Uludağ Üniversitesi Tıp Fakültesi Dergisi
Journal of Uludağ University Medical Faculty
50 (3) 443-447, 2024
DOI: https://doi.org/10.32708/uutfd.1543370

ORIGINAL RESEARCH

A Comparative Analysis of GPT-3.5, GPT-4 and GPT-4.o in Heart Failure

Şeyda GÜNAY-POLATKAN[1], Deniz SIĞIRLI[2]

[1] Bursa Uludag University Faculty of Medicine, Department of Cardiology, Bursa, Türkiye.
[2] Bursa Uludag University Faculty of Medicine, Department of Biostatistics, Bursa, Türkiye.

ABSTRACT

Digitalization have increasingly penetrated in healthcare. Generative artificial intelligence (AI) is a type of AI technology that can generate new content. Patients can use AI-powered chatbots to get medical information. Heart failure is a syndrome with high morbidity and mortality. Patients search about heart failure in many web sites commonly. This study aimed to assess Large Language Models (LLMs) - ChatGPT 3.5, GPT-4 and GPT-4.o- in terms of their accuracy in answering the questions about heart failure (HF). Thirteen questions regarding to the definition, causes, signs and symptoms, complications, treatment and lifestyle recommendations of the HF were evaluated. These questions to assess the knowledge and awareness of medical students about heart failure were taken from a previous study in literature. Of the students who participated in this study, 158 (58.7%) were first-year students, while 111 (41.3%) were sixth-year students and were taking their cardiology internship in their fourth year. The questions were entered in Turkish language and 2 cardiologists with over ten years of experience evaluated the responses generated by different models including GPT-3.5, GPT-4 and GPT-4.o. ChatGPT-3.5 yielded “correct” responses to 8/13 (61.5%) of the questions whereas, GPT-4 yielded “correct” responses to 11/13 (84.6%) of the questions. All of the responses of GPT-4.o were accurate and complete. Performance of medical students did not include 100% correct answers for any question. This study revealed that performance of GPT-4.o was superior to GPT-3.5, but similar with GPT-4

Keywords: Artificial intelligence. Heart failure. Medical knowledge.

Kalp Yetersizliğinde GPT-3,5, GPT-4 ve GPT-4.o Performansının Karşılaştırılması

ÖZET

Dijitalleşme sağlık hizmetleri alanında giderek daha fazla yer almaktadır. Üretken yapay zeka yeni içerik üretebilen bir yapay zeka teknolojisi türüdür. Hastalar tıbbi bilgi almak için yapay zeka destekli sohbet robotlarını kullanabilmektedir. Kalp yetersizliği, yüksek morbidite ve mortaliteye sahip bir sendromdur. Hastalar genellikle birçok web sitesinde kalp yetersizliği hakkında arama yapmaktadır. Bu çalışma, kalp yetersizliği hakkındaki soruları yanıtlamadaki doğrulukları açısından Büyük Dil Modelleri (LLM'ler) - ChatGPT 3.5, GPT-4 ve GPT-4.o'yu karşılaştırmayı amaçlamaktadır. Çalışmada kalp yetersizliğinin tanımı, nedenleri, belirti ve semptomları, komplikasyonları, tedavisi ve yaşam tarzı önerileriyle ilgili on üç soru soruldu. Bu sorular, tıp fakültesi öğrencilerinin kalp yetmezliği hakkındaki bilgi ve farkındalığını değerlendirmek için yapılan önceki bir çalışmadan alındı. Bu çalışmaya katılmış olan öğrencilerin 158 tanesi (%58,7) 1. Sınıf öğrencisi iken, 111 tanesi (%41,3) 6. Sınıf öğrencisiydi ve kardiyoloji stajı 4. sınıfta alınmaktaydı. Sorular yapay zeka destekli modellere Türkçe dilinde soruldu ve on yılı aşkın deneyime sahip 2 kardiyolog, GPT-3.5, GPT-4 ve GPT-4.o tarafından üretilen yanıtları değerlendirdi. ChatGPT-3.5 soruların 8/13'üne (61.5%) "doğru" yanıt verirken, GPT-4 soruların 11/13'üne (84.6%) "doğru" yanıt verdi. GPT-4.o'nun tüm yanıtları doğru ve eksiksizdi. Tıp fakültesi öğrencilerinin performansı hiçbir soru için %100 doğru yanıt içermiyordu. Bu çalışma GPT-4.o' nun performansının GPT-3.5'ten üstün olduğunu ancak GPT-4 ile benzer olduğunu ortaya koydu.

Anahtar Kelimeler: Yapay zeka. Kalp yetersizliği. Tıbbi bilgi.

Date Received: September 04, 2024
Date Accepted: November 018, 2024

Dr. Şeyda GÜNAY-POLATKAN
Bursa Uludag University Faculty of Medicine,
Department of Cardiology,
Bursa, Türkiye.
Phone: 0505 278 07 77
E-mail: seydagunay@uludag.edu.tr.

Authors' ORCID Information:
Şeyda GÜNAY-POLATKAN: 0000-0003-0012-345X
Deniz SIĞIRLI: 0000-0002-4006-3263

Heart failure (HF) can be defined as a condition in which the heart cannot pump enough blood to meet the body's needs or an abnormality of cardiac structure or function leading to failure of the heart to deliver oxygen at a rate commensurate with the requirements of the metabolizing tissues[1-3]. HF poses a significant health burden by causing recurrent and frequent hospitalizations, deterioration in quality of life, increase in health care costs and premature deaths[4]. With an estimated prevalence of more than 56 million individuals worldwide, one in five people aged 40 and over is expected to develop heart failure[5,6]. If knowledge and awareness levels are not sufficient, the symptoms of heart failure may be noticed late and there may be delays in starting the appropriate treatment. In most patients, diagnosis is made when symptoms are advanced[7].

Advances in digitalization have penetrated almost every aspect of daily life, including healthcare, ensuring people to easily access a range of digital tools and platforms that enable them to access information about diseases and medical issues. Health information is an increasingly accessible topic to the more than 3.2 billion people who have access to the internet worldwide[8]. The current decade is witnessing the emergence of generative artificial intelligence (AI), a type of AI technology that can generate new content. With the revolution in digital data processing processes, generative AI is now becoming effective in every area, and the scientific arena is witnessing a new innovation. However, two-thirds of people find this information unreliable and one-third report confusion after the search[9,10].

AI-powered chatbots are sophisticated systems which are designed to mimic human conversation using text or voice interaction, providing information in a conversational manner, are part of a fresh ware of generative AI. Generative Pre-trained Transformer (GPT) is a deep learning model that is pre-trained on the unlabeled text data and can be used to enable specific tasks specifically like language generation, language modelling, and text completion[11-13]. Models are systems based on statistical models that are used to construct a probability distribution function that assigns a probability to every string in the language and predict the likelihood of a string, word, or group of words[14,15]. Large language models are language models pre-trained on large amounts of text with bulk parameter sizes, making them sensitive to minor changes in input[15-17]. The tech company OpenAI launched ChatPT using GPT-3.5 which is a general-purpose chatbot based on large language models late in 2022[18,19]. OpenAI recently introduced GPT-4 on March 2023[20] and GPT-4.o on May 2024[21] that can process image inputs as well as text input.

Results of validation analysis of GPT-3.5 on numerous medical examinations have been published[22-27]. GPT-3.5 and GPT-4 were already validated on, to our knowledge, several national medical tests like the United States Medical Licensing Examination (USMLE)[24] and Chinese National Medical Licensing Examinations[25,26].

There are also studies in the literature reporting the insufficiency of the awareness and knowledge levels of the public about heart failure in many countries[29-31]. Those who want to have sufficient information about heart failure will resort to artificial intelligence applications more and more in time. The studies investigating whether the information people can obtain about heart failure using artificial intelligence applications is accurate and complete are limited[32,33]. In this study, we aimed to assess Large Language Models (LLMs) -ChatGPT 3.5, GPT-4 and GPT-4.o in terms of their accuracy in answering the questions about heart failure (HF).

Material and Method

Thirteen questions regarding to the definition, causes, signs and symptoms, complications, treatment and lifestyle recommendations of the HF were evaluated. These questions were taken from a study conducted by Gunay-Polatkan et al.[34] to assess the knowledge and awareness of medical students about heart failure. The previous study included 269 students, 158 (58.7%) of them were first-year medical students and 111 (41.3%) of them were in their final year of medical education. Cardiology internship is performed in the fourth year of education. The answers obtained from different artificial intelligence applications were compared with each other. This survey included 5 multiple choice questions with only one correct answer about the definition and epidemiology of heart failure, 1 question with 2 options (yes, no) questioning 9 etiological causes,1 question with 2 options (yes, no) questioning 10 heart failure symptoms, 1 question with 2 options (yes, no) questioning 9 heart failure complications, 1 question with 2 options (yes, no) questioning 6 heart failure treatment methods, 1 question with 2 options (yes, no) was asked to evaluate 4 preventive methods and 3 multiple choice questions with only one correct answer were asked to evaluate knowledge and awareness about lifestyle recommendations in heart failure. The questions were entered in Turkish language and 2 cardiologists with over ten years of experience evaluated the responses generated by each model. If the answers given by artificial intelligence applications were correct but incomplete, they were categorized as "partially correct". In addition, the accuracy rate of the answers given by humans was calculated.

Statistical analysis

The Fisher–Freeman–Halton and Fisher exact chi-square tests were used to compare the response rates between different language models including ChatGPT-3.5, GPT-4 and GPT-4.o. Categorical variables were presented as percentage (%). Significance level was accepted as 0.05 for two sided hypothesis test. Statistical analyses were performed using the IBM SPSS Statistics package program (IBM SPSS Statistics for Windows, Version 28.0. Armonk, NY: IBM Corp.).

Results

ChatGPT-3.5 yielded “correct” responses to 8/13 (61.5%) of the questions whereas, rates of “partially correct” responses were 5/13 (38.5%). GPT-4 yielded “correct” responses to 11/13 (84.6%) of the questions whereas, rates of “partially correct” responses were 2/11 (15.4%). All of the responses of GPT-4.o were accurate and complete. (Table I). When we compared the responses of three models, we found statistically significant difference between them (p=0.048). also in pair wise comparisons there was a significant difference between ChatGPT-3.5 and GPT-4o (p=0.039). But the differences between ChatGPT-3.5 and GPT-4 (p=0.378), GPT-4 and GPT-4o (p=0.480) were not statistically significant (Table II). Human performance did not include 100% correct answers for any question. The correct answer percentages of medical students are shown in Figure 1.

Table I. Responses of langage models and human performance.

Questions		ChatGPT	GPT-4	GPT-4.o
Q1	Which of the following diseases is the most deadly?	partially correct	partially correct	correct
Q2	Which of the following is the most common cause of repeated hospitalizations?	partially correct	correct	correct
Q3	Which of the following could be a symptom of heart failure? (more than one option can be selected)	correct	correct	correct
Q4	Which of the following may cause heart failure? (more than one option can be selected)	correct	correct	correct
Q5	Which of the following may develop due to heart failure? (more than one option may be selected)	partially correct	correct	correct
Q6	Which of the following should a patient whose heart failure symptoms are under control with treatment pay attention to in order to prevent decompensation? (more than one option can be selected)	correct	correct	correct
Q7	Which of the following is a treatment option of heart failure? (more than one option can be selected)	partially correct	correct	correct
Q8	How much salt should a heart failure patient consume daily?	correct	correct	correct
Q9	How much fluid should a patient with heart failure consume daily?	correct	correct	correct
Q10	Which of the following statements is correct regarding whether a heart failure patient can perform exercise?	correct	correct	correct
Q11	Which of the following statements is the best definition of heart failure?	correct	partially correct	correct
Q12	May heart failure develop at any age?	correct	correct	correct
Q13	What is the lifelong risk of developing heart failure in a healthy individual?	partially correct	correct	correct

Table II. Comparison of responses between language models

	ChatGPT-3.5 (n,%)	GPT-4 (n,%)	GPT-4.o (n,%)	p value
Correct	8 (61.5)	11 (84.6)	13 (100)	0.048
Partial correct	5 (38.5)	2 (15.4)	0 (0)	

ChatGPT-3.5 vs GPT-4: p=0.378
ChatGPT-3.5 vs GPT-4o: p=0.039
GPT-4 vs GPT-4o: p=0.480

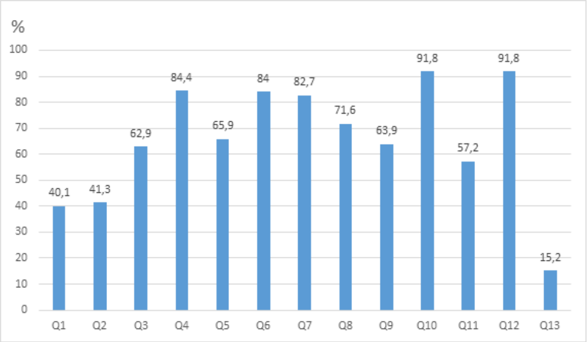

Figure 1.
Correct response percentages of human performance

Discussion and Conclusion

In our study, GPT-4.o consistently outperformed GPT-3.5 and GPT-4 in terms of the number of correct answers. Statistically, performance of GPT-4.o was superior to GPT-3.5, but similar with GPT-4. Medical knowledge represented by the GPT-4.o model indicates an improvement compared to the previous versions. Our results obtained in Turkish language are in line with other studies conducted on different tests and languages which indicated the improvement of the leverage of the medical knowledge from the training dataset by GPT LLMs alongside with the development of the consecutive versions[28].

Studies investigating AI-provided answers to questions about heart failure are available in the literature. Dimitriadis et al. reported that ChatGPT was able to adequately answer all questions posed to it[32]. Also, King et al. reported that GPT-3.5 and GPT-4 answered the majority of heart failure-related questions accurately and reliably[33]. There are increasing numbers of studies evaluating the performance of LLMs in other medical conditions. For instance, a previous study by Gencer et al. showed that ChatGPT passed the thoracic surgery exam[22]. In a research letter, the performance of medical students and chatbots were compared on free-response clinical reasoning examinations and it was shown that the study model scored more than students[23].

Previous studies evaluating GPT performance have also investigated whether similar success was achieved in languages other than English[25,35,36]. To the best of our knowledge, our study is the only one conducted in Turkish language. Although, the ChatGPT had a lower success rate for exams in which the question language was not English[35,37], in our study GPT-4.o version responded all Turkish questions correctly. There may be several potential reasons for the imperfect performance and providing incorrect answers by the tested models. First of all, both models are general-purpose LLMs that are capable of answering questions from various fields and are not dedicated to medical applications. Also, since how the question is asked determines the answer, Chatbot's responses can be sensitive to rewording of prompt. In spite of this, a study has shown that chatbot responses were preferred over physician responses on a social media forum[38].

While the results of this study demonstrated the potential utility of AI language models in the medical field, important limitations should be acknowledged. First of all, the study focused solely on a Turkish questionnaire which was previously generated for a study about heart failure including a small sample size of medical students. This limits the generalizability of the findings to other medical issues or languages. Additionally, for the studies conducted with multiple choice questions including more than one correct answer, it should be kept in mind that it may be suboptimal to evaluate the accuracy of the answers given by the AI. It may be more appropriate to use questions with an only single correct answer. Another limitation of our study is that since people who are not health professionals were not included in the study, a comparison was not made between medical school students and people who did not receive education in the field of health.

Such powerful tools might have a considerable impact on the shape of the public health and medical education. In the future, LLMs may also provide decision-making recommendations on a detailed defined problem beyond only the presentation of current information. In conclusion, this study highlights the advances in AI language models' performance on medical information. Study results revealed that the performance of GPT-4.o was superior to GPT-3.5, but similar with GPT-4. Future research should focus on exploring their potential applications in medical decision making and public health education.

Ethics Committee Approval Information:
This research was conducted by asking questions to artificial intelligence models. There are no human participants. As in similar studies in the literature, this research does not require ethics committee approval.

Researcher Contribution Statement:
Idea and design: Ş.G.P., D.S.; Data collection and processing: Ş.G.P., D.S.; Analysis and interpretation of data: D.S.; Writing of significant parts of the article: Ş.G.P.

Support and Acknowledgement Statement:
This study received no financial support.

Conflict of Interest Statement:
The authors of the article have no conflict of interest declarations.

References

1- Braunwald E., Heart Failure, Journal of the American College of Cardiology: Heart Failure, (2013). 1(1): 1-20.

2- Wagner S & Cohn K. Heart failure. A proposed definition and classification. Arch Intern Med. 1977; 137: 675-678.

3- Biykem B. et al. Universal Definition and Classification of Heart Failure, Journal of Cardiac Failure, (2021) 27 (4), 387-413.

4- Khan, M.S., Shahid, I., Bennis, A. et al. Global epidemiology of heart failure. Nat Rev Cardiol (2024). https://doi.org/10.1038/s41569-024-01046-6

5- GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018; 392: 1789– 1858.

6- Lloyd-Jones DM, Larson MG, Leip EP, et al. Lifetime risk for developing congestive heart failure: the Framingham Heart Study. Circulation. 2002;106(24):3068-3072.

7- Johansson S, Wallander M.A., Ruigomez A., Garcia Rodriguez L.A. Incidence of newly diagnosed heart failure in UK general practice. Eur J Heart Fail. 2001; 3 (2): 225–231.

8- ITU releases 2015 ICT figures. Statistics confirm ICT revolution of the past 15 years. http://www.itu.int/net/pressoffice/press_releases/2015/17.aspx#.

9- Torrent-Sellens J, Díaz-Chao Á, Soler-Ramos I, et al. Modelling and predicting eHealth usage in Europe: a multidimensional approach from an online survey of 13,000 european union internet users. *J Med Internet Res.* 2016;18(7):e188.

10- Klerings I, Weinhandl AS, Thaler KJ. Information overload in healthcare: too much of a good thing? Z Evid Fortbild Qual Gesundhwes. 2015;109(4-5):285-90.

11- Labadze, L., Grigolia, M., Machaidze, L. Role of AI chatbots in education: systematic literature review. *Int J Educ Technol High Educ* **20**, 56 (2023).

12- Dwivedi Y.K., et al. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy, International Journal of Information Management, 71, 2023, https://doi.org/10.1016/j.ijinfomgt.2023.102642.

13- Yenduri G. GPT (Generative Pre-Trained Transformer)— A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. IEEE Access, 12, 2024. https://doi.org/10.1109/ACCESS.2024.3389497.

14- Venkat N. Gudivada, Dhana Rao, Vijay V. Raghavan. Chapter 9 - Big Data Driven Natural Language Processing Research and Applications. Editor(s): Venu Govindaraju, Vijay V. Raghavan, C.R. Rao, Handbook of Statistics,Elsevier, 2015, Pages 203-238, https://doi.org/10.1016/B978-0-444-63492-4.00009-5.

15- Picazo-Sanchez, P., Ortiz-Martin, L. Analysing the impact of ChatGPT in research. Appl Intell (2024). 4172–4188.

16- Chiang H.H., Lee H.Y. Can Large Language Models Be an Alternative to Human Evaluations? Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pages 15607–15631, 2023.

17- Ferrara E, The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness, Machine Learning with Applications, 15, 2024, doi.org/10.1016/j.mlwa.2024.100525.

18- Saka, A., Taiwo, R., Saka, N., Salami, B., Ajayi, S., Akande, K., Kazemi, H. GPT Models in Construction Industry: Opportunities, Limitations, and a Use Case Validation. Developments in the Built Environment. 2024, 17, 1-29. https://doi.org/10.1016/j.dibe.2023.100300

19- Urbina F, Lentzos F, Invernizzi C, Ekins S. Dual Use of Artificial Intelligence-powered Drug Discovery. Nat Mach Intell. 2022 Mar;4(3):189-191.

20- GPT-4 Technical Report. OpenAI (2023). https://cdn.openai.com/papers/gpt-4.pdf

21- OpenAI. Introducing GPT-4o and more tools to ChatGPT free users. https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/.

22- Gencer A, Aydin S. Can ChatGPT pass the thoracic surgery exam? Am J Med Sci. 2023 Oct;366(4):291-295.

23- Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, Chen JH. Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations. JAMA Intern Med. 2023 Sep 1;183(9):1028-1030.

24- Beam K, Sharma P, Kumar B, Wang C, Brodsky D, Martin CR, Beam A. Performance of a Large Language Model on Practice Questions for the Neonatal Board Examination. JAMA Pediatr. 2023 Sep 1;177(9):977-979.

25- 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. J Med Syst. 2023 Aug 15;47(1):86.

26- Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, Jiang Y, Wu Y, Chen Y, Zhou J, Zhu Z, Yan Z, Yu P, Liu X. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health. 2023 Dec 1;2(12):e0000397. doi: 10.1371/journal.pdig.0000397.

27- Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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 Med Educ. 2023 Feb 8;9:e45312. doi: 10.2196/45312. Erratum in: JMIR Med Educ. 2024 Feb 27;10:e57594. doi: 10.2196/57594.

28- Kung TH, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Dig. Health.* 2023;**2**:e0000198

29- WJ, McMurray JJ, Rauch B, Zannad F, Keukelaar K, CohenSolal A, Lopez-Sendon J, Hobbs FD, Grobbee DE, Boccanelli A, Cline C, Macarie C, Dietz R, Ruzyllo W. Public awareness of heart failure in Europe: first results from SHAPE. Eur Heart J. 2005 Nov;26(22):2413-21.

30- Zelenak C, Radenovic S, Musial-Bright L, Tahirovic E, Sacirovic M, Lee CB, Jahandar-Lashki D, Inkrot S, Trippel TD, Busjahn A, Hashemi D, Wachter R, Pankuweit S, Störk S, Pieske B, Edelmann F, Düngen HD. Heart failure awareness survey in Germany: general knowledge on heart failure remains poor. ESC Heart Fail. 2017 Aug;4(3):224-231.

31- Nowak K, Stępień K, Furczyńska P, Owsianka I, Włodarczyk A, Zalewski J, Nessler J, Gackowski A. The awareness and knowledge about heart failure in Poland - lessons from the Heart Failure Awareness Day and internet surveys. Folia Med Cracov. 2019;59(2):93-109.

32- Dimitriadis F, Alkagiet S, Tsigkriki L, Kleitsioti P, Sidiropoulos G, Efstratiou D, Askalidi T, Tsaousidis A, Siarkos M, Giannakopoulou P, Mavrogianni AD, Zarifis J, Koulaouzidis G. ChatGPT and Patients With Heart Failure. Angiology. 2024 Mar 7:33197241238403.

33- King RC, Samaan JS, Yeo YH, Mody B, Lombardo DM, Ghashghaei R. Appropriateness of ChatGPT in Answering Heart Failure Related Questions. Heart Lung Circ. 2024 May 30:S1443-9506(24)00165-3. doi: 10.1016/j.hlc.2024.03.005.

34- Gunay-Polatkan S, Sigirli D, Alak C, Senturk T. Assessment of Knowledge and Awareness on Heart Failure among Medical Students. Journal of Uludag Medical Faculty.2023;49(3):305-12.

35- Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. *Capabilities of GPT-4 on Medical Challenge Problems*. (2023).

36- Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep. 2023 Nov 22;13(1):20512.

37- Oner S.K., Ocak B., Sahbat Y. Kurnaz R.Y. and Cilingir E. Performance of Chat Gpt on a Turkish Board of Orthopaedic Surgery Examination.(2024). DOI: 10.21203/rs.3.rs-4637339/v1

38- Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023 Jun 1;183(6):589-596.