ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA

DOI: 10.5336/ophthal.2024-102632

Evaluation of the Performance of Large Language Models (ChatGPT-3.5, ChatGPT-4, Bing and Bard) in Turkish Ophthalmology Chief-Assistant Exams: A Comparative Study

Göz Hastalıkları Alanında Türkçe Başasistanlık Sınavlarında Geniş Dil Modellerinin (ChatGPT-3.5, ChatGPT-4, Bing ve Bard) Performansların Değerlendirilmesi: Karşılaştırmalı Bir Çalışma

Mehmet CANLEBLEBİCİ[a], Ali DAL[b], Murat ERDAĞ[c]

[a]Kayseri State Hospital, Clinic of Ophthalmology, Kayseri, Türkiye
[b]Uğur Eye Hospital, Clinic of Ophthalmology, Kahramanmaraş, Türkiye
[c]Van Training and Research Hospital, Clinic of Ophthalmology, Van, Türkiye

ABSTRACT Objective: Recent strides in artificial intelligence, particularly in large language models (LLM), have prompted their exploration in medical education. This study investigates the proficiency of LLMs in Turkish chief-assistant ophthalmology exams, assessing ChatGPT-3.5, ChatGPT-4.0, Bing, and Bard. The aim is comparing their accuracy in answering 200 questions spanning six critical ophthalmology topics, providing insights into their potential applications in medical education. **Material and Methods:** The questions were asked in Turkish and obtained from the chief-assistant exam administered by the Ministry of National Education from internet. A total of 200 questions were presented to each LLM as fallows ChatGPT-3.5, ChatGPT-4.0, Bing, and Bard, in October 2023. The questions covered six groups, including Retina and Vitreous as Group-1, Cornea, Cataract and Anterior Segment as Group-2, Glaucoma as Group-3, Pediatric Ophthalmology, Genetics and Clinical Refraction as Group-4, Adnexa, Uvea and Oculoplastic as Group-5, and Neuro-Ophthalmology and Strabismus as Group-6. The primary outcome measure was response accuracy, with topics grouped under these six main headings. Statistical analyses were employed to assess the accuracy and reliability of the responses with Pearson's chi-square test. **Results:** ChatGPT-4.0 emerges as the most accurate LLM with a 77.5% correct response rate, followed by Bing at 63.0%. In contrast, ChatGPT-3.5 and Bard exhibit lower accuracy at 51% and 45.5%, respectively. Subgroup analyses emphasize ChatGPT-4.0's superiority across all branches, showcasing its efficacy in diverse ophthalmology topics. **Conclusion:** Despite promising results, the study acknowledges challenges in accuracy and underscores the imperative for continual improvements in LLMs, especially in the realm of clinical applications and education.

Keywords: Artificial intelligence; education and training; large language models; ophthalmology

ÖZET Amaç: Yapay zekâ alanında, özellikle de büyük dil modellerindeki [large language models (LLM)] son gelişmeler, bunların tıp eğitiminde araştırılmasına yol açmıştır. Bu çalışma, ChatGPT-3.5, ChatGPT-4.0, Bing ve Bard'ı değerlendirerek, LLM'lerin Türkçe başasistan oftalmoloji sınavlarındaki yeterliliğini araştırmaktadır. Çalışmanın amacı, 6 ana oftalmoloji konusunu kapsayan 200 soruyu yanıtlamadaki doğruluk oranlarını karşılaştırmak ve tıp eğitimindeki potansiyel uygulamalarına ilişkin alanları tartışmaktır. **Gereç ve Yöntemler:** Araştırmada kullanılmak üzere, önceki yıllarda Millî Eğitim Bakanlığı tarafından uygulanan ve Türkçe sorulan başasistanlık sınavı soruları internet üzerinden elde edildi. Toplam 200 soru Ekim 2023'te her bir LLM'ye ChatGPT-3.5, ChatGPT-4.0, Bing ve Bard olmak üzere tek tek sunulmuştur. Sorular, Grup 1 olarak Retina ve Vitreus, Grup 2 olarak Kornea, Katarakt ve Ön Segment, Grup 3 olarak Glokom, Grup 4 olarak Pediatrik Oftalmoloji, Genetik ve Klinik Refraksiyon, Grup 5 olarak Adneksa, Uvea ve Oküloplastik ve Grup 6 olarak Nöro-Oftalmoloji ve Şaşılık olmak üzere 6 grubu kapsamaktadır. Birincil değerlendirme ölçütü, bu 6 ana başlık altında gruplandırılan konularla birlikte doğru yanıtlama oranıdır. Yanıtların doğruluğunu ve güvenilirliğini değerlendirmek için Pearson'ın ki-kare testi ile istatistiksel analizler yapılmıştır. **Bulgular:** ChatGPT-4.0 %77,5 doğruluk oranıyla en iyi performansı gösteren LLM olmuştur ve onu %63,0 ile Bing takip etmektedir. Buna karşılık, ChatGPT-3.5 ve Bard sırasıyla %51 ve %45,5 ile daha düşük doğruluk oranı sergilemektedir. Alt grup analizleri ChatGPT-4.0'ın tüm branşlardaki üstünlüğünü vurgulayarak çeşitli oftalmoloji konularındaki etkinliğini ortaya koymaktadır. **Sonuç:** Umut verici sonuçlara rağmen bu çalışma doğruluk konusundaki sorunları göstermekte ve özellikle eğitim ve klinik uygulamalar alanında LLM'lerde sürekli iyileştirmeler yapılması zorunluluğunun altını çizmektedir.

Anahtar Kelimeler: Yapay zekâ; eğitim ve öğretim; geniş dil modelleri; oftalmoloji

TO CITE THIS ARTICLE:
Canleblebici M, Dal A, Erdağ M. Evaluation of the performance of large language models (ChatGPT-3.5, ChatGPT-4, Bing and Bard) in Turkish ophthalmology chief-assistant exams: A comparative study. Turkiye Klinikleri J Ophthalmol. 2024;33(3):163-70.

Correspondence: Mehmet CANLEBLEBİCİ
Kayseri State Hospital, Clinic of Ophthalmology, Kayseri, Türkiye
E-mail: mehmetcl@hotmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Ophthalmology.

Received: 06 Mar 2024 ***Received in revised form:*** 02 May 2024 ***Accepted:*** 21 May 2024 ***Available online:*** 11 Jun 2024

Significant advancements have been made in the field of artificial intelligence (AI) in recent times. In particular, thanks to advances in deep learning, natural language processing and large language models (LLM), AI can perform human-like language-based operations such as text comprehension, question answering and text generation. The most extensively used LLM-powered chatbots are Chat Generative Pre-Trained Transformer (ChatGPT) by Open AI Inc. (San Francisco, U.S.), Bing by Microsof Corp. (Washington, U.S.), and Bard by Google LLC. (California, U.S.). In the medical field, these LLMs have a wide range of applications.

These domains are open to further development and are utilized at various levels, including by patients, education and training, health professionals and health systems, ophthalmology in particular, and research.[1,2] Although there are various biases and inaccuracies related to the responses and evaluations of these chatbots, research is nowadays intensively conducted in the field of ophthalmology, especially on exams, patient information, patient eye care and competencies in subspecialties.[3-7] In the research that was done in preparation for the Fellowship of Royal College of ophthalmologists and Fellowship of European Board of Ophthalmology exams, it was found that chatbots were effective.[3,8]

Examinations for board certification in the medical specialty of ophthalmology are given in our nation under the auspices of the Turkish Ophthalmology Association. Candidates who pass the International Council of Ophthalmology Visual Sciences; Optics & Refraction and Instruments; and Clinical Ophthalmology exams apply for this exam and it is held in objective structured clinical examination format. As a consequence, it was not feasible to work with chatbots on our national board exam. In previous years a multiple-choice question style ophthalmology examination for the chief assistant statue was held in our country. Consequently, in our study, we aimed to compare the performance of the 4 most used LLM-powered chatbots over these exam questions on the basis of ophthalmology in Turkish language. Also, these chatbots are only accessible chatbots that can understand and answer the Turkish language at the present. To our knowledge this is the first study to evaluate the effectiveness of LLMs with Turkish questions.

MATERIAL AND METHODS

The questions for the chief assistant exam held by the Ministry of National Education were obtained from the website. After a detailed internet search, made public and available online exam questions from 2010 and 2015 were obtained.[9,10] 200 questions were asked to ChatGPT3.5, ChatGPT4.0, Bing and Bard respectively in October 2023. Bing offers three usage options and was used in “more precise” mode in our study. The questions were asked to LLMs one by one. No new chat page was opened. When the daily usage limit for ChatGPT-4.0 and Bing expired, the chat was continued on the same tab the next day. Primary outcome was response accuracy. The topics were grouped under six main headings and divided into sub-groups. They are defined as fallows; Group 1: Retina and vitreous, Group 2: Cornea, cataract and anterior segment, Group 3: Glaucoma, Group 4: Pediatric ophthalmology, genetics and clinic refraction, Group 5: Adnexa, uvea and oculoplastic, Group 6: Neuro-ophthalmology and strabismus. The second outcome was determined according to the correct response rates of these subgroups.

Each question was asked only once and correct and incorrect answers were recorded. Prompting was never done. Questions were asked directly and were not categorized in terms of difficulty, as this would have distorted objectivity. If the chatbot did not give the correct answer, it was recorded as incorrect. Since the study does not involve living beings such as human and animal objects, ethics committee approval is not required.

STATISTICAL ANALYSIS

Statistical analysis of the results was performed with the program “SPSS for Macintosh Client 25.0 (2016, IBM, Chicago, IL)”. Descriptive statistics and n (%) for categorical variables were used. Pearson’s chi-squared test was used to evaluate the groups among themselves and together. The reliability value of the questions was calculated with Cronbach’s alpha value. Statistical significance was taken as a p value less than 0.05.

RESULTS

ChatGPT-4.0 had the highest accuracy of any chatbot at 77.5% correct responses. Bing, ChatGPT-3.5 and Bard followed (63%, 51% and 45.5%). Statistical analysis of the groups' AI performances revealed that ChatGPT-4.0 outperformed all others, achieving an accuracy rate exceeding 70% across all branches ($p<0.05$). Statistical analysis of the groups' LLM performances revealed that ChatGPT-4.0 outperformed all others, achieving an accuracy rate exceeding 70% across all branches ($p<0.05$). In group 5, Bing achieved a statistically significant advantage ($p=0.042$). Bing provided an approximate 60% correct response rate for the remaining categories; however, no statistically significant results were detected. ChatGPT-3.5 provided responses for the groups with an approximate range of 35% to 65% accuracy. There was no significant difference observed in comparison to other AIs ($p>0.05$). Among all the chatbots, Bard exhibited the lowest accurate response rate. Furthermore, the correct response rates varied the most between groups, spanning from 27.3% to 71.4%. Bard received the least number of responses ($p<0.01$) for Group 3, which comprised topics related to glaucoma.

The mean correct response rates for the question groups were 65.7% for Group 1, 55.68% for Group 2, 52.28% for Group 3, 67.83% for Group 4, 67.25% for Group 5 and 54.63% for Group 6. LLMs performed above 60 percent on average for groups 1, 4 and 5. These groups encompass issues related to the retina-vitreous, pediatric ophthalmology, genetics and clinic refraction and adnexa, uvea-oculoplastic. The most difficult group was the 3rd group with Glaucoma questions except ChatGPT-4.0. Detailed results, statistical values and comparative graph for the groups are summarized in Table 1 and Figure 1.

When the chatbots were compared with each other in groups of two, no statistically significant difference was observed between ChatGPT-4.0 and Bing, while a difference was observed between the other LLMs. When ChatGPT-3.5 and Bard were compared with other LLMs, statistically significant differences were observed with all of them. Table 2 provides a concise overview of the comparison between pairs of AIs.

The Cronbach's alpha coefficient for the questions was 0.712, suggesting that the questions were sufficient in evaluating various subjects.

DISCUSSION

Our research examines the use of several chatbots to assess ophthalmic credentials in the Turkish language. ChatGPT-3.5, ChatGPT-4.0, Bing, and Bard, four of the most popular LLMs right now, were used. Our research showed that ChatGPT-4.0 provided a high degree of question accuracy. Bing is in second place with 63% correct answers, although statistically significant. Generally, 60% is used as a passing grade

TABLE 1: Correct response rates according to LLMs and statistical comparison of groups together.

	Question numbers	ChatGPT-3.5		ChatGPT-4.0		Bing		Bard		Overall values for each group
		Accuracy	p value	Accuracy	p value	Accuracy	p value	Accuracy	p value	Accuracy
Group 1	35	65.7%	p=0.063	80.0%	**p<0.001**	65.7%	p=0.063	51.4%	p=0.866	65.70%
Group 2	62	46.8%	p=0.611	72.6%	**p<0.001**	58.1%	p=0.204	45.2%	p=0.446	55.68%
Group 3	33	36.4%	p=0.117	81.8%	**p<0.001**	63.6%	p=0.117	27.3%	**p=0.009**	52.28%
Group 4	14	57.1%	p=0.286	85.7%	**p=0.008**	57.1%	p=0.593	71.4%	p=0.109	67.83%
Group 5	29	62.1%	p=0.593	79.3%	**p=0.002**	69.0%	**p=0.041**	58.6%	p=0.353	67.25%
Group 6	27	44.4%	p=0.564	74.1%	**p=0.012**	66.7%	p=0.083	33.3%	p=0.083	54.63%
Overall values for each LLM	200	51.0%	p=0.777	77.5%	**p<0.001**	63.0%	**p<0.001**	45.5%	p=0.203	59.25%

Group 1: Retina and vitreous, Group 2: Cornea, cataract and anterior segment, Group 3: Glaucoma, Group 4: Pediatric ophthalmology, genetics and clinic refraction, Group 5: Adnexa, uvea and oculoplastic, Group 6: Neuro-ophthalmology and strabismus; p values with statically significant are shown with bold characters; The all groups are compared with Pearson chi-square test ; LLM: Large language model.

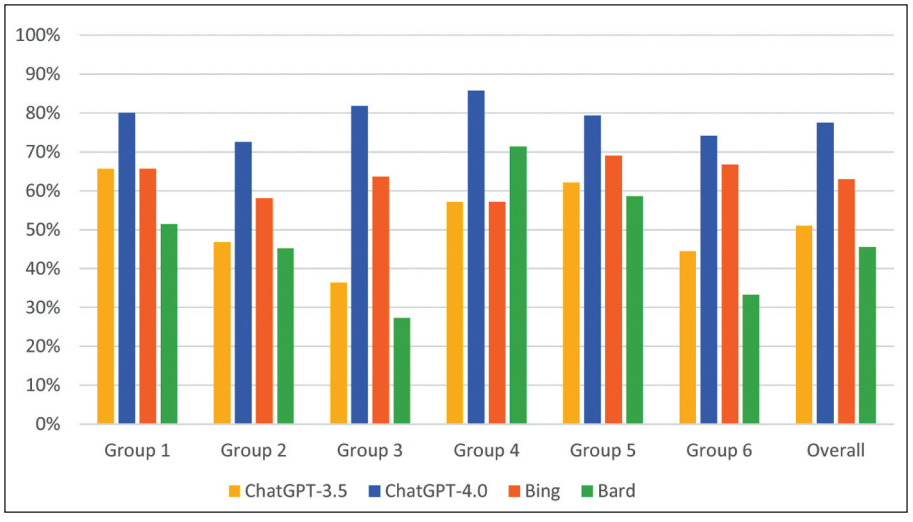

FIGURE 1: Comparison of large language model -powered chatbots' accuracy for all groups and overall results. The x-axis shows the percentage and the y-axis shows the groups. Group 1: Retina and vitreous, Group 2: Cornea, cataract and anterior segment, Group 3: Glaucoma, Group 4: Pediatric ophthalmology, genetics and clinic refraction, Group 5: Adnexa, uvea and oculoplastic, Group 6: Neuro-ophthalmology and strabismus.

TABLE 2: Statistical comparison of correct answer rates one by one according to LLMs with Pearson chi-square test.

LLM Chatbots		p value	LLM Chatbots		p value
ChatGPT-3.5 vs.	ChatGPT-4	**p<0.001**	ChatGPT-4 vs.	Bing	p=0.127
	Bing	**p<0.001**		Bard	**p<0.001**
	Bard	**p<0.001**		ChatGPT-3.5	**p<0.001**
LLM Chatbots		**p value**	**LLM Chatbots**		**p value**
Bing vs.	Bard	**p=0.024**	Bard vs.	ChatGPT-3.5	**p<0.001**
	ChatGPT-3.5	**p<0.001**		ChatGPT-4	**p<0.001**
	ChatGPT-4	p=0.127		Bing	**p=0.024**

LLM: Large language model.

in exams. At this point, it's safe to assume that Bing will pass the vast majority of tests. Neither ChatGPT-3.5 nor Bard were able to provide answers that were right. A lot of work has been done on ChatGPT-4.0 and it is one of the most successful LLMs-based chatbots for examinations. All four AIs we tested were employed by Raimondi et al. in the Royal College of Ophthalmologists' fellowship exams, with ChatGPT-4.0 achieving the highest percentage of accuracy (82.9%).[8] Similar to our own research, Bing's accuracy here was satisfactory. ChatGPT-3.5 and Bard, on the other hand, stayed around 50%. Another study found that when simply ChatGPT-4.0 was used to analyze the European Board of Ophthalmology fellowship exam in French, the LLM had a 91% success rate.[3]

A recent study assessed human participants as well as the LLMs Bing, ChatGPT-3.5, and ChatGPT-4.0. The study included 250 questions sourced from the Basic Science and Clinical Science Self-Assessment Program of the American Academy of Ophthalmology. The human participants achieved a score of 72.2%, while ChatGPT-3.5, ChatGPT-4.0, and Bing Chat achieved scores of 58.8%, 71.6%, and 71.2% respectively.[4] Antaki et al. generated two 260-question simulated exams from the Basic and Clinical Science Course Self-Assessment Program and the OphthoQuestions online question bank for evaluate

ChatGPT-3.5 and 4.0.[11] Although lower than other studies, ChatGPT-4.0 gave a correct answer rate of approximately 60% in the first group of questions and 50% in the second group of questions and also better than its older version. Recently, a United Kingdom magazine conducted a comparison between Bard and ChatGPT for Fellowship in Ophthalmology, specifically focusing on examination part 1. ChatGPT outperformed the human average by a significant margin, although Bard's performance fell short of it.[12] The findings of our investigation align with the conclusions presented in these five studies. ChatGPT-4.0 exams have demonstrated remarkable success in the field of ophthalmic examinations. Bing has a high success rate. ChatGPT-3.5 and Bard yield subpar outcomes.

In order to comprehend the performances of LLMs, it is imperative to assess the operational methodologies of these four distinct LLMs. ChatGPT-4.0 differs from version 3.5 by being multimodal, much more processing power, much more nuanced, more accurate, less prone to hallucinations.[13] Bing was developed by Microsoft but uses ChatGPT-4.0 as its LLM base.[14] However, Bing differs from ChatGPT-4.0 by being leaner, less sophisticated and less trainable than ChatGPT-4.0.[15] Bing chat uses GPT-4.0 AI architecture with Bing search engine, indexing and data ranking. Basically, Bing was designed this way because it has a different purpose than ChatGPT-4.0 and being connected to the internet aims to provide a more refined experience for the user. Simpler conversations, with simpler and faster response rates, can be used for free with the GPT-4.0 model thanks to Bing chat.[16] It offers different experiences with three different usage options. On the other hand, ChatGPT-4.0 offers a more advanced, deep and sophisticated, deeply layered and modifiable experience, but at a cost. The AI tool Bard, a competitor to OpenAI's ChatGPT, uses PaLM2, a LLM developed by Google.[17] At this point Bard is in an experimental phase and is constantly receiving updates and evolving. The database of ChatGPT was closed to data from September 2021 onwards, but the company has recently announced that current data will be available. At the time of our study, data before September 2021 was valid for ChatGPT 3.5 and 4.0.[18] Given the continuous updates and training capabilities of AIs, it is likely that the effectiveness of these LLMs will improve in the future. In our study, as we mentioned above about the current status of AI, the correct answer rates to the exam questions were similarly observed. ChatGPT-4.0, which is evaluated as the most up-to-date and powerful LLM, was the most successful AI. Bing uses the GPT-4.0 model and ranked second. The older version, ChatGPT-3.5, which is currently available for free, and Bard, which is developing and experimental, had low success rates. However, a future database update for ChatGPT-3.5 and 4.0 and future enhancements to Bing and Bard may change these rankings. AI technology is perpetually receptive to advancement.

ChatGPT-4.0 demonstrated superior performance while assessing subspecialties. Results in groups 1, 3, 4, and 5 reached or above 80%. In contrast, Bing yielded outcomes that closely aligned with its own average of 63%. Although ChatGPT-3.5 achieved a success rate of approximately 60% in Groups 1, 4, and 5, it notably struggled with the glaucoma-related queries. Bard answered correctly approximately 30% of the questions on glaucoma, neuro-ophthalmology and strabismus, while other groups it had variable results. Jiao et al. used ChatGPT-3.5 and 4.0 in their study for ophthalmology subspecialties using multiple-choice ophthalmic clinical cases provided by the American Academy of Ophthalmology.[19] In this study, version 4.0 was 75% to 46% more successful than version 3.5. Even in subspecialties, version 4.0 was better than 3.5 in every area. In our study, ChatGPT-4.0 was the most successful in every branch in subgroup analyses. Especially neuro-ophthalmology, cornea/anterior segment and pediatric ophthalmology questions were answered correctly by ChatGPT-4.0 in this study. In the above-mentioned study by Raimondi et al., LLMs were found to be successful, especially in cornea, cataract and external eye disease questions.[8] These differences may be due to reasons such as the difficulty levels of the questions, their understandability, whether they contain photographs, and the number of questions.

Recently, there has been a shift towards publishing separate studies for each discipline instead of reviewing each specialism inside a single study.

Holmes et al. assessed the performance of ChatGPT-3.5, 4.0, and PaLM2 (a different AI language model developed by Google, now assisting Bard) on a set of 100 inquiries related to pediatric ophthalmology.[20] In this article, while ChatGPT-4.0 showed similar performance to human participants as expected, other LLMs showed lower performance. Sensoy and Citirik utilized ChatGPT-3.5, Bing, and Bard to assess the study questions part of the American Academy of Ophthalmology's 2022-2023 Basic and Clinical Science Course on Ophthalmic Pathology and Intraocular Tumor.[21] In this study, correct response rates were 58.6%, 63.9% and 69.4%, respectively, but no statistically significant difference was observed. Although it is a current study, ChatGPT-4.0 has not been evaluated. Delsoz et al. evaluated ChatGPT-3.5 and 4.0 on corneal diseases.[22] ChatGPT-4.0 has demonstrated improved accuracy in generating suitable responses, showing great potential in this domain. There is likely to be a rise in the quantity of research conducted on these specific subjects in the near future. So, the success of AI powered LLMs in every field will be better understood.

Although AI powered LLMs have achieved impressive outcomes and offer potential opportunities, they also encounter some issues. There may be lack of accuracy and consistency, hallucinations, falsehood mimicry and biases in the answers given by AI powered LLMs, and these issues can be problematic since they may result in the dissemination of inaccurate information, particularly if patients are granted access to such responses.[5,23,24] Another issue is the problem of accessing and interpreting current data since some of these chatbots do not have up-to-date access and some of them are in the development phase.[25] Data pollution and source access containing incorrect information are also possible.[26] In such cases, responses may need to be moderated.[27] For such reasons, chatbots are not currently suitable for clinical use.[28,29] Further research is necessary to have a deeper understanding of the benefits and drawbacks of AI powered chatbots, as this area appears to be expanding rapidly.[30]

The study is limited by the inability to make a comparison between the exam results and those of humans, the absence of questions featuring photos, and the non-uniform distribution of specialist groups. This comparison is not feasible due to the inability to obtain the average of past exam scores, which would introduce bias if we were to attempt to solve the test ourselves. The question distribution within the group was not a variable that could be changed. Prompting can increase the correct response rates of questions, but it can also lead to misdirection when it is not used correctly.[31] In fact, which prompting is effective and how it affects the results may interestingly be the subject of another study. Chatbots that can speak the user's native language fluently can better understand queries and provide more precise information. Therefore, a chatbot's language ability can significantly affect the accuracy of its responses. For the reasons mentioned above, we found it appropriate to evaluate LLMs in their natural state in our study.

CONCLUSION

The most successful of the four AI powered LLMs currently available for use in ophthalmic examinations using the Turkish language is ChatGPT-4.0. Bing, despite its lower rank, can be regarded a moderately successful search engine. The ChatGPT-3.5 and the Bard skill did not reach adequate levels. In the not-too-distant future, chatbots may be able to make significant contributions to advancements in the field of ophthalmological education. However, there is a need for future improvements in terms of competency and reliability.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Mehmet Canleblebici, Ali Dal, Murat Erdağ;* ***Design:*** *Mehmet Canleblebici, Ali Dal, Murat Erdağ;* ***Control/Su-***

pervision: *Mehmet Canleblebici, Murat Erdağ;* ***Data Collection and/or Processing:*** *Mehmet Canleblebici, Ali Dal;* ***Analysis and/or Interpretation:*** *Mehmet Canleblebici, Ali Dal, Murat Erdağ;* ***Literature Review:*** *Mehmet Canleblebici, Ali Dal, Murat Erdağ;* ***Writing the Article:*** *Mehmet Canleblebici, Ali Dal;* ***Critical Review:*** *Mehmet Canleblebici, Murat Erdağ;* ***References and Fundings:*** *Mehmet Canleblebici;* ***Materials:*** *Mehmet Canleblebici.*

REFERENCES

1. Ting DSJ, Tan TF, Ting DSW. ChatGPT in ophthalmology: the dawn of a new era? Eye (Lond). 2024;38(1):4-7. [Crossref] [PubMed] [PMC]
2. Honavar SG. Eye of the AI storm: exploring the impact of AI tools in ophthalmology. Indian J Ophthalmol. 2023;71(6):2328-40. [Crossref] [PubMed] [PMC]
3. Panthier C, Gatinel D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: a novel approach to medical knowledge assessment. J Fr Ophtalmol. 2023;46(7):706-11. [Crossref] [PubMed]
4. Cai LZ, Shaheen A, Jin A, Fukui R, Yi JS, Yannuzzi N, et al. Performance of generative large language models on ophthalmology board-style questions. Am J Ophthalmol. 2023;254:141-9. [Crossref] [PubMed]
5. Tan TF, Thirunavukarasu AJ, Campbell JP, Keane PA, Pasquale LR, Abramoff MD, et al. Generative artificial intelligence through ChatGPT and other large language models in ophthalmology: clinical applications and challenges. Ophthalmol Sci. 2023;3(4):100394. [Crossref] [PubMed] [PMC]
6. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine. 2023;95:104770. [Crossref] [PubMed] [PMC]
7. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open. 2023;6(8):e2330320. [Crossref] [PubMed] [PMC]
8. Raimondi R, Tzoumas N, Salisbury T, Di Simplicio S, Romano MR; North East Trainee Research in Ophthalmology Network (NETRiON). Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. Eye (Lond). 2023;37(17):3530-3. [Crossref] [PubMed] [PMC]
9. Oftalmologlar [İnternet]. [Erişim tarihi: 17 Kasım 2023]. Nisan 2010 Dönemi Başasistanlık Sınav Soruları. Erişim linki: [Link]
10. Türk Oftalmoloji Başasistanlığı Sınavı 2015. Erişim tarihi: 17 Kasım 2023. Erişim linki: [Link]
11. 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. Ophthalmol Sci. 2023;3(4):100324. [Crossref] [PubMed] [PMC]
12. Fowler T, Pullen S, Birkett L. Performance of ChatGPT and Bard on the official part 1 FRCOphth practice questions. Br J Ophthalmol. 2023:bjo-2023-324091. [Crossref] [PubMed]
13. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. International Journal of Oral Science. 2023;15:29. [Crossref] [PubMed] [PMC]
14. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. J Med Internet Res. 2023;25:e51580. [Crossref] [PubMed] [PMC]
15. Motlagh NY, Khajavi M, Sharifi A, Ahmadi M. The impact of artificial intelligence on the evolution of digital education: a comparative study of openAI text generation tools including ChatGPT, Bing Chat, Bard, and Ernie. arXiv. 2023. [Link]
16. Lozić E, Štular B. ChatGPT v Bard v Bing v Claude 2 v Aria v human-expert. How good are AI chatbots at scientific writing? arXiv. 2023. [Link]
17. Waisberg E, Ong J, Masalkhi M, Zaman N, Sarker P, Lee AG, et al. Google's AI chatbot "Bard": a side-by-side comparison with ChatGPT and its utilization in ophthalmology. Eye (Lond). 2024;38(4):642-5. [Crossref] [PubMed] [PMC]
18. Vaishya R, Misra A, Vaish A. ChatGPT: is this version good for healthcare and research? Diabetes Metab Syndr. 2023;17(4):102744. [Crossref] [PubMed]
19. Jiao C, Edupuganti NR, Patel PA, Bui T, Sheth V. Evaluating the artificial intelligence performance growth in ophthalmic knowledge. Cureus. 2023;15(9):e45700. [Crossref] [PubMed] [PMC]
20. Holmes J, Peng R, Li Y, Hu J, Liu Z, Wu Z, et al. Evaluating multiple large language models in pediatric ophthalmology. arXiv. 2023. [Link]
21. Sensoy E, Citirik M. A comparative study on the knowledge levels of artificial intelligence programs in diagnosing ophthalmic pathologies and intraocular tumors evaluated their superiority and potential utility. Int Ophthalmol. 2023;43(12):4905-9. [Crossref] [PubMed]
22. Delsoz M, Madadi Y, Munir WM, Tamm B, Mehravaran S, Soleimani M, et al. Performance of ChatGPT in diagnosis of corneal eye diseases. medRxiv [Preprint]. 2023:2023.08.25.23294635. Update in: Cornea. 2024;43(5):664-70. [Crossref] [PubMed] [PMC]
23. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care. 2023;27(1):120. [Crossref] [PubMed] [PMC]
24. Ashraf H, Ashfaq H. The role of chatgpt in medical research: progress and limitations. Ann Biomed Eng. 2024;52(3):458-61. [Crossref] [PubMed]
25. Hadi MU, Qureshi R, Shah A, Irfan M, Zafar A, Shaikh M, et al. A survey on large language models: applications, challenges, limitations, and practical usage. TechRxiv. 2023. [Crossref]
26. Waters MR, Aneja S, Hong JC. Unlocking the power of ChatGPT, artificial intelligence, and large language models: practical suggestions for radiation oncologists. Pract Radiat Oncol. 2023;13(6):e484-e90. [Crossref] [PubMed]

27. Bang J, Lee B-T, Park P. Examination of Ethical Principles for LLM-Based Recommendations in Conversational AI. 2023 International Conference on Platform Technology and Service (PlatCon): IEEE. 2023:109-13. [Crossref] [PubMed]

28. Au Yeung J, Kraljevic Z, Luintel A, Balston A, Idowu E, Dobson RJ, et al. AI chatbots not yet ready for clinical use. Front Digit Health. 2023;5:1161098. [Crossref] [PubMed] [PMC]

29. Korot E, Wagner SK, Faes L, Liu X, Huemer J, Ferraz D, et al. Will AI replace ophthalmologists? Transl Vis Sci Technol. 2020;9(2):2. Erratum in: Transl Vis Sci Technol. 2021;10(8):6. [Crossref] [PubMed] [PMC]

30. Ghadiri N. Comment on: 'Comparison of GPT-3.5, GPT-4, and human user performance on a practice ophthalmology written examination' and 'ChatGPT in ophthalmology: the dawn of a new era?'. Eye (Lond). 2024;38(4):654-5. [Crossref] [PubMed] [PMC]

31. Kleinig O, Gao C, Kovoor JG, Gupta AK, Bacchi S, Chan WO. How to use large language models in ophthalmology: from prompt engineering to protecting confidentiality. Eye (Lond). 2024;38(4):649-53. [Crossref] [PubMed] [PMC]