Comparing generative and retrieval-based chatbots in answering patient questions regarding age-related macular degeneration and diabetic retinopathy

Kai Xiong Cheong, FRCOphth,[1] Chenxi Zhang, MD,[2] Tien-En Tan, FRCOphth,[1,3] Beau J. Fenner, PhD,[1,3] Wendy Meihua Wong, FRCOphth,[4,5,6] Kelvin Yi Chong Teo, PhD,[1,3] Ya Xing Wang, MD,[7] Sobha Sivaprasad, MD,[8] Pearse A. Keane, MD,[8] Cecilia S. Lee, MD,[9] Aaron Y. Lee, MD,[9] Gemmy Cheung, FRCOphth,[1,3] Tien Yin Wong, PhD,[1,10] Yun-Gyung Cheong, PhD,[11] Su Jeong Song, PhD,[12,13†] Yih Chung Tham, PhD[1,3,5,6†]

[1] Singapore Eye Research Institute, Singapore National Eye Centre, Singapore.

[2] Department of Ophthalmology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, China.

[3] Ophthalmology & Visual Sciences Academic Clinical Program (Eye ACP), Duke-NUS Medical School, Singapore.

[4] Department of Ophthalmology, National University Hospital, National University Health System, Singapore.

[5] Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.

[6] Centre for Innovation & Precision Eye Health, Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore.

[7] Beijing Institute of Ophthalmology, Beijing Key Laboratory of Ophthalmology and Visual Sciences, Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing, China.

[8] NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust, London, UK Institute of Ophthalmology, University College London, London, UK.

[9] Department of Ophthalmology, University of Washington, Seattle, WA, USA.

[10] Tsinghua Medicine, Tsinghua University, Beijing, China.

[11] Department of Artificial Intelligence, Sungkyunkwan University, Suwon, Republic of Korea.

[12] Department of Ophthalmology, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea.

[13] Biomedical Institute for Convergence (BICS), Sungkyunkwan University, Suwon, Republic of Korea.

[†]Contributed equally and are joint last authors

Correspondence and reprint requests to:

Dr Su Jeong Song

Department of Ophthalmology, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

Email: sjsong7@gmail.com

Dr Yih Chung Tham

Singapore Eye Research Institute, Singapore National Eye Centre, Singapore

The Academia, 20 College Rd, Level 6 Discovery Tower, Singapore 169856

Phone: (65) 6576 7200

Email: thamyc@nus.edu.sg

Supported by:

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub).

Conflict of Interest:

No conflicting relationship related to this work exists for any author.

Running Head:

Chatbots in age-related macular degeneration and diabetic retinopathy

Key Words:

Chatbot, generative, retrieval-based, large language model, age-related macular degeneration, diabetic retinopathy

Word Count

3071 words

Abstract

Importance: Artificial intelligence (AI) chatbots demonstrate considerable potential to improve healthcare. However, studies comparing generative and retrieval-based AI chatbots in medical domain are notably absent.

Objective: To compare the performance of generative versus retrieval-based chatbots in answering patient inquiries regarding age-related macular degeneration (AMD) and diabetic retinopathy (DR).

Design: Cross-sectional study.

Settings: We evaluated four chatbots: generative models (ChatGPT-4, ChatGPT-3.5, and Google Bard), and a retrieval-based model (OcularBERT). Their responses to 15 AMD- and 15 DR-related question were evaluated and compared.

Participants: Three retinal specialists with at least seven years of retinal subspecialty experience.

Exposure: Four chatbots were evaluated between 10th May to 30th June 2023.

Main Outcomes and Measures: The primary outcome was the total accuracy score. Three masked retinal specialists graded the responses using a three-point Likert scale: either 2 (good, error-free), 1 (borderline), or 0 (poor with significant inaccuracies). The scores were aggregated, ranging from 0 to 6. Based on majority consensus across the three graders, the responses were also classified as “Good”, “Borderline”, or “Poor” quality.

Results: ChatGPT-4 and ChatGPT-3.5 outperformed the other chatbots, both achieving median scores (interquartile range) of 6 (1), compared with 4.5 (2) in Google Bard, and 2 (1) in OcularBERT (all $p \leq 8.4 \times 10^{-3}$). Based on the consensus approach, 83.3% of ChatGPT-4's responses and 86.7% of ChatGPT-3.5's responses were rated as "Good", surpassing Google Bard (50%) and OcularBERT (10%) (all $p \leq 1.4 \times 10^{-2}$). ChatGPT-4 and ChatGPT-3.5 had no “Poor” rated responses, but Google Bard produced 6.7% Poor responses and OcularBERT produced 20%. Across AMD- and DR-specific categories, these performances remained largely similar.

Conclusions and Relevance: ChatGPT-4 and ChatGPT-3.5 demonstrated superior performance, followed by Google Bard and OcularBERT. Generative chatbots are capable of answering domain-specific questions outside their original training. Further studies are required for real-world implementation.

Introduction

In healthcare, artificial intelligence (AI) chatbots demonstrate considerable potential to improve access to information, facilitate timely communication, enhance self-care management, and provide contextually relevant support to patients.[1–4] These potential capabilities may lead to informed decision-making by patients, improved patient experience, reduced hospital caseload, and enable physicians to focus on complex tasks.

Broadly, chatbots utilise machine learning and natural language processing (NLP) techniques to process and understand human language in digital form.[5–7] Chatbots can be categorised as generative or retrieval-based. Generative chatbots, which can generate responses *de novo*, have gained traction recently, especially with the rapid advancements in large language models (LLMs). Chat Generative Pre-trained Transformer (ChatGPT; OpenAI), which is based on the GPT architecture, and Google Bard, which is based on LaMDA (Language Model for Dialogue Applications), have both experienced a significant surge in usage due to advancements in LLMs.[8] These LLMs are pre-trained on large datasets that encompass existing language patterns using self-supervision at scale, and can be used for multiple tasks,[9] including question answering,[10–13] language translation,[13,14] clinical documentation,[15,16] research writing,[17,18] and diagnostic assistance.[19–21]

In contrast, retrieval-based chatbots, such as OcularBERT (Pre-Trained BERT for Ophthalmic Multi-Step Retrieval), utilise a pre-existing repository of responses from a specific knowledge base.[22–25] Such chatbots match user inputs to the most relevant pre-defined responses based on similarity or relevance metrics. OcularBERT, one of the first validated domain-specific chatbots, specialises in ophthalmology-related question-and-answer tasks, particularly on age-related macular degeneration (AMD) and diabetic retinopathy (DR).[22]

The use of AI chatbots in healthcare settings has garnered increasing interest. Generative chatbots, like ChatGPT and Google Bard that are trained on general data, may lack domain-specific capabilities.[26] Conversely, retrieval-based chatbots, which are trained on curated data, may excel only within their specific training domains.[27] Each medical specialty presents unique challenges requiring distinct contextualisation and clinical reasoning. This underscores the importance of evaluating the performance of these AI chatbots in different medical domains.

Notably, studies comparing generative and retrieval-based AI chatbots in addressing questions related to retinal diseases are lacking. AMD and DR are leading causes of visual impairment and are topics frequently asked by patients.[28] The performance of AI chatbots in subspecialties of ophthalmology, particularly in addressing questions about AMD and DR, remains under-explored.[29] Recognising these knowledge gaps, our study aimed to compare the performance of generative and retrieval-based AI chatbots in answering common patient questions regarding AMD and DR.

Methods

Study design

This was a cross-sectional study that was conducted between 10th May to 30th June 2023. No ethics approval was required as no identifiable patient data were used. Four chatbots were assessed: ChatGPT-4 and ChatGPT-3.5 (OpenAI; San Francisco, CA, USA), Google Bard (Google LLC; Mountain View, CA, USA), and OcularBERT. OcularBERT's design has been described in detail in our previous paper.[22] ChatGPT-3.5 and Google Bard are available for use by the public. ChatGPT-4 requires paid-subscription. OcularBERT is licensed-owned, and its access requires permission from the developers (YGC, SJS).

A total of 45 questions from three categories were assessed: AMD (15), DR (15), and Others (15). These questions are commonly asked by patients and were randomly selected from the question-and-answer paired data that was originally used to train OcularBERT.[22] **Supplementary Table 1** presents the full list of questions and responses by each chatbot.

Figure 1 illustrates the overall study design. In brief, the questions were input individually as standalone queries for each chatbot. Each chatbot response was generated in a new session after clearing the preceding question and answer, to prevent undue influence from concurrent conversations. To mask the chatbot identify, the responses were formatted into plain text, randomly ordered, and stripped of information which might reveal the identity of the chatbot. The responses by the four chatbots to all 45 questions were randomly distributed into four grading sets. Each set of 45 responses was graded in a separate session, with a 24-hour washout between sessions. The grading was independently performed by three fellowship-trained retinal specialists (BF, TET, and WMW), each having at least seven years of subspecialty experience.

Evaluation of performance

The primary outcome was the accuracy of the chatbot responses. The responses were graded using a three-point Likert scale as follows: score of 0 denotes "Poor" (unacceptable inaccuracy, highly likely to mislead patients and cause harm), score of 1 denotes "Borderline" (borderline accuracy with potential factual errors, but unlikely to mislead the average patient or cause harm), and score of 2 denotes "Good" (error-free). The total accuracy score was summed up across the three graders, with a minimum of 0 and maximum of 6.

We also utilised a majority consensus approach, determining the final rating for each chatbot response based on the most common grade among the three graders. In the situation where each grader assigned a different grade and a common consensus was not reached, we defaulted to a stringent approach and the lowest score of 0 (Poor) was assigned to the chatbot response.

Evaluation of self-correction capabilities

Amongst the generative chatbots, the original "Poor" rated responses were further prompted to self-correct using the following line: "This does not seem right. Could you kindly review your response?" The revised responses by the chatbots were subsequently re-assessed by the three graders for accuracy. For this, the chatbot's identity and original responses were also masked.

Additional qualitative evaluation of poorly rated responses

To elucidate the potential limitations and risks of using generative chatbot responses for information, poorly rated responses were analysed. The graders provided explanations for their gradings.

Statistical analysis

All statistical analyses were performed in R statistical software, version 4.3.0 (R Project for Statistical Computing, R Foundation, Vienna, Austria). The Shapiro-Wilk test was used to assess normality of data distribution. The Kruskal-Wallis H test was used to compare the median scores. The two-tailed Chi-squared test was used to compare the proportions of “Good”, “Borderline”, and “Poor” responses. The one-way analysis of variance test was used to compare mean length of the responses. When performing pairwise comparisons, we applied Bonferroni corrections. A p-value of less than 0.05 was used as the significance value.

Results

ChatGPT-3.5 produced the longest response (292.0±73.4 words), followed by ChatGPT-4 (214.0±57.4 words), Google Bard (183±56.7 words), and then OcularBERT (36.5±23.6 words), overall $p<2.0x10^{-16}$. Pairwise comparisons indicated that the difference in word counts among the chatbots were significant, except for that between GPT-4 and Bard. See **Table 1**.

Response accuracy

Overall, for AMD and DR, ChatGPT-4 and ChatGPT-3.5 outperformed the other chatbots, both achieving median scores (interquartile range) of 6 (1), compared with 4.5 (2) in Google Bard, and 2 (1) in OcularBERT (all $p\leq8.4x10^{-3}$) (**Table 2**). In the pairwise comparisons, the scores of ChatGPT-4 and ChatGPT3-5 were significantly higher than that of Google Bard (ChatGPT-4 versus Google Bard: $p=6.8x10^{-4}$; ChatGPT-3.5 versus Google Bard: $p=8.4x10^{-3}$), which was in turn higher than that of OcularBERT (Google Bard versus OcularBERT: $p=8.0x10^{-6}$). There were no differences between ChatGPT-4 and ChatGPT-3.5 (**Figure 2**). The total score for each question is shown in **Supplementary Table 2**.

Table 3 shows the sub-analysis of response scores by question type (AMD [**Supplementary Figure 1**], DR [**Supplementary Figure 2**], and Others [**Supplementary Figure 3**]). ChatGPT-4, ChatGPT-3.5, and Google Bard still outperformed OcularBERT for all question types. There were also no differences between ChatGPT-4 and ChatGPT-3.5. There were, however, minor differences between ChatGPT-4 and ChatGPT-3.5 versus Google Bard among question types after accounting for multiple comparisons. ChatGPT-4 outperformed Google Bard for AMD (6 [2] versus 4 [2], $p=2.7x10^{-3}$), but had similar performance for DR (6 [0] versus 6 [2]) and Others (6 [1] versus 5 [2]). ChatGPT-3.5 outperformed Google Bard for DR (6 [0] versus 6 [2], $p=2.4x10^{-2}$) and Others (6 [0.5)] versus 5 [2], $p=2.7x10^{-2}$), but had a similar performance for AMD (5 [2] versus 4 [2]).

Consensus grading

Overall, for AMD and DR, 83.3% of ChatGPT-4's responses and 86.7% of ChatGPT-3.5's responses were rated as "Good", surpassing Google Bard (50%) and OcularBERT (10%) (all $p\leq1.4x10^{-2}$) (**Table 2**). Whilst ChatGPT-4 and ChatGPT-3.5 had no “Poor” rated responses, Google Bard produced 6.7% Poor responses and OcularBERT produced 20%. Similarly, in the pairwise comparisons, both ChatGPT-4 and ChatGPT3-5 produced significantly higher proportions of Good responses compared with Google Bard (ChatGPT-4 versus Google Bard: $p=1.4x10^{-2}$; ChatGPT-3.5 versus Google Bard: $p=5.5x10^{-3}$), which was in turn higher than that of OcularBERT (Google Bard versus OcularBERT: $p=1.9x10^{-3}$). There were no differences between ChatGPT-4 and Chat GPT-3.5. (**Figure 3**). The consensus grading for each question is shown in **Supplementary Table 2.**

Table 3 shows the sub-analysis of the consensus grading by question type (AMD [**Supplementary Figure 4**], DR [**Supplementary Figure 5**], and Others [**Supplementary Figure 6**]). ChatGPT-4 and ChatGPT-3.5 still had significantly higher proportions of Good responses compared with OcularBERT for all question types. There were also no differences between ChatGPT-4 and ChatGPT-3.5. However, both ChatGPT-4 and ChatGPT-3.5 performed similarly to Google Bard for all question types after accounting for multiple comparisons. Google Bard still had a higher proportion of Good responses compared with OcularBERT for DR (60.0% versus 6.7%, $6.7x10^{-3}$), but not for AMD and Others.

Correction of poor responses by generative chatbots

As ChatGPT-4 and ChatGPT-3.5 did not produce any Poor response, we evaluated Google Bard’s ability to correct its poorly rated responses after prompting. The consensus grading improved for three of five (60%) questions: from Poor to Borderline for two questions, and from Poor to Good for one question. **Supplementary Table 3** shows the initial and corrected responses by Google Bard.

Qualitative evaluation of poor responses by Bard and OcularBERT

In all, Google Bard and OcularBERT had five and seven Poor responses, respectively, among all question types. **Supplementary Table 4** illustrates these responses and the graders' explanations for the gradings. In general, these responses were graded as Poor because the responses were inaccurate and misleading, and/or had incomplete information.

Discussion

We evaluated the performance of four chatbots (ChatGPT-4, ChatGPT-3.5, Google Bard, and OcularBERT) in answering patient questions regarding AMD and DR. To date, this represents the first study that compares the performance of both generative and retrieval-based chatbots in Ophthalmology. ChatGPT-4 and ChatGPT-3.5 had the best performance, followed by Google Bard, and then OcularBERT for accuracy score and consensus grading. ChatGPT-4 and ChatGPT-3.5 performed similarly. Generative chatbots can effectively answer domain-specific questions that they were not originally trained for and are potentially capable of correcting responses after prompting. These findings provide useful insights into the potential roles chatbots can play in healthcare, such as assisting healthcare practitioners in communicating with patients.

One striking observation was that the generative models (ChatGPT-4, ChatGPT-3.5, and Google Bard), although not specifically trained to respond to such questions, outperformed OcularBERT, a retrieval-based model that had been designed to focus on medical question-and-answer tasks. This is an interesting finding as a common criticism of generative chatbots are that although they have impressive capabilities in generating text across a wide range of topics, they may lack the necessary knowledge in specific domains. The performance of ChatGPT and Google Bard are likely due to the size of the training data. ChatGPT was trained on multiple extremely large datasets, of which

Common Crawl[a] constituted the bulk of the datasets.[30,31] Google Bard was trained on Infiniset[b] and continually draws information from the internet.[30,31] The sheer size of the training datasets is likely what enabled ChatGPT-4, ChatGPT-3.5 and Google Bard to outperform OcularBERT. Another key contributory factor is the chatbot architecture. ChatGPT and Google Bard are based on transformers[c] and take advantage of a variety of supervised, self-supervised, and fine-tuning approaches using large amounts of data. This equips them to perform tasks that they were not specifically trained for, including the answering of medical questions. In contrast, OcularBERT is based on a pre-trained model (BERT; Bidirectional Encoder Representations from Transformers) that was trained with the Wikipedia corpus, then further trained with 1176 ophthalmic disease question and answer paired data.[22] OcularBERT relies on training data curation and the responses are pre-defined.

ChatGPT-4 and ChatGPT-3.5 demonstrated similar performance, although ChatGPT-4 is larger than ChatGPT-3.5, which further improves upon its NLP capabilities and is purported to produce more accurate and relevant responses.[32,33] While OpenAI has not released a detailed technical report on ChatGPT-4's architecture, and ChatGPT-4 generally retains the same transformer-based encoder-decoder architecture of its predecessors in the ChatGPT family, a few differences between ChatGPT-4 and ChatGPT-3.5 have been identified. These include the model size: 170 trillion versus 175 billion parameters; modality: text and images versus text alone; and context window length: 8192 (32768 is offered to select test users at time of writing) versus 4096 tokens[d], respectively.[32,33] Our results are different from those of Raimondi et al.[34] and Ali et al.,[35] who demonstrated that ChatGPT-4 outperformed ChatGPT-3.5 in neurosurgery and ophthalmology examinations, respectively. This may be attributed to differences between examination and patient questions, in which the former may be

[a] Open repository of web crawl data that contains petabytes of raw web page data, extracted metadata, and text extractions since 2008.

[b] Data comprising public forum dialogue, Common Crawl's web crawl, publicly available programming documents, English Wikipedia, and other public web documents.

[c] Neural network architecture first described in 2017 that learns context and meaning by tracking relationships in sequential data, and use encoders and decoders.

[d] Basic units of text or code that an LLM AI model uses to process and generate language.

more factual in nature, and the latter being more subjective. Another reason for differences in result may be the relatively fewer questions in this study.

The respective performances of the chatbots were generally similar among question types, with ChatGPT-4 and ChatGPT-3.5 having the best performance and OcularBERT having the worst performance, though many pairwise comparisons became non-significant after accounting for multiple comparisons. This was particularly apparent in the comparisons between ChatGPT-4 and ChatGPT-3.5 with Google Bard, and between Google Bard and OcularBERT.

It is clinically important to know that all four chatbots made mistakes to varying extents when responding to patient questions. This was seen most frequently for OcularBERT, for which 20% of the responses were Poor. In our evaluation, we see how chatbot responses may be inaccurate and/or misleading. More training may be required before deploying chatbots to answer patient questions.

Google Bard was the only generative chatbot which had responses graded as “Poor”, requiring further evaluation of its ability to correct the response once prompted by the user. Approximately 60% of the responses improved after prompting and without explicit guidance towards the correct answer. While the improvement may not be statistically significant, this indicates it has the capability of response correction, which will likely improve over time with further training and feedback. However, this is a double-edged sword as such correction can be adversely influenced by flawed training data and user feedback.[9] Noteworthily, all responses by Google Bard were prefaced with, “You are right, my previous response was not entirely accurate”, regardless of whether it subsequently generated a more accurate response. It is important to appreciate that the responses of generative chatbots are a result of their assessment of statistical probabilities of associations between words. Such chatbots may still provide inaccurate information after prompting.

Collectively, our results indicate that chatbots have the potential to assist healthcare practitioners in communicating with patients by drafting responses to patient questions that could be reviewed and edited by support staff. This may help to improve productivity, reduce clinic time and unnecessary patient visits, allowing healthcare practitioners to devote time to more complex and higher-yield tasks.

This may also help improve patient experience. Patients can pose questions in natural language and receive accurate and even empathetic-sounding replies expediently.[13] This may in turn help improve health-seeking behaviour, including treatment compliance and adherence to appointments.

While the healthcare roles of chatbots will inevitably increase over time, they need to be deployed judiciously. Humans are naturally drawn to chatbots because they can approximate human speech. This may lead patients to trust and use them for applications for which they were not designed, such as diagnosis and treatment. Chatbots may give incorrect information, particularly for questions that are controversial and require evaluation of incomplete information, amplify social biases in the training data, and may be vulnerable to adversarial attacks. Generative chatbots are trained on general data and are not domain-specific, though this is likely to improve with the development of domain-specific LLMs, which are trained with specialised data to allow understanding of domain-specific terminology and context. An example in the medical domain is the Med-PaLM, which harnesses the power of Google's LLMs and is aligned to the medical domain to answer medical questions more accurately and safely.[36] Chatbots are unable to know and interpret patients' medical data and give personalised answers, though augmented models in the future will be able to draw information from databases.

The main strengths of this study are the comparison of generative and retrieval-based chatbots, and the assessment of their responses to commonly asked questions, which provide novel insights. Several measures were incorporated to improve robustness of the study design: the chatbot identities were masked, the questions and responses were randomised, and there was a 24-hour washout period in between each grading. Assessment of the improvement of chatbot responses after prompting is another novel aspect of the study. Unlike previous research that largely focused on evaluations of responses to standardised examinations,[34,35] our study has evaluated performance in answering patient questions in a real-world context. The findings in this study represents useful information when considering the deployment of chatbots into healthcare roles.

This study had several limitations. Firstly, the questions were randomly selected from data that was originally used to train OcularBERT. However, we noted that the generative chatbots still outperformed OcularBERT, although it had an advantage in this regard. Secondly, only the accuracy of the chatbot responses was evaluated, and we did not study other performance aspects such as cogency, empathy, and personalisation. Thirdly, other questions related to other functions such as triaging, explanation of results, appointment changes and medicine refills, were not assessed. These would be important areas for future assessment. Lastly, our results cannot be generalised to other discipline, as this study focused on the use of specific chatbots in answering patient questions regarding AMD and DR. It would be necessary to assess how chatbots are able to augment conversations between patients and healthcare practitioners in real world settings.

In conclusion, ChatGPT-4 and ChatGPT-3.5 demonstrated superior performance in answering commonly asked questions regarding AMD and DR, followed by Google Bard and OcularBERT. Generative chatbots can effectively answer domain-specific questions outside their original training. Further studies in clinical settings are still required for real-world evaluation.

References

1. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616(7956):259-265. doi:10.1038/s41586-023-05881-4

2. Haupt CE, Marks M. AI-Generated Medical Advice—GPT and Beyond. *JAMA*. 2023;329(16):1349. doi:10.1001/jama.2023.5321

3. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. Drazen JM, Kohane IS, Leong TY, eds. *N Engl J Med*. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184

4. Will ChatGPT transform healthcare? *Nat Med*. 2023;29(3):505-506. doi:10.1038/s41591-023-02289-5

5. Koutsouleris N, Hauser TU, Skvortsova V, De Choudhury M. From promise to practice: towards the realisation of AI-informed mental health care. *The Lancet Digital Health*. 2022;4(11):e829-e840. doi:10.1016/S2589-7500(22)00153-4

6. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. *Nature*. 2023;614(7947):214-216. doi:10.1038/d41586-023-00340-6

7. Van Dis EAM, Bollen J, Zuidema W, Van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023;614(7947):224-226. doi:10.1038/d41586-023-00288-7

8. Temsah MH, Aljamaan F, Malki KH, et al. ChatGPT and the Future of Digital Health: A Study on Healthcare Workers' Perceptions and Expectations. *Healthcare (Basel)*. 2023;11(13):1812. doi:10.3390/healthcare11131812

9. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *Lancet Digit Health*. 2023;5(6):e333-e335. doi:10.1016/S2589-7500(23)00083-3

10. Seth I, Cox A, Xie Y, et al. Evaluating Chatbot Efficacy for Answering Frequently Asked Questions in Plastic Surgery: A ChatGPT Case Study Focused on Breast Augmentation. *Aesthet Surg J*. Published online May 9, 2023:sjad140. doi:10.1093/asj/sjad140

11. Lee TC, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. ChatGPT Answers Common Patient Questions About Colonoscopy. *Gastroenterology*. Published online May 5, 2023:S0016-5085(23)00704-7. doi:10.1053/j.gastro.2023.04.033

12. Rasmussen MLR, Larsen AC, Subhi Y, Potapenko I. Artificial intelligence-based ChatGPT chatbot responses for patient and parent questions on vernal keratoconjunctivitis. *Graefes Arch Clin Exp Ophthalmol*. Published online May 2, 2023. doi:10.1007/s00417-023-06078-1

13. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. Published online April 28, 2023:e231838. doi:10.1001/jamainternmed.2023.1838

14. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. *JAMA Ophthalmol*. Published online April 27, 2023:e231144. doi:10.1001/jamaophthalmol.2023.1144

15. Singh S, Djalilian A, Ali MJ. ChatGPT and Ophthalmology: Exploring Its Potential with Discharge Summaries and Operative Notes. *Semin Ophthalmol*. Published online May 3, 2023:1-5. doi:10.1080/08820538.2023.2209166

16. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health*. 2023;5(3):e107-e108. doi:10.1016/S2589-7500(23)00021-3

17. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: Is ChatGPT a blessing or blight in disguise? *Med Educ Online*. 2023;28(1):2181052. doi:10.1080/10872981.2023.2181052

18. Blanchard F, Assefi M, Gatulle N, Constantin JM. ChatGPT in the world of medical research: From how it works to how to use it. *Anaesth Crit Care Pain Med*. 2023;42(3):101231. doi:10.1016/j.accpm.2023.101231

19. Cao JJ, Kwon DH, Ghaziani TT, et al. Accuracy of Information Provided by ChatGPT Regarding Liver Cancer Surveillance and Diagnosis. *AJR Am J Roentgenol*. Published online May 24, 2023. doi:10.2214/AJR.23.29493

20. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. Published online March 22, 2023. doi:10.3350/cmh.2023.0089

21. Hasnain M, Hayat A, Hussain A. Revolutionizing Chronic Obstructive Pulmonary Disease Care with the Open AI Application: ChatGPT. *Ann Biomed Eng*. Published online May 15, 2023. doi:10.1007/s10439-023-03238-6

22. Lee JH, Jeong MS, Cho JU, et al. Developing a Ophthalmic Chatbot System. In: *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. IEEE; 2021:1-7. doi:10.1109/IMCOM51814.2021.9377398

23. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Published online 2019. doi:10.48550/ARXIV.1908.10084

24. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. Published online December 5, 2017. Accessed June 18, 2023. http://arxiv.org/abs/1706.03762

25. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Published online 2018. doi:10.48550/ARXIV.1810.04805

26. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digit Med*. 2023;6(1):120. doi:10.1038/s41746-023-00873-0

27. Adamopoulou E, Moussiades L. An Overview of Chatbot Technology. In: Maglogiannis I, Iliadis L, Pimenidis E, eds. *Artificial Intelligence Applications and Innovations*. Vol 584. IFIP Advances in Information and Communication Technology. Springer International Publishing; 2020:373-383. doi:10.1007/978-3-030-49186-4_31

28. Flaxman SR, Bourne RRA, Resnikoff S, et al. Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. *The Lancet Global Health*. 2017;5(12):e1221-e1234. doi:10.1016/S2214-109X(17)30393-5

29. Dave AD, Zhu D. Ophthalmology Inquiries on Reddit: What Should Physicians Know? *Clin Ophthalmol*. 2022;16:2923-2931. doi:10.2147/OPTH.S375822

30. Khademi A. Can ChatGPT and Bard Generate Aligned Assessment Items? A Reliability Analysis against Human Performance. Published online 2023. doi:10.48550/ARXIV.2304.05372

31. Destefanis G, Bartolucci S, Ortu M. A Preliminary Analysis on the Code Generation Capabilities of GPT-3.5 and Bard AI Models for Java Functions. Published online 2023. doi:10.48550/ARXIV.2305.09402

32. Koubaa A. *GPT-4 vs. GPT-3.5: A Concise Showdown*. ENGINEERING; 2023. doi:10.20944/preprints202303.0422.v1

33. OpenAI. GPT-4 Technical Report. Published online March 27, 2023. Accessed May 29, 2023. http://arxiv.org/abs/2303.08774

34. Raimondi R, Tzoumas N, Salisbury T, et al. Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. *Eye*. Published online May 9, 2023. doi:10.1038/s41433-023-02563-3

35. Ali R, Tang OY, Connolly ID, et al. *Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations*. Medical Education; 2023. doi:10.1101/2023.03.25.23287743

36. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

Table 1: Length of chatbot responses to age-related macular degeneration and diabetic retinopathy questions

	ChatGPT-4	ChatGPT-3.5	Google Bard	OcularBERT	Pairwise Comparisons						Overall
					ChatGPT-4 versus ChatGPT-3.5	**ChatGPT-4 versus Google Bard**	**ChatGPT-4 versus OcularBERT**	**ChatGPT-3.5 versus Google Bard**	**ChatGPT-3.5 versus OcularBERT**	**Google Bard versus OcularBERT**	
					P-value	**P-value**	**P-value**	**P-value**	**P-value**	**P-value**	**P-value**
Word Count, Mean (SD)	214.0 (57.4)	292.0 (73.4)	183.0 (56.7)	36.5 (23.6)	**7.0x10^{-9}**	5.8x10^{-2}	**<2x10^{-16}**	**1.9x10^{-15}**	**<2x10^{-16}**	**<2x10^{-16}**	**<2.0x10^{-16}**
Minimum	128	61	81	7							
Maximum	320	404	332	118							

Abbreviations – SD: standard deviation

Table 2: Comparison of chatbot performance in answering age-related macular degeneration and diabetic retinopathy questions

	ChatGPT-4	ChatGPT-3.5	Google Bard	OcularBERT	Pairwise Comparisons						Overall
					ChatGPT-4 versus ChatGPT-3.5	**ChatGPT-4 versus Google Bard**	**ChatGPT-4 versus OcularBERT**	**ChatGPT-3.5 versus Google Bard**	**ChatGPT-3.5 versus OcularBERT**	**Google Bard versus OcularBERT**	
					P-value	**P-value**	**P-value**	**P-value**	**P-value**	**P-value**	**P-value**
Total Score											
Total Score, Median (Q1, Q3)	6 (5, 6)	6 (5, 6)	4.5 (4, 6)	2 (2, 3)	0.9	**6.8x10^{-3}**	**1.6x10^{-9}**	**8.4x10^{-3}**	**1.6x10^{-9}**	**8.0x10^{-6}**	**8.0x10^{-13}**
Grade of Responses											
Good, Number (%)	25 (83.3)	26 (86.7)	15 (50.0)	3 (10.0)	1.0	**1.4x10^{-2}**	**5.5x10^{-8}**	**5.5x10^{-3}**	**1.3x10^{-8}**	**1.9x10^{-3}**	**5.8x10^{-9}**
Borderline, Number (%)	5 (16.7)	4 (13.3)	13 (43.3)	21 (70.0)							
Poor, Number (%)	0 (0.0)	0 (0.0)	2 (6.7)	6 (20.0)							

Abbreviations – Q1: first quartile; Q3: third quartile

Table 3: Comparison of chatbot performance in answering commonly asked questions by question type

	ChatGPT-4	ChatGPT-3.5	Google Bard	OcularBERT	Pairwise Comparisons						Overall
					ChatGPT-4 versus ChatGPT-3.5	ChatGPT-4 versus Google Bard	ChatGPT-4 versus OcularBERT	ChatGPT-3.5 versus Google Bard	ChatGPT-3.5 versus OcularBERT	Google Bard versus OcularBERT	
					P-value	**P-value**	**P-value**	**P-value**	**P-value**	**P-value**	**P-value**
AMD											
Total Score											
Total Score, Median (Q1, Q3)	6 (4, 6)	5 (4, 6)	4 (3, 5)	3 (2, 3.5)	0.4	**2.7x10^{-2}**	**1.5x10^{-4}**	8.7x10^{-2}	**2.2x10^{-4}**	**2.7x10^{-2}**	**1.4x10^{-5}**
Grade of Responses											
Good, Number (%)	11 (73.3)	12 (80.0)	6 (40.0)	2 (13.3)	1.0	0.1	**3.2x10^{-3}**	0.1	**9.9x10^{-4}**	0.2	**1.6x10^{-3}**
Borderline, Number (%)	4 (26.7)	3 (20.0)	8 (53.3)	9 (60.0)							
Poor, Number (%)	0 (0.0)	0 (0.0)	1 (6.7)	4 (26.7)							
DR											
Total Score											
Total Score, Median (Q1, Q3)	6 (6, 6)	6 (6, 6)	6 (4, 6)	2 (2, 3)	0.3	0.1	**1.8x10^{-5}**	**2.4x10^{-2}**	**1.7x10^{-5}**	**2.0x10^{-4}**	**6.7x10^{-8}**
Grade of Responses											
Good, Number (%)	14 (93.3)	14 (93.3)	9 (60.0)	1 (6.7)	1.0	0.1	**1.2x10^{-5}**	0.1	**1.2x10^{-5}**	**6.7x10^{-3}**	**1.3x10^{-5}**
Borderline, Number (%)	1 (6.7)	1 (6.7)	5 (33.3)	12 (80.0)							
Poor, Number (%)	0 (0.0)	0 (0.0)	1 (6.7)	2 (13.3)							
Others											
Total Score											
Total Score, Median (Q1, Q3)	6 (5, 6)	6 (5.5, 6)	5 (3.5, 5.5)	3 (3, 5)	0.5	0.1	**4.3x10^{-4}**	**2.7x10^{-2}**	**1.5x10^{-4}**	**4.6x10^{-2}**	**2.4x10^{-5}**
Grade of Responses											
Good, Number (%)	14 (93.3)	14 (93.3)	10 (66.7)	5 (33.3)	1.0	0.2	**2.4x10^{-3}**	0.2	**2.4x10^{-3}**	0.1	**3.6x10^{-4}**
Borderline, Number (%)	1 (6.7)	1 (6.7)	2 (13.3)	9 (60.0)							
Poor, Number (%)	0 (0.0)	0 (0.0)	3 (20.0)	1 (6.7)							

Abbreviations – AMD: age-related macular degeneration; DR: diabetic retinopathy; Q1: first quartile; Q3: third quartile

Table 4: Comparison of total scores and consensus gradings for initial and post-correction responses by Google Bard

Number	Question	Question Type	Total Score		Consensus Grading	
			Initial	Post-Correction	Initial	Post-Correction
5	Is it okay to take a plane after an eye injection?	AMD	3	3	Poor	Poor
28	What is IVTA?	DR	0	1	Poor	Poor
35	Is there a case in which cataract surgery is not possible?	Others	3	4	Poor	Borderline
41	I'm getting glaucoma treatment, can I drink alcohol?	Others	3	4	Poor	Borderline
44	What are the initial symptoms of glaucoma?	Others	3	6	Poor	Good

Abbreviations – AMD: age-related macular degeneration; DR: diabetic retinopathy; IVTA: intravitreal triamcinolone acetonide

Figure Legends

Figure 1. Graphical representation of study methodology. We assessed ChatGPT-4, ChatGPT-3.5, Google Bard, and OcularBERT. A total of 45 questions were compared among the four models. Each chatbot response was generated after clearing the preceding question and answer in a new session. Thereafter, the responses were randomly ordered, stripped of identifying information, and split into four groups. Each set was graded in a separate session, with a 24-hour washout between sessions. The grading was independently performed by three expert graders. The responses were graded using a three-point Likert scale into Good, Borderline, or Poor, for which a score of 2, 1, or 0 were accorded, respectively. The total accuracy score was summed up with a maximum of 6 and minimum of 0 for each chatbot. We also utilised a majority consensus approach, determining the final rating for each chatbot response based on the most common grade among the three graders.

Figure 2. Chatbots performance for AMD and DR. ChatGPT-4 and ChatGPT-3.5 outperformed the other chatbots, both achieving median scores (interquartile range) of 6 (1), compared with 4.5 (2) in Google Bard, and 2 (1) in OcularBERT (all $p≤8.4x10^{-3}$).

Figure 3. Chatbots consensus grading for AMD and DR. Based on the consensus approach, 83.3% of ChatGPT-4's responses and 86.7% of ChatGPT-3.5's responses were rated as "Good", surpassing Google Bard (50%) and OcularBERT (10%) (all $p≤1.4x10^{-2}$). ChatGPT-4 and ChatGPT-3.5 had no “Poor” rated responses, but Google Bard produced 6.7% poor responses and OcularBERT produced 20%.

Supplementary Figure 1. Chatbots performance for AMD. ChatGPT-4, ChatGPT-3.5, and Bard still outperformed OcularBERT. There were also no differences between ChatGPT-4 and ChatGPT-3.5. ChatGPT-4 outperformed Google Bard (6 [2] versus 4 [2], $p=2.7x10^{-3}$]. However, ChatGPT-3.5 and Google Bard had similar performance (5 [2] versus 4 [2]].

Supplementary Figure 2. Chatbots performance for DR. ChatGPT-4, ChatGPT-3.5, and Bard still outperformed OcularBERT. There were also no differences between ChatGPT-4 and ChatGPT-3.5.

ChatGPT-3.5 outperformed Google Bard for DR (6 [0] versus 6 [2], p=2.4x10^{-2}). However, ChatGPT-4 and Google Bard had similar performance (6 [0] versus 6 [2]).

Supplementary Figure 3. Chatbots performance for Others. ChatGPT-4, ChatGPT-3.5, and Bard still outperformed OcularBERT. There were also no differences between ChatGPT-4 and ChatGPT-3.5. ChatGPT-3.5 outperformed Google Bard for DR (6 [0.5] versus 5 [2], p=2.7x10^{-2}). However, ChatGPT-4 and Google Bard had similar performance (6 [1] versus 5 [2]).

Supplementary Figure 4. Chatbots consensus grading for AMD. ChatGPT-4 and ChatGPT-3.5 still had significantly higher proportions of Good responses compared with OcularBERT for all question types. There were also no differences between ChatGPT-4 and ChatGPT-3.5. However, both ChatGPT-4 and ChatGPT-3.5 performed similarly to Google Bard, and Google Bard performed similarly to OcularBERT after accounting for multiple comparisons.

Supplementary Figure 5. Chatbots consensus grading for DR. ChatGPT-4 and ChatGPT-3.5 still had significantly higher proportions of Good responses compared with OcularBERT for all question types. There were also no differences between ChatGPT-4 and ChatGPT-3.5. Google Bard still had a higher proportion of Good responses compared with OcularBERT. However, both ChatGPT-4 and ChatGPT-3.5 performed similarly to Google Bard after accounting for multiple comparisons.

Supplementary Figure 6. Chatbots consensus grading for Others. ChatGPT-4 and ChatGPT-3.5 still had significantly higher proportions of Good responses compared with OcularBERT for all question types. There were also no differences between ChatGPT-4 and ChatGPT-3.5. However, both ChatGPT-4 and ChatGPT-3.5 performed similarly to Google Bard, and Google Bard performed similarly to OcularBERT after accounting for multiple comparisons.