information it provided was more detailed and explanatory compared to the other programs. By supporting its answers with reference materials and visuals, Gemini distinguished itself in this regard. Nevertheless, for the 10 questions it answered incorrectly, it managed to correct four after re-prompting, but failed to provide correct answers for the remaining six. One of these questions, related to the neural innervation of the palate muscles, remained unanswered, and the explanatory information was partially incorrect. Therefore, while Gemini provided some success in answering anatomy questions, its 13.51% error rate indicates that its overall performance was limited. Compared to the other programs, there was a significant difference in responses between both Gemini-Copilot and Gemini-ChatGPT ($p<0.005$).

Microsoft's Copilot provided nearly 99% correct answers to the questions. For a question about the function of the tongue muscles from the 2013 DUS exam, which Gemini failed to answer correctly, Copilot initially gave an incorrect answer but corrected it after being asked to reconsider. With its speed and accuracy in generating responses, and supplementary information similar to that provided by ChatGPT, Copilot was considered the most successful program. There was no significant difference between the success of ChatGPT and Copilot ($p>0.005$). However, there was a significant difference between Gemini and Copilot and Gemini and ChatGPT ($p<0.005$).

Our study, along with other literature, demonstrates that AI-based chat programs are not 100% accurate in solving questions and providing supplementary information. Some studies have shown that these programs perform better than humans in question-solving, while others indicate that they are less successful [5,15,17,18]. Therefore, it is important to verify the content generated by AI-based chat programs, as they have limited use in learning anatomy

4. Conclusions

Copilot, ChatGPT, and Gemini are considered suitable tools for students and educators, given their high accuracy in answering complex anatomy questions despite the clinical content and complicated terminology. However, due to inaccuracies in the supplementary information provided by these programs, their use in self-directed learning is recommended with caution.

References

[1] Arslan, K.. Eğitimde Yapay Zekâ Ve Uygulamaları. Batı Anadolu Eğitim Bilimleri Dergisi,; 11, 71-88. (2020)

[2] Büyükada, S.. Akademik Yazımda Yapay Zekâ Kullanımının Etik Açıdan İncelenmesi: Chatgpt Örneği. Rize İlahiyat Dergisi,; 1-12. (2024)

[3] Gilson, A., Safranek, C., Huang, T., Socrates, V., Chi, L., Taylor, R. A. & Chartash, D. 2022. How Does Chatgpt Perform On The Medical Licensing Exams? The Implications Of Large Language Models For Medical Education And Knowledge Assessment. Medrxiv,.12. 23.22283901. (2022)

[4] Yağar, S. D. Chatgpt'nin Sağlık Alanındaki Potansiyel Kullanımına Ilişkin Çıkarımlar. Table LegendsAygul, Y, Olucoglu, M And Alpkocak, A. Tipta Uzmanlik Sinavinda (Tus) Büyük Dil Modelleri Insanlardan Daha Mi Başarili Arxiv E-Prints, Arxiv: 2408.12305. (2024)

[5] Şensoy, E. & Çıtırık, M. Okülofasiyal Plastik Ve Orbital Cerrahide İngilizce Ve Türkçe Dil Çeşitliliğinin Yapay Zeka Chatbot Performansına Etkisi: Chatgpt-3.5, Copilot Ve Gemini Üzerine Bir Çalışma. Osmangazi Tıp Dergisi,;46, 781-786. (2024)

[6] Flores-Cohaila Ja, García-Vicente A, Vizcarra-Jiménez Sf, De La Cruz-Galán Jp, Gutiérrez-Arratia Jd, Torres Bgq And Taype-Rondan A. Performance Of Chatgpt On The Peruvian National Licensing Medical Examination: Cross-Sectional Study. Jmir Medical Education,;9, E48039. (2023)

[7] Wang, X., Gong, Z., Wang, G., Jia, J., Xu, Y., Zhao, J., Fan, Q., Wu, S., Hu, W. & Li, X. Chatgpt Performs On The Chinese National Medical Licensing Examination. Journal Of Medical Systems,; 47, 86. (2023)

[8] Çulhaoğlu, AK, Kiliçarslan, MA and Deniz, KZ. Diş Hekimliğinde Uzmanlik Sinavinin Farkli Eğitim Seviyelerdeki Algi Ve Tercih Durumlarinin Değerlendirilmesi. Atatürk Üniversitesi Diş Hekimliği Fakültesi Dergisi,;31, 420-426. (2021)

[9] İncemen, S. & Öztürk, G. Farklı Eğitim Alanlarında Yapay Zekâ: Uygulama Örnekleri. International Journal Of Computers In Education. 7, 27-49. (2024)

[10] Totlis, T., Natsis, K., Filos, D., Ediaroglou, V., Mantzou, N., Duparc, F. & Piagkou, M. The Potential Role Of Chatgpt And Artificial Intelligence In Anatomy Education: A Conversation With Chatgpt. Surgical And Radiologic Anatomy. 45, 1321-1329. (2023)

[11] Uzun, Y., Tümtürk, A. Y. & Öztürk, H. Günümüzde Ve Gelecekte Eğitim Alanında Kullanılan Yapay Zeka. 1st International Conference On Applied Engineering And Natural Sciences,. 1-3. (2021)

[12] Uzun, Arpacı, M. F., Kaya, A. & Aydın, M. Öğrencilerin Anatomi Dersine Ayırdıkları Zaman Ile Anatomi Dersi Hakkındaki Görüşlerinin Ilişkisinin Değerlendirilmesi. Journal Of Medical Topics And Updates.; 1, 20-27. (2022)

[13] Leng, L. Challenge, Integration, And Change: Chatgpt And Future Anatomical Education. Medical Education Online,;29, 2304973. (2024)

[14] Mantzou, N., Ediaroglou, V., Drakonaki, E., Syggelos, S. A., Karageorgos, F. F. & Totlis, T. Chatgpt Efficacy For Answering Musculoskeletal Anatomy Questions: A Study Evaluating Quality And Consistency Between Raters And Timepoints. Surgical And Radiologic Anatomy,;1-6. (2024)

[15] Mogali, S. R. Initial Impressions Of Chatgpt For Anatomy Education. Anatomical Sciences Education,;7, 444-447. (2024)

[16] Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A. & Chartash, D. How Does Chatgpt Perform On The United States Medical Licensing Examination (Usmle)? The Implications Of Large Language Models For Medical Education And Knowledge Assessment. Jmir Medical Education,;9, E45312. (2023)

[17] Meral, G., Ateş, S., Günay, S., Öztürk, A. & Kuşdoğan, M. Comparative Analysis Of Chatgpt, Gemini And Emergency Medicine Specialist In Esi Triage Assessment. The American Journal Of Emergency Medicine,;81, 146-150. (2024)