Original research article

EJO European Journal of Ophthalmology

Utility of an LLM-powered experts-in-the-loop chatbot for pre- and post-operative care of cataract surgery patients

European Journal of Ophthalmology
2026, Vol. 36(3) 561–568

DOI: 10.1177/11206721251396664
journals.sagepub.com/home/ejo
Mary Ann Liebert
A Part of Sage

Bhuvan Sachdeva[1,2,*], Pragnya Ramjee[1,*], Rahul Sharma[1], Mithun Thulasidas[2], Sowmya Raveendra Murthy[2], Geeta Fulari[2], Kaushik Murali[2] and Mohit Jain[1]

Abstract

Purpose: To evaluate the utility of *CataractBot*, an LLM (Large Language Model)-powered chatbot that provides doctor-verified answers to patient questions about cataract surgery. We examine its use by both end-users (patients and attendants) and medical experts.

Methods: A 24-week study was conducted to evaluate *CataractBot* among patients, their attendants, doctors, and patient coordinators. The bot responded instantly to questions by querying a knowledge base curated by medical professionals. Each response was asynchronously verified by an ophthalmologist (for medical questions) or a patient coordinator (for logistical questions), and their edits contributed to updating the knowledge base, thereby minimizing future expert intervention. A mixed-methods analysis was conducted on interaction logs, including patient and attendant questions, chatbot answers, and expert verifications.

Results: A total of 318 patients and attendants sent 1,992 messages, and LLM-generated answers were verified by five doctors and two coordinators. Questions asked pre-surgery were significantly more than post-surgery ($p < 0.001$). Participants asked significantly more medical than logistical questions ($t_{309} = 7.3$, $p < 0.001$). Doctors rated 84.5% of *CataractBot*'s answers to medical questions as accurate and complete. Their edits, which mainly involved adding information, increased the acceptance of the bot's answers by 19.0% over time.

Conclusion: *CataractBot* was predominantly used to address medical questions. It incorporated expert corrections to improve its answers and reduce the experts' bot-related workload over time. This study highlights the potential of LLM-powered chatbots to support patient-provider communication in ophthalmology.

Keywords

WhatsApp Chatbot, GPT-4, Artificial Intelligence, Question Answering Bot, Ophthalmology

Received: 26 May 2025; accepted: 8 October 2025

Introduction

Since their advent in late 2022, Large Language Models (LLMs) have experienced rapid and widespread adoption. OpenAI's ChatGPT, for instance, surpassed 100 million users within two months of its launch, making it the fastest-growing consumer application in history.[1] The success of LLMs is largely attributed to their ease of use, ability to understand natural language, and extensive knowledge base, enabling them to answer a broad range of questions effectively.[2] In healthcare, LLMs have been applied in various domains, including addressing patient queries,[3–6] diagnosis,[7–9] processing electronic health records,[10] summarizing radiology reports,[11] and even training healthcare providers.[12,13] For instance, researchers found that ChatGPT provided accurate and reliable information in most cases when answering questions about pediatric ophthalmology,

[1]Microsoft Research, Bangalore, India
[2]Sankara Eye Hospital, Bangalore, India

*Equal contributor

Corresponding author:
Mohit Jain, Microsoft Research, Bangalore, India.
Email: mohja@microsoft.com

strabismus, and glaucoma.[3,4] However, despite their utility, LLMs have limitations, including hallucinating information, providing incomplete or outdated responses, struggling with complex questions, and exhibiting inconsistencies.[12,14–16] Such issues are concerning especially in healthcare, where accuracy and trustworthiness are critical.

To address these concerns, researchers introduced 'Build Your Own expert Bot' (BYOeB), an open-source platform for developing expert-in-the-loop, LLM-powered chatbots.[17] The platform's first application was *CataractBot*,[18] a WhatsApp-based chatbot designed to help patients and their attendants with queries related to cataract surgery. Unlike generic LLM-powered chatbots, *CataractBot* relied on expert verification for all LLM-generated responses, with doctors reviewing medical answers and patient coordinators handling logistical ones. Expert-provided edits were used to update the custom knowledge base to minimize future expert intervention. A small-scale pilot deployment was previously conducted[18] with 55 users, relying on interview-based methods, which can give rise to participant bias.[19]

In our work, we extend prior research by conducting a large-scale, in-the-wild deployment study of *CataractBot*. We investigate the hypothesis that *CataractBot* can support communication between cataract patients and healthcare providers. Specifically, we examine these research questions: How did the *CataractBot* system perform in terms of the quality of LLM-generated responses and the expert verification process? How did end-users (patients and attendants) and experts (doctors and patient coordinators) interact with *CataractBot*?

Methods

A cross-sectional longitudinal study was conducted between December 2023 - May 2024 at Sankara Eye Hospital, Bengaluru, India, which serves patients from diverse linguistic, educational, and technical backgrounds. Approval from the Scientific and Ethics Committees was obtained prior to the study, and informed consent was obtained from each patient in accordance with the tenets of the Declaration of Helsinki.

As per hospital protocol, once a patient opts for cataract surgery based on a doctor's recommendation, the patient and their attendant meet with a patient coordinator. The coordinator schedules the surgery and provides guidance on pre- and post-operative measures. At the end of this interaction, the coordinator assessed the patient's eligibility for *CataractBot* based on these criteria: aged 18 or above, fluent in one of the five languages supported by *CataractBot* (English, Hindi, Kannada, Tamil, or Telugu), and scheduled for surgery with one of the four operating doctors. If these conditions were met, the coordinator introduced *CataractBot* to address their surgery-related queries. Upon obtaining consent, the coordinator filled a web-based onboarding form. Post-onboarding, participants were instructed to ask a trial question, and the coordinator briefly explained the chatbot's icons and expert verification system. Additionally, participants received reminder messages on WhatsApp at 4pm on five specific days–the day after onboarding, the day before surgery, the surgery day, the day after surgery, and five days post-surgery–reinforcing *CataractBot*'s availability for surgery-related questions.

CataractBot *system*

The features of *CataractBot* have been previously described in detail;[18] a summary is provided here (Figure 1). *CataractBot* supports three interaction modalities: text, speech, and tap. For every voice message, *CataractBot* provides both a text and an audio response. After each response, the bot suggests three follow-up questions randomly generated by an LLM, enabling users to tap and continue the conversation. Upon receiving a message, *CataractBot* classifies it into one of three categories–medical question (e.g., post-surgery care), logistical question (e.g., scheduling), or small-talk (e.g., "Hello")–and responds in real-time. For medical and logistical questions, the bot strictly employs the knowledge base curated by the hospital to generate an appropriate response, which is marked as unverified. If the knowledge base lacks an answer, *CataractBot* provides a template "I don't know" response. For small-talk messages, it provides corresponding small-talk responses.

For each medical question, the operating doctor (Table 1) receives a message containing the question asked, *CataractBot*'s response, and patient's demographics. The doctor is asked, "Is the answer accurate and complete?" with three options: 'Yes', 'No', or 'Send to Patient Coordinator'. Selecting 'Yes' notifies the patient that the answer has been verified. Selecting 'No' alerts the patient to await a corrected response. The doctor is asked to provide a correction in free-form text, which *CataractBot* automatically combines with the bot's initial answer to create a new response, delivered to the patient. If a question is misclassified, i.e., a logistical question is sent to the doctor, they can select 'Send to Patient Coordinator'. (Note: Classification errors were rare, with only 9 questions misclassified; therefore, we excluded them from our analysis.) Patient coordinators follow a similar workflow, verifying and correcting responses for logistical questions.

Edits provided by experts (cataract surgeons and patient coordinators) were used to update the knowledge base, increasing the likelihood of future 'Yes' responses from experts. A senior cataract surgeon, serving as the '*knowledge base expert*' (Table 1), reviewed and selected expert-verified question-answer pairs for inclusion in the knowledge base.

Data analysis

Data entry was conducted using Microsoft Excel with coded variables. Statistical analysis was performed using

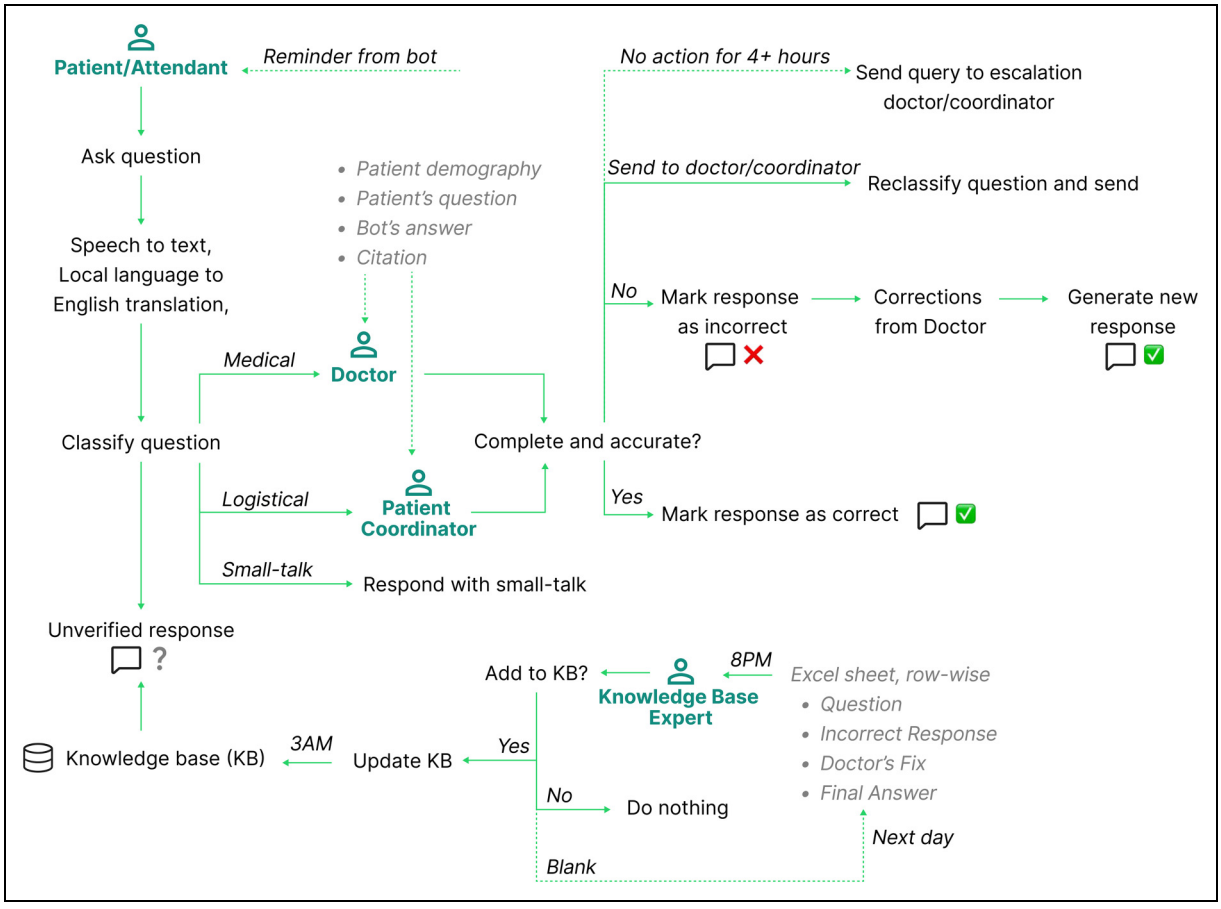

Figure 1. Overview of *CataractBot* features, stakeholder workflows, and component interactions.

Statistical Package for Social Sciences (SPSS-24, IBM). Quantitative analysis included descriptive statistics, t-tests, and (RM-)ANOVAs with checks for normality. A qualitative thematic analysis was conducted through open-coding by a single researcher. As a result, inter-rater agreement was not applicable. The analysis covered patient questions, *CataractBot*'s responses, and expert edits made by both the operating experts and the knowledge base update expert. For all statistical tests, a p-value less than 0.05 was considered statistically significant.

Results

Although 550 individuals were onboarded to *CataractBot*, 318 (57.8%) sent at least one message, forming the participant set.[i] This group included 154 patients and 164 attendants, comprising a total of 271 patient-attendant pairs. A notable demographic difference was that attendants were generally younger, more fluent in English, and better educated compared to patients (Table 2). Additionally, the study involved five doctors (four operating doctors and one as knowledge base update expert) and two patient coordinators. Our dataset comprised 1,992 messages sent by patients and attendants, with LLM-generated answers verified by experts.

System performance

Most LLM-generated responses were found to be accurate and complete for both medical (84.5%) and logistical (69.5%) questions, excluding those without expert verification (Table 3). Corrections were provided for 187 (14.8%) medical and 114 (28.1%) logistical answers. On average, the chatbot responded in 9.3±4.9 seconds. *CataractBot* provided an "I don't know" response for 9.3% of medical questions and 23.3% logistical questions. Further manual analysis revealed that 37.5% of these responses resulted from gaps in the knowledge base, while 29.4% were due to patient-specific questions that required access to patient's medical records. Over the 24-week study, as the knowledge base was updated, the proportion of "I don't know" responses decreased by 7.8% (Figure 2(a)), and the number of LLM-generated answers marked as 'accurate and complete' by experts increased by 19.0%. Manual analysis identified hallucinated content–defined as information not present in the retrieved documents[20] in only five answers (0.3%).

Table 1. Stakeholders and their roles in the *CataractBot* socio-technical system.

Stakeholder			Role
End-user	Patient		Person scheduled for cataract surgery. Asks *CataractBot* surgery-related questions.
	Attendant		Person accompanying the patient (e.g., child of the patient). Asks *CataractBot* surgery-related questions.
Expert	Doctor	Operating Doctor	Surgeon scheduled to operate on the patient. Verifies *CataractBot*'s answers to users' medical questions.
		Knowledge Base Expert	Senior surgeon. Selects and edits verified answers for addition to *CataractBot*'s knowledge base.
	Patient Coordinator	Operating Coordinator	Liaison between patients/attendants and operating doctors. Verifies *CataractBot*'s answers to users' logistical questions.

Table 2. Demography and usage details of study participants.

Data collected	Patients (n=154)	Attendants (n=164)
Age (years)	63.8±9.7	37.9±10.7
Gender, n (%)	73 female (47.4%)	46 female (29.8%)
Education, n (%)	22 ≤ Grade 10 (14.3%) 18 Grade 12 (11.7%) 49 Bachelors (31.8%) 10 Masters (6.5%) 55 Unknown (35.7%)	6 ≤ Grade 10 (3.7%) 4 Grade 12 (2.4%) 25 Bachelors (27.4%) 47 Masters (28.7%) 1 PhD (0.6%) 61 Unknown (37.2%)
Language, n (%)	120 English (77.9%) 11 Kannada (7.1%) 10 Hindi (6.5%) 7 Tamil (4.5%) 6 Telugu (3.9%)	148 English (90.2%) 7 Kannada (4.3%) 3 Hindi (1.8%) 2 Tamil (1.2%) 4 Telugu (2.4%)
Message Type, n (%)	653 Medical (62.4%) 272 Logistical (26.0%) 121 Small-talk (11.6%)	615 Medical (65.0%) 257 Logistical (27.2%) 74 Small-talk (7.8%)
Message Modality, n (%)	618 Text (59.1%) 357 Tap (34.1%) 71 Audio (6.8%)	544 Text (57.5%) 379 Tap (40.1%) 23 Audio (2.4%)
Message Language, n (%)	878 English (83.9%) 39 Kannada (3.7%) 46 Hindi (4.4%) 59 Tamil (5.6%) 24 Telugu (2.3%)	850 English (89.9%) 23 Kannada (2.4%) 18 Hindi (1.9%) 15 Tamil (1.6%) 40 Telugu (4.2%)

Patient and attendant interaction

Participants asked significantly more medical questions (4.1±5.6 questions/participant) compared to logistical questions (1.7±2.2 questions/participant) ($t_{309} = 7.3, p < 0.001$). This confirms that *CataractBot* was primarily used for medical concerns. The most common medical questions related to 'Post-surgery dos and don'ts' (11.5% of messages), 'Medication' (7.4%), and 'Surgery preparation' (6.4%). Among logistical questions, the top three were 'Surgery schedule' (12.0%), 'Hospital contact number' (5.3%), and 'Appointment scheduling' (4.4%).

A one-way RM-ANOVA of interaction modalities found a significant difference ($F_{2,618} = 66.7, p < 0.001$), with text (3.2±3.6) being the most used modality, significantly

Table 3. Summary statistics of LLM-generated responses, verification and edits by experts (doctors and coordinators), and acceptance and edits by the knowledge base expert.

	LLM response	Expert verification and edit	Knowledge base expert edit
Medical: 1268 (63.6%)	Valid Ans: 1150 (90.7%)	Yes: 1033 (89.8%). No: 112 (10.8%), Edit Dist: 44.7%. Ignored: 5 (0.4%).	Yes (without edit): 47 (42.0%). Yes (with edit): 31 (27.7%), Edit Dist: 25.3%. No: 14 (12.5%). Ignored: 20 (17.8%).
	IDK: 118 (9.3%)	Yes: 32 (27.1%). No: 83 (70.3%), Edit Dist: 75.4%. Ignored: 3 (2.6%).	Yes (without edit): 30 (36.2%). Yes (without edit): 30 (36.2%). No: 22 (26.5%). Ignored: 9 (10.8%).
Logistical: 529 (26.6%)	Valid Ans: 406 (76.7%)	Yes: 266 (65.5%). No: 62 (15.3%), Edit Dist: 57.7%. Ignored: 78 (19.2%).	Yes (without edit): 11 (17.7%). Yes (with edit): 14 (22.6%), Edit Dist: 63.2%. No: 25 (40.3%). Ignored: 12 (19.4%).
	IDK:123 (23.3%)	Yes: 16 (13.0%). No: 62 (50.4%), Edit Dist: 70.3%. Ignored: 45 (36.6%).	Yes (without edit): 6 (9.7%). Yes (with edit): 28 (45.2%), Edit Dist: 70.3%. No: 25 (40.3%). Ignored: 3 (4.8%).
Small-talk: 195 (9.8%)	NA		

more than taps (2.4±4.4) and speech (0.3±1.1) ($p < 0.01$). Taps were also preferred over speech ($p < 0.001$).

Regarding temporal trends, the highest number of questions were asked on the day before surgery (Figure 2(b, c)), particularly logistical questions (38.2%) related to the next day's schedule. Medical questions were distributed more evenly across the pre- and post-surgery periods, primarily within 7 days (±3 day) of surgery (Figure 2(b)). In contrast, logistical questions were concentrated within 3 days (±1 day) of surgery, peaking sharply on the day before surgery (Figure 2(c)). Questions asked pre-surgery (2.7±3.2 questions/participant) were significantly more than post-surgery (1.2±2.8) and on-surgery (0.9±2.0) ($p < 0.001$).

Although no statistically significant correlation was found between the day of the week and the number of questions asked, Saturday had the highest query volume. This may be attributed to Saturday being a working day for medical staff but a holiday for most patients and attendants, allowing end-users more time at home to engage with the chatbot. Time-of-day analysis indicated a significant effect on the number of questions ($F_{3,918} = 9.0$, $p < 0.001$). Participants asked significantly more questions in the evening 3pm-6pm (1.9±2.7) compared to the afternoon 11am-2pm (1.2±2.3) and night 7pm-10pm (1.0±2.6) ($p < 0.01$). This evening peak aligns with the *CataractBot*'s 4pm reminder message.

No significant difference was found in the number of questions asked by males and females. However, education level showed a marginally significant effect: participants with a Bachelor's degree or higher asked more questions (6.2±0.5 questions/participant) than those with 12th grade education or below (4.3±0.8) ($t_{197} = 1.9$, $p = 0.06$). Additionally, no significant differences were observed across different languages, supporting the decision to offer multilingual functionality.

Expert interaction

Of the 1797 bot-generated answers reviewed by experts, 75.0% were marked as 'Yes' for being 'accurate and complete', while 17.8% were marked as 'No', with nearly half (45.5%) being "I don't know" responses (Table 3). Experts more often marked logistical responses as incorrect (30.5%) compared to medical responses (15.5%), primarily due to patient-specific logistical questions (19.9%). In contrast, only 9.4% of incorrect answers pertained to patient-specific medical questions. The bot lacked access to patient's health records, limiting its ability to provide case-specific guidance. Also, patient-specific corrections could not be added to the knowledge base as they were not universally applicable.

A thematic analysis of expert corrections revealed nine distinct types, with the top four described here. First, the most frequent correction type was 'Adding new information' for both medical (65.9%) and logistical (30.5%) questions, wherein experts addressed gaps in the bot's knowledge. It contributed to resolving 49.3% of "I don't know" responses. Second, experts performed 'Factual corrections' more frequently for medical answers (7.8%) than logistical answers (2.1%). These medical corrections often reflected expert-specific preferences rather than outright inaccuracies. Third, experts included 'Clarifying questions' in 9.4% of corrections, primarily because 22.4% of the bot's "I don't know" responses resulted from unclear or incomplete questions. Fourth, experts responded with 'Redirection' (4.2%), wherein experts recommended patients to visit the hospital for further evaluation.

The knowledge base update expert reviewed the experts' corrections, and 73.5% were approved for inclusion in the knowledge base. Of these, 49.7% were accepted without

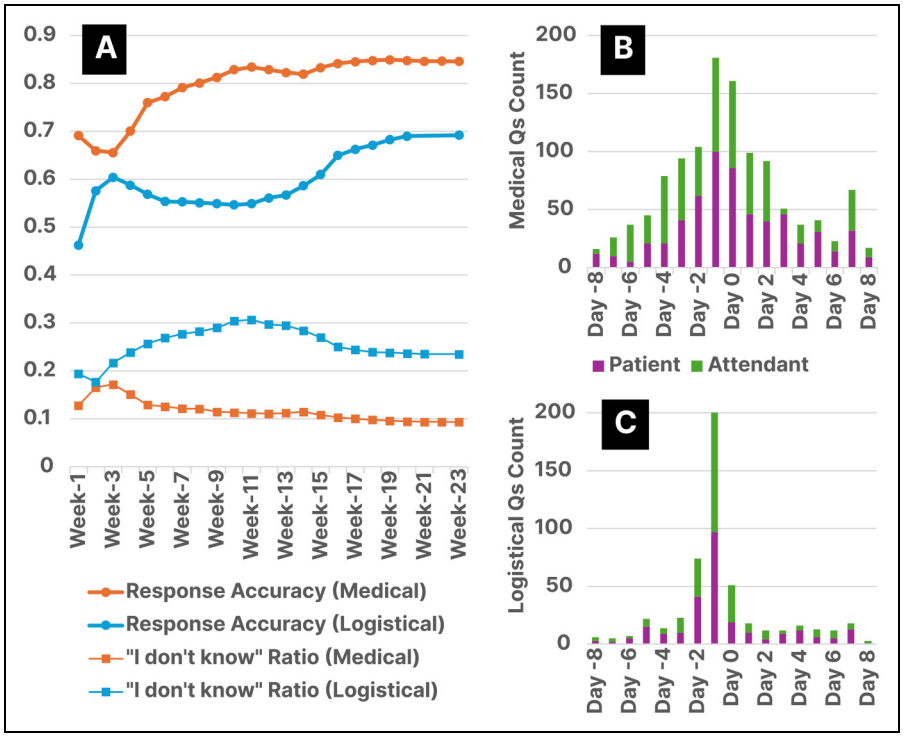

Figure 2. (A) Distribution of accurate and 'I don't know' responses by *CataractBot*. (B) and (C) Distribution of medical and logistical questions asked by patients and attendants relative to the day of surgery.

modifications. For the rest, we identified two key types of edits. First, the most frequent modification was 'Adding new information' (78.9%). Second, in 13.3% of cases, the update expert 'Generalized information' by replacing specific dates, times, or staff names, with broadly applicable terms relevant to all patients. These findings highlight the iterative nature of expert involvement in refining the bot's responses.

Discussion

Prior studies have investigated the role of chatbots in responding to patient questions in ophthalmology.[3–6,21–26] To mitigate the risks of errors, hallucinations, and bias[12,14] in AI, recently, a human-AI collaboration approach was proposed with *CataractBot*,[18] where LLM-generated answers to healthcare queries were verified by experts. That study[18] included a small-scale pilot with 55 users. In this work, we strengthen and expand their findings by conducting a large-scale, in-the-wild 24-week deployment study of *CataractBot* with 318 patients and attendants, examining its use and performance in a real clinical setting.

We present several novel findings. First, while the original study[18] observed that attendants, being younger and tech-savvy, asked more questions than patients, we found no significant difference in the number of questions asked between the two groups. This suggests that the bot was equally accessible to all users and its WhatsApp-based interface engaged older patients effectively. Second, contrary to the earlier study's hypothesis that less educated, non-english speakers would prefer audio messages in Indic languages,[18] we found audio to be the least preferred modality across user groups. Moreover, we found no significant correlation between language choice and modality, or between education level and modality, suggesting significant scope for improvements in translation and transcription technologies to improve accessibility. Third, we found the highest number of both medical and logistical questions were asked on the day before surgery, rather than on the day of surgery as earlier reported.[18] This highlights the importance of pre-surgery informational support, especially where direct access to hospital staff may be limited. Finally, we contribute a qualitative analysis of expert edits to *CataractBot*'s responses, revealing that most corrections involved adding new information, with repetitions and contradictions emerging as the knowledge base expanded. Below, we discuss design implications for LLM-powered doctors-in-the-loop chatbots.

Patients and attendants often sought similar types of information at specific times relative to their surgery date. For future deployments of similar informational chatbots, we recommend proactively engaging with different end-users via push notifications at key moments, delivering relevant information based on observed information-seeking behaviour. This approach could replace vanilla reminder messages and potentially improve user engagement.

We observed that chatbot conversations often failed due to unclear or incomplete questions, leading to "I don't know" responses, and causing experts to struggle with providing answers without additional context or follow-up information. To address this issue, future bots should incorporate mechanisms to identify ambiguous queries, potentially using a Small Language Model to minimize cost and latency. Such bots should then facilitate multi-turn conversations to gather necessary clarifications or enable users to build complex queries, before relaying the exchange to an expert for verification. Additionally, *CataractBot* currently considers only the last two queries when generating responses. Prior work[2] highlights the benefits of integrating long-term conversational history into LLM systems, enabling proactive follow-ups and personalized recommendations. However, incorporating a long conversation history is challenging due to token limitations in LLM inputs. As an alternative, generating conversational summaries could be explored to retain context efficiently.

Our deployment revealed instances where different answers to similar questions were added to the knowledge base. This inconsistency led to varying responses for patients, as *CataractBot* generated answers based on three selected data chunks from the knowledge base. To address this issue, we recommend that knowledge base update experts have a structured overview of existing content before adding new content, ensuring consistency and reducing redundancy. Additionally, previously rejected questions should not be resubmitted to the update expert. We also noticed that these discrepancies in responses often stemmed from differences in individual doctors' recommendations. To avoid contradictions, we suggest creating doctor-specific partitions within the knowledge base, enabling doctors to provide personalized recommendations while maintaining overall coherence. This decentralization of knowledge base control would reduce the burden on a single update expert and democratize such doctor-in-the-loop systems to accommodate diverse medical opinions.

This study has a few limitations. First, the reliance on manual explanations from patient coordinators may have contributed to some onboarded participants not engaging with the bot, resulting in their exclusion from the study. Also, the accuracy of participant data, such as phone numbers, could not be ensured during onboarding. Second, this study lacks a comparative baseline and a control group, which limits our ability to comprehensively evaluate *CataractBot*'s impact on patient satisfaction and the workload of healthcare professionals compared to existing patient support methods. Further randomized controlled trials with a larger sample size may provide more practical results.

In conclusion, *CataractBot* serves as a 24/7, accessible resource, providing cataract patients with expert-verified answers with minimal addition to ophthalmologists' workload. The bot accurately classified questions, generated responses, and incorporated expert corrections. Over time, experts rated an increasing proportion of its answers as accurate and complete, demonstrating the system's ability to improve through iterative updates. While LLM-based systems are not designed to replace human ophthalmologists, they hold potential to augment their work by enhancing patient education and streamlining communication under appropriate supervision. We hope these insights will inform the design and development of LLM-powered doctor-in-loop chatbots in the field of ophthalmology and beyond.

Acknowledgments

Many thanks to the Sankara Eye Hospital staff for their time and patience.

ORCID iDs

Bhuvan Sachdeva https://orcid.org/0009-0002-1946-684X
Pragnya Ramjee https://orcid.org/0000-0003-0061-2624
Rahul Sharma https://orcid.org/0009-0007-7533-1794
Mithun Thulasidas https://orcid.org/0000-0002-0623-4612
Sowmya Raveendra Murthy https://orcid.org/0000-0003-1037-3583
Geeta Fulari https://orcid.org/0009-0003-2358-810X
Kaushik Murali https://orcid.org/0000-0002-1385-3227
Mohit Jain https://orcid.org/0000-0002-7106-164X

Ethical considerations

This study was approved by the Institutional Ethics Committee of Sankara Eye Hospital Bangalore (Ethics Code Approval Number: SEH/BLR/EC/2023/93) on July 8th, 2023. This research was conducted ethically in accordance with the World Medical Association Declaration of Helsinki.

Consent to participate

All participants provided written informed consent prior to participating.

Consent for publication

Not applicable.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interest

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Any individual associated with Microsoft Research, India or Sankara Eye Hospital, India has a potential conflict of interest with respect to the research, authorship, and/or publication of this article.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Note

i. Among the remaining 232 individuals, there were 86 unique patient-attendant pairs in which neither person used the bot. For all other cases, at least one member of the pair–either the patient or the attendant–used the bot to ask one or more questions. No significant demographic differences were observed between users and non-users of the bot.

References

1. Gordon C. ChatGPT is the fastest-growing app in the history of web applications. *Forbes* 2023. https://www.forbes.com/sites/cindygordon/2023/02/02/chatgpt-is-the-fastest-growing-ap-in-the/-history-of-web-applications/.
2. Yang Z, Xu X, Yao B, et al. Talk2Care: An LLM-based voice assistant for communication between healthcare providers and older adults. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2024; 8: 1–35. DOI: 10.1145/3659625.
3. Ahmed HS and Thrishulamurthy CJ. Evaluating ChatGPT's efficacy and readability to common pediatric ophthalmology and strabismus-related questions. *Eur J Ophthalmol* 2025; 35: 466–473.
4. Dogan L and Ibrahim Edhem Y. The performance of ChatGPT-4 and bing chat in frequently asked questions about glaucoma. *Eur J Ophthalmol* 2025; 0: 11206721251321197.
5. Gurnani B, Kaur K, Gireesh P, et al. Evaluating the novel role of ChatGPT-4 in addressing corneal ulcer queries: An AI-powered insight. *Eur J Ophthalmol* 2025; 0: 11206721251337290.
6. Sharma R, Ramjee P, Murali K, et al. Editing with AI: How doctors refine LLM-generated answers to patient queries. In: *The second workshop on GenAI for health: Potential, trust, and policy compliance*. https://openreview.net/forum?id=WyaJ7jBTed.
7. Danilov G, Kotik K, Shevchenko E, et al. Length of stay prediction in neurosurgery with Russian GPT-3 language model compared to human expectations. In: *Informatics and technology in clinical care and public health*, 2022, pp.156–159. IOS press.
8. Nguyen T, Ong J, Jonnakuti V, et al. Artificial intelligence in the diagnosis and management of refractive errors. *Eur J Ophthalmol* 2025; 0: 11206721251318384.
9. Mohammadpour M, Heidari Z, Hashemi H, et al. Comparison of artificial intelligence-based machine learning classifiers for early detection of keratoconus. *Eur J Ophthalmol* 2022; 32: 1352–1360.
10. Wang SY, Huang J, Hwang H, et al. Leveraging weak supervision to perform named entity recognition in electronic health records progress notes to identify the ophthalmology exam. *Int J Med Inform* 2022; 167: 104864.
11. Yan A, McAuley J, Lu X, et al. Radbert: Adapting transformer-based language models to radiology. *Radiol: Artif Intell* 2022; 4: e210258.
12. Gurnani B and Kaur K. Leveraging ChatGPT for ophthalmic education: A critical appraisal. *Eur J Ophthalmol* 2024; 34: 323–327.
13. Ramjee P, Chhokar M, Sachdeva B, et al. ASHABot: An LLM-powered chatbot to support the informational needs of community health workers. In: *Proceedings of the 2025 CHI conference on human factors in computing systems*. *CHI '25*, New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/3706598.3713680.
14. Au Yeung J, Kraljevic Z, Luintel A, et al. AI chatbots not yet ready for clinical use. *Front Digit Health* 2023; 5: 1161098. DOI: 10.3389/fdgth.2023.1161098.
15. Denecke K, May R and Rivera Romero O. Potential of large language models in health care: Delphi study. *J Med Internet Res* 2024; 26: e52399.
16. Fogel AL and Kvedar JC. Artificial intelligence powers digital medicine. *npj Digit Med* 2018; 1: 5.
17. Microsoft. BYOeB: Build Your Own expert Bot. https://github.com/microsoft/byoeb (2024, accessed 18 June 2024).
18. Ramjee P, Sachdeva B, Golechha S, et al. CataractBot: An LLM-powered expert-in-the-loop chatbot for cataract patients. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2025; 9: 1–31. DOI: 10.1145/3729479.
19. Dell N, Vaidyanathan V, Medhi I, et al. "Yours is better!": Participant response bias in HCI. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, *CHI '12*, 2012, pp.1321–1330. New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/2207676.2208589.
20. Ayala O and Bechard P. Reducing hallucination in structured outputs via retrieval-augmented generation. In: *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (Volume 6: Industry Track)*, 2024, p.228–238. Association for Computational Linguistics. DOI: 10.18653/v1/2024.naacl-industry.19.
21. Bernstein IA, Zhang YV, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Network Open* 2023; 6: e2330320.
22. Azzopardi M, Ng B, Logeswaran A, et al. Artificial intelligence chatbots as sources of patient education material for cataract surgery: ChatGPT-4 versus google bard. *BMJ Open Ophthalmol* 2024; 9: e001824.
23. Su Z, Jin K, Wu H, et al. Assessment of large language models in cataract care information provision: A quantitative comparison. *Ophthalmol Ther* 2025; 14: 103–116.
24. Betzler BK, Chen H, Cheng CY, et al. Large language models and their impact in ophthalmology. *Lancet Digit Health* 2023; 5: e917–e924.
25. Thompson P, Thornton R and Ramsden CM. Assessing chatbots ability to produce leaflets on cataract surgery: Bing AI, ChatGPT 3.5, ChatGPT 4o, ChatSonic, Google Bard, Perplexity, and Pi. *J Cataract Refract Surgery* 2025; 51: 371–375.
26. Ittarat M, Cheungpasitporn W and Chansangpetch S. Personalized care in eye health: Exploring opportunities, challenges, and the road ahead for chatbots. *J Pers Med* 2023; 13: 1679. DOI: 10.3390/jpm13121679.