Title: Assessing Large Language Models' Accuracy in Providing Patient Support for Choroidal Melanoma

Rodrigo Anguita,[1,2] Catriona Downie,[1] Lorenzo Ferro Desideri[2], Mandeep S. Sagoo[1,3]

Affiliations:

1: Moorfields Eye Hospital NHS Foundation Trust, City Road London, EC1V 2PD, UK

2: Department of Ophthalmology, Inselspital, University Hospital of Bern, Bern Switzerland

3: NIHR Biomedical Research Centre for Ophthalmology at Moorfields Eye Hospital and University College London Institute of Ophthalmology, London, UK.

Corresponding author:

Mr Rodrigo Anguita

Email: rodrigoanguita@gmail.com

Synopsis/Precis:

Large language models provided accurate answers to most choroidal melanoma patient questions, showing a potential role for patient support. Nevertheless, it is crucial to consider addressing inaccuracies, ethical concerns, specialized fine-tuning, and data protection

ABSTRACT:

Purpose:

This study aimed to evaluate the accuracy of information that patients can obtain from large language models (LLMs) when seeking answers to common questions about choroidal melanoma.

Methods:

Comparative study comparing frequently-asked questions from choroidal melanoma patients and queried 3 major LLMs – ChatGPT 3.5, Bing AI, and DocsGPT. Answers were reviewed by 3 ocular oncology experts and scored as accurate, partially accurate, or inaccurate. Statistical analysis compared the quality of responses across models.

Results:

For medical advice questions, ChatGPT gave 92% accurate responses compared to 58% for Bing AI and DocsGPT. For pre/post-op questions, ChatGPT and Bing AI were 86% accurate while DocsGPT was 73% accurate. There were no statistically significant differences between models. ChatGPT responses were the longest while Bing AI were the shortest, but length did not impact accuracy. All LLMs appropriately directed patients to seek medical advice from professionals.

Conclusion:

LLMs show promising capability to address common choroidal melanoma patient questions at generally acceptable accuracy levels. However, inconsistent, and inaccurate responses do occur, highlighting need for improved fine-tuning and oversight before integration into clinical practice. Ethical and legal considerations around data privacy also need to be addressed given the sensitive nature of medical information.

Introduction:

The diagnosis of eye cancer causes great distress, fear and anxiety for patients [1]. After being diagnosed with cancer, patients often become confused and forget important information communicated in their medical appointments. This is secondary to attentional narrowing and state-dependent learning. Attentional narrowing is when the central message (diagnosis) becomes the primary focus, limiting attention to peripheral information (treatment and prognosis) and state dependency refers to the phenomenon that information acquired in a specific state is best recalled when the individual is in a similar state [2] Moreover, the time expended with patients discussing treatment options and prognosis in medical consultations is often short [3] This leaves patients with many questions that need to be answered after their initial diagnosis.

Good oncological practice encourages cancer patients to become active participants in their own care, and information-seeking has been shown to play a crucial role in helping patients cope with the uncertainties associated with a cancer diagnosis and treatment[4]. When searching for information, the Internet is the preferred source for patients[5] and increasingly it is being used to access resources such as large language models (LLMs). LLMs represent a class of sophisticated artificial intelligence (AI) models trained on vast quantities of textual data [5]. These models are built using deep learning techniques and are designed to understand and mimic human responses by analysing patterns, relationships, and contexts within the fine-tuning data. LLMs can generate coherent and contextually relevant responses to various prompts or queries. The effectiveness of LLMs lies in their capacity to learn intricate linguistic structures, understand contextual nuances, and generate coherent responses based on the patterns gleaned from extensive fine-tuning

datasets [6]. [7] Nowadays it is not uncommon that patients refer to LLMs to seek clinical information. For this reason, several, recent studies have investigated the applicability of LLMs in addressing patients' concerns with different ocular diseases[7-11].

Recently, a study on the applicability of LLMs in answering multiple-choice exam questions in ophthalmology, showed relative low performance of ChatGPT 3.5 in ophthalmic pathology and intraocular tumours in comparison with other subfields [11].

Despite this study, there is still little evidence in the literature on the accuracy and quality of information that patients can obtain from such tools in the field of ocular oncology. Therefore, to evaluate the information that can be obtained from LLMs by the public, we collected the most frequently asked questions from choroidal melanoma patients in our practice and asked these questions to LLMs. We compared answers from three commonly available platforms. The answers from the LLMs were scrutinised by ocular oncology experts for accuracy of information.

Material and methods:

We interrogated the most widely used and open-access LLMs with a set of questions commonly encountered in our clinical practice managing patients with malignant melanoma of the choroid, as well as questions from the clinic email inbox monitored by the specialist ocular oncology nurses. We evaluated ChatGPT 3.5 (OpenAI, San Francisco, CA, USA), Bing AI (powered by GPT-4 and Microsoft) and Docs-GPT Beta (Doximity, San Francisco, USA).

ChatGPT is engineered for chatbot interactions, employing sophisticated natural language processing algorithms to generate responses in conversational contexts. Conversely, Bing

harnesses large LLM technology to enhance search functionality, facilitating the retrieval of pertinent information. Docs-GPT specializes in tasks such as document summarization, content generation, and text comprehension, utilizing advanced language understanding capabilities. The set of questions were classified into two categories: medical advice (Table 1) with 12 queries and pre- and post-operative questions (Table 2) with 15 queries. All the answers given by the LLMs were reviewed by 3 different ocular oncologist specialists and they were scored as:

1) Accurate and sufficient if the content was correct and no important information was missing.

2) Partially accurate and sufficient, if some of the information was incorrect, but this did not affect the overall content of the answer and there was a good amount of information to understand the answer.

3) Inaccurate, when the answer was completely wrong, or a fundamental part of the answer was incorrect.

To avoid memory retention biases, a new session was started for every question for all the 3 LLMs. Subsequently, each question was posed multiple times (at least 3 times) and every corresponding response thoroughly assessed. The reviewers (RA, CD and MSS) were masked to the LLMs they were evaluating and to each other's scores. Where there was a discrepancy in the responses of the reviewers, the majority view among the three reviewers was considered to be correct. All statistical analyses were performed using GraphPad Prism® 8.0.1 Test. A p value of <0.05 was considered statistically significant for all tests. Inter-grader reliability among the reviewers (RA, CD, and MSS) and reliability between triplicate answers were assessed using Cronbach's alpha coefficient.

No human subjects were involved in our study, and the questions used did not include any personal information about patients.

Results:

The three distinct LLMs underwent interrogation using a combined total of twenty-seven questions, which were divided into two categories: medical advice and pre- and post-operative advice (Table 1). In the group of medical advice questions, ChatGPT provided 92% accurate and sufficient answers (11 questions). In comparison, Bing AI and DocsGPT managed to offer accurate and sufficient responses for 58% of the questions (7 questions). Partially accurate and sufficient answers were provided in 42% (5 questions) and 34% (4 questions) in DocsGPT and Bing AI respectively. Both ChatGPT and Bing AI provided inaccurate answers to 8% of the questions (1 question each). No statistically significant difference was found among the 3 different models ($p=0.3$, one-way ANOVA test). Reliability statistics referred a Cronbach's α of 0.914 (95% confidence interval:0.872-0.9.44), showing optimal degree of agreement between the 3 readers.

In the category of pre- and post-operative advice questions; ChatGPT and Bing AI responded accurately and sufficiently in 86% (13) of the questions while DocsGPT in 73% (11) of questions. DocsGPT provided partially accurate and sufficient answers in 20% (3 questions), followed by Bing AI with 14% (2 questions), and ChatGPT with 7% (1 question). Regarding inaccurate answers, both ChatGPT and DocsGPT had 7% (1 each) of their responses classified as inaccurate. No statistically significant difference was found among the 3 different models ($p=0.6$, one-way ANOVA test).

It was revealed that 57% of responses varied across triplicated queries (Cohen's kappa = 0.43, p

< 0.05), indicating moderate variability in the reliability of the LLMs' responses.

The LLM that gave more detailed information was ChatGPT with an average of 274 words per question followed by DocsGPT with 216 words and Bing AI with 126 words. However, this had no impact on the accuracy of the responses to the questions. Importantly, the three examined LLMs clearly stated they were not medical doctors. Their recommendation was to seek advice and treatment from an eye care professional for an accurate diagnosis of symptoms and the most appropriate guidance and care.

Discussion:

A cancer diagnosis can be an immensely distressing experience for patients. Information plays an important role in the development of coping strategies and treatment choices. Effective patient support, including providing accessible information can help in reducing anxiety and stress and contribute to the overall psychological well-being of the patient [4,12,13] From this perspective, LLMs have the capability to help us in tasks such as summarizing topics for patients, responding to patients' queries and emails, and improving communication and understanding.[8] However, before implementing its regular use in clinical practice and exposing patients to the technology, there are many issues that need to be addressed.

Despite our evaluation of the LLMs showing a generally good performance, there were four questions answered inaccurately. Inaccurate information can lead to confusion, misunderstanding, and more importantly, a risk of harm and undue stress. Furthermore, our study revealed a significant 57% variability between responses from triplicate queries, emphasizing the necessity for exploring factors impacting consistency and reliability in LLM-generated

responses. Although studies directly addressing triplicate query variability in LLMs are scarce, related research in natural language processing and machine learning offers valuable insights into assessing reliability and consistency, drawing parallels from studies on inter-rater reliability and model evaluation techniques[14].

A previous study evaluated the accuracy of ChatGPT across various sections of an ophthalmology examination by comparing its outputs with answer keys. The findings revealed superior performance of the legacy model in general medicine, contrasting with its weaker performance observed in neuro-ophthalmology, ocular pathology and oncology. The authors concluded that specializing LLMs through domain-specific pretraining may be needed to improve their performance in ophthalmic subspecialties[15]. Furthermore, it is important to note that the performance of LLMs can vary across different ophthalmic subspecialties. LLMs have been found to give accurate and comprehensive responses to myopia-related queries [16] and appropriate responses that are comparable to physician-written responses in terms of information accuracy [11], but poor performance queries related to vitreoretinal conditions [17]. These differences could be explained by the different depth on data available for each topic on the internet [17]. It is worth considering that ChatGPT was trained on internet data up until January 2022 which could influence its proficiency in certain medical questions. In our study, we found that the 3 different LLMs have an acceptable performance in most of the questions related to choroidal melanoma. Nonetheless, this needs to be thoroughly reviewed and widely tested. Answers from LLMs scoured from the internet may not be totally relevant to practice in different geographical locations, especially as most of the medical literature arises from the Western Hemisphere.

In addition, we believe that appropriate fine-tuning models in ocular oncology need to be developed to improve accuracy and cover more complex topics to make LLMs reliable and

accurate. This will depend on not only fine-tuning the LLMs with depth of information but also improving the capability of the LLMs themselves. Lim et al analysed the performance of Google Bard, ChatGPT 3.5 and ChatGPT 4 answering questions myopia care-related questions. They found that the 3 LLMs were able to deliver accurate and comprehensive responses to myopia-related queries; however, ChatGPT 4.0 demonstrated superior accuracy, self-correction capabilities and performed better in the topic of 'treatment and prevention'[16].

In the context of integrating LLM technology into medical practice, ethical and legal considerations surrounding data collection and sharing are paramount. While our study centered on generic queries, the possibility of patients providing personalized, sensitive health information presents ethical dilemmas. Establishing clear guidelines on data usage and obtaining informed consent is essential to safeguard patient privacy and ensure robust data protection. However, challenges arise regarding the handling and storage of information by commercial entities, necessitating careful regulation and oversight [18,19.] Therefore, authorities should establish clear guidelines for data usage and obtain informed consent for handling sensitive personal information, ensuring robust data protection controls.

Undoubtedly, LLMs hold significant potential as valuable tools in medicine. They can assist us in various tasks like addressing patient inquiries, and enhancing communication between doctors and patients, including for non-English speaking patients by providing translations into native languages, among others.[13] These functions are especially vital in virtual or remote clinics, where patients can seek clarifications following virtual medical evaluations[20,21] Additionally, LLMs offer the advantage of accessibility, providing quick access to answers with no fixed schedule, which is particularly beneficial in remote or isolated areas.

The main challenge for the future is to find the best way to integrate LLMs into everyday

medical practice in a way that is safe and beneficial for patients.

Conflicts of interest

The authors report no conflicts of interest

Author Contribution Statement

RA, and MS conceived and designed the research. RA, CD and MS create the questions. LFD interrogated the LLMs. RA, CD and MS review and classify the answers. RA, CD, LFD and MS analyzed and interpreted the literature. RA, CD, LFD, and MS drafted the manuscript and made critical revision of the manuscript.

Data Availability Statement

The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request.

Funding statement:

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. The research was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Acknowledgement

We thank Sr Sinéad Hanrahan and Sr Nana Gyasi-Twum, Senior Specialist Ocular Oncology Nurses for their help in compiling the lists of questions commonly asked by patients.

References

1. Damato, B.*, et al.* Patient-Reported Outcomes and Quality of Life after Treatment for Choroidal Melanoma. *Ocul Oncol Pathol* **5**, 402-411 (2019).
2. Kessels, R.P. Patients' memory for medical information. *J R Soc Med* **96**, 219-222 (2003).
3. Singh, S.*, et al.* Characterizing the Nature of Scan Results Discussions: Insights Into Why Patients Misunderstand Their Prognosis. *J Oncol Pract* **13**, e231-e239 (2017).
4. Chua, G.P., Tan, H.K. & Gandhi, M. Information sources and online information seeking behaviours of cancer patients in Singapore. *Ecancermedicalscience* **12**, 880 (2018).
5. Madadi, Y.*, et al.* Applications of artificial intelligence-enabled robots and chatbots in ophthalmology: recent advances and future trends. *Curr Opin Ophthalmol* (2024).
6. Contreras Kallens, P., Kristensen-McLachlan, R.D. & Christiansen, M.H. Large Language Models Demonstrate the Potential of Statistical Learning in Language. *Cogn Sci* **47**, e13256 (2023).
7. Ferro Desideri, L., Roth, J., Zinkernagel, M. & Anguita, R. "Application and accuracy of artificial intelligence-derived large language models in patients with age related macular degeneration". *Int J Retina Vitreous* **9**, 71 (2023).
8. Ong, J., Hariprasad, S.M. & Chhablani, J. ChatGPT and GPT-4 in Ophthalmology: Applications of Large Language Model Artificial Intelligence in Retina. *Ophthalmic Surg Lasers Imaging Retina* **54**, 557-562 (2023).
9. Anguita, R., Makuloluwa, A., Hind, J. & Wickham, L. Large language models in vitreoretinal surgery. *Eye (Lond)* (2023).
10. Jin, K., Yuan, L., Wu, H., Grzybowski, A. & Ye, J. Exploring large language model for next generation of artificial intelligence in ophthalmology. *Front Med (Lausanne)* **10**, 1291404 (2023).
11. Bernstein, I.A.*, et al.* Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. *JAMA Netw Open* **6**, e2330320 (2023).
12. Lang, E.V., Berbaum, K.S. & Lutgendorf, S.K. Large-core breast biopsy: abnormal salivary cortisol profiles associated with uncertainty of diagnosis. *Radiology* **250**, 631-637 (2009).
13. Davison, B.J. & Breckon, E.N. Impact of health information-seeking behavior and personal factors on preferred role in treatment decision making in men with newly diagnosed prostate cancer. *Cancer Nurs* **35**, 411-418 (2012).
14. Chen, J.S. & Baxter, S.L. Applications of natural language processing in ophthalmology: present and future. *Front Med (Lausanne)* **9**, 906554 (2022).
15. 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* **3**, 100324 (2023).
16. Lim, Z.W.*, et al.* Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine* **95**, 104770 (2023).
17. Caranfa, J.T., Bommakanti, N.K., Young, B.K. & Zhao, P.Y. Accuracy of Vitreoretinal Disease Information From an Artificial Intelligence Chatbot. *JAMA Ophthalmol* **141**, 906-907 (2023).
18. Li, H.*, et al.* Ethics of large language models in medicine and medical research. *Lancet Digit Health* **5**, e333-e335 (2023).

19. Kleinig, O.*, et al.* How to use large language models in ophthalmology: from prompt engineering to protecting confidentiality. *Eye (Lond)* **38**, 649-653 (2024).
20. Hanumunthadu, D.*, et al.* Outcomes following implementation of a high-volume medical retina virtual clinic utilising a diagnostic hub during COVID-19. *Eye (Lond)* **36**, 627-633 (2022).
21. Shahid, S.M., Anguita, R. & daCruz, L. Telemedicine for postoperative consultations following vitrectomy for retinal detachment repair during the COVID-19 crisis: a patient satisfaction survey. *Can J Ophthalmol* **56**, e46-e48 (2021).

Table 1. Medical advice questions

	ChatGPT 3.5	Bing AI	Docs-GPT Beta
1.is choroidal melanoma a cancer?	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient
2. Did the sun cause choroidal melanoma, what caused this?	Accurate and sufficient	Accurate and sufficient	Partially accurate and sufficient
3. What is the risk of spread of choroidal melanoma?	Accurate and sufficient	Inaccurate	Accurate and sufficient
4. Am I going to die because of choroidal melanoma?	Accurate and sufficient	Partially accurate and sufficient	Partially accurate and sufficient
5. Could I have passed the choroidal melanoma on to my children?	Inaccurate	Partially accurate and sufficient	Partially accurate and sufficient
6. Why me? Is it lifestyle choices resulting in me having a choroidal melanoma in the eye	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient
7. Can the choroidal melanoma transfer to the good eye (untreated eye)?	Accurate and sufficient	Accurate and sufficient	Partially accurate and sufficient
8. How do I communicate to my children that I have eye cancer (choroidal melanoma)?	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient
9. what symptoms should I expect if there is a liver metastasis secondary to my choroidal melanoma?	Accurate and sufficient	Partially accurate and sufficient	Accurate and sufficient
10. How long have I had the choroidal melanoma before diagnosis	Accurate and sufficient	Partially accurate and sufficient	Accurate and sufficient

11. Will my choroidal melanoma spread?	Accurate and sufficient	Accurate and sufficient	Partially accurate and sufficient
12. Is the choroidal melanoma a primary or secondary tumour?	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient

Table 2. Pre and post-operative advice questions

	ChatGPT 3.5	Bing AI	Docs-GPT Beta
1. Can I have choroidal melanoma treatment if the genetic testing shows I have the wrong type of gene?	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient
2. Will I have pain after the enucleation?	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient
3. Will the eye be covered when I am in hospital after the brachytherapy treatment?	Accurate and sufficient	Partially accurate and sufficient	Partially accurate and sufficient
4. am I allowed to drive after radiation or enucleation treatment for choroidal melanoma?	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient
5. Can choroidal melanoma radiation treatment cause eye deformity?	Partially accurate and sufficient	Accurate and sufficient	Partially accurate and sufficient
6. Can I still have sex once I have the choroidal melanoma treatment?	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient
7. How long will I have to stay in hospital after brachytherapy for my choroidal melanoma?	Accurate and sufficient	Partially accurate and sufficient	Accurate and sufficient
8. can my family member stay in the same room with me when I have brachytherapy treatment for my choroidal melanoma?	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient
9. Can the radiation treatment for choroidal melanoma makes me infertile?	Inaccurate	Accurate and sufficient	Inaccurate
10. Would I lose my hair after radiation treatment for choroidal melanoma?	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient
11. When will I know the radiation treatment for choroidal melanoma has worked?	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient
12. How will my vision be affected after radiation treatment for choroidal melanoma?	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient
13. will I be cured after treatment for choroidal melanoma?	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient
14. how can I know if I have metastasis secondary to choroidal melanoma?	Accurate and sufficient	Accurate and sufficient	Partially accurate and sufficient

15. what happens if I develop metastasis secondary to choroidal melanoma?	Accurate and sufficient	Accurate and sufficient	Accurate and sufficient