

CataractBot: An LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients

PRAGNYA RAMJEE*, Microsoft Research, India

BHUVAN SACHDEVA*, Microsoft Research, India

SATVIK GOLECHHA*, Microsoft Research, India

SHREYAS KULKARNI, Microsoft Research, India

GEETA FULARI, Sankara Eye Foundation India, India

KAUSHIK MURALI, Sankara Eye Foundation India, India

MOHIT JAIN, Microsoft Research, India

The healthcare landscape is evolving, with patients seeking reliable information about their health conditions and available treatment options. Despite the abundance of information sources, the digital age overwhelms individuals with excess, often inaccurate information. Patients primarily trust medical professionals, highlighting the need for expert-endorsed health information. However, increased patient loads on experts has led to reduced communication time, impacting information sharing. To address this gap, we developed *CataractBot*¹. *CataractBot* answers cataract surgery related questions instantly using an LLM to query a curated knowledge base, and provides expert-verified responses asynchronously. It has multimodal and multilingual capabilities. In an in-the-wild deployment study with 49 patients and attendants, 4 doctors, and 2 patient coordinators, *CataractBot* demonstrated potential, providing anytime accessibility, saving time, accommodating diverse literacy levels, alleviating power differences, and adding a privacy layer between patients and doctors. Users reported that their trust in the system was established through expert verification. Broadly, our results could inform future work on designing expert-mediated LLM bots.

CCS Concepts: • Human-centered computing → Interaction design; Ubiquitous and mobile computing systems and tools; • Applied computing → Health care information systems.

Additional Key Words and Phrases: GPT-4, Generative AI, LLM, Question Answering Bot, Medical, Healthcare, Surgery

ACM Reference Format:

Pragnya Ramjee, Bhuvan Sachdeva, Satvik Golechha, Shreyas Kulkarni, Geeta Fulari, Kaushik Murali, and Mohit Jain. 2025. CataractBot: An LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients. 1, 1 (February 2025), 30 pages. <https://doi.org/xx.xxxx/xxxxxx.xxxxxx>

1 INTRODUCTION

The evolving landscape of healthcare witnesses a significant shift with patients assuming more proactive roles in their care journeys [46, 47], resulting in an increased demand for information. Patients and their caregivers seek accessible, comprehensive, and reliable information about their symptoms, diagnoses, treatment options,

*Equal contribution.

¹Our code is publicly available at <https://github.com/microsoft/BYOeB>.

Authors' addresses: Pragnya Ramjee, Microsoft Research, Bangalore, India, t-pramjee@microsoft.com; Bhuvan Sachdeva, Microsoft Research, Bangalore, India, b-bsachdeva@microsoft.com; Satvik Golechha, Microsoft Research, Bangalore, India, zsatvik@gmail.com; Shreyas Kulkarni, Microsoft Research, Bangalore, India, shreyaskulkarni.sak@gmail.com; Geeta Fulari, Sankara Eye Foundation India, Bangalore, India, quality@sankaraeye.com; Kaushik Murali, Sankara Eye Foundation India, Bangalore, India, kaushik@sankaraeye.com; Mohit Jain, Microsoft Research, Bangalore, India, mohja@microsoft.com.

potential risks, and preventive measures [18]. To satisfy these information needs, patients actively explore diverse sources, including online resources, friends and family, support groups, and direct communication with healthcare professionals [78]. In particular, for major treatments like surgery, satisfying the demand for information becomes crucial [18], with questions spanning pre-, during-, and post-treatment phases. Studies highlight the anxiety patients and caregivers experience regarding such treatments [48], emphasizing the correlation between anxiety and negative clinical outcomes [14, 23]. Similarly, access to information has been found to significantly reduce anxiety [94] and improve patient satisfaction and clinical outcomes [51, 94], thus underlining its pivotal role.

Despite the abundance of available information sources, the digital age often overwhelms individuals with an excess of information, much of which is inaccurate or unreliable [21, 46]. This poses a challenge in discerning trustworthy sources from misleading ones. Patients also frequently encounter difficulties finding information online at an appropriate level, ranging from oversimplified to overly technical [95]. As a result, patients tend to rely on the experiences of friends and family members who have undergone similar treatments, although identifying them can be challenging [20]. Therefore, patients primarily trust their doctor and the hospital staff responsible for their treatment, turning to them for any medical or logistical queries [6, 21]. Studies examining patients' information needs reveal a strong desire for doctor-endorsed health information [83].

However, the escalating pressure on doctors to accommodate more patients has led to reduced time per patient, affecting communication and information sharing. In developing nations, a lower doctor-to-patient ratio hinders personalized attention and comprehensive guidance [91], while in developed countries, managing Electronic Health Records (EHR) often diminishes direct patient communication quality and available time [15]. Previous studies [46, 87] have found that doctors often underestimate patients' information needs and overestimate the amount of information they provide. However, providing all potentially relevant information risks overwhelming patients with cognitive overload [18]. Surveys consistently highlight patients' desire for improved communication with their healthcare providers [27]. This information exchange extends beyond doctor-patient interactions to encompass relationships between various medical professionals and patients, such as patient-nurse interactions [57, 90]. Moreover, enhanced access to information not only benefits patients but also has the potential to reduce unnecessary visits, alleviating the burden on doctors and medical staff. To summarize, as Tang and Lansky [82] correctly stated: "*Patients have little access to information and knowledge that can help them participate in, let alone guide, their own care... A simple, non-urgent exchange of questions and answers is often all that is required.*".

To address patients' information needs, particularly during surgical treatments, we propose a chatbot solution integrated with medical professionals to provide accessible, reliable and comprehensive responses. Leveraging the capabilities of large-language models (LLMs) and the widespread use of smartphones and instant messaging services such as WhatsApp, this chatbot serves as a 24/7 resource capable of understanding intricate human queries and providing accurate information. In collaboration with a tertiary eye hospital in Bangalore, India, we exemplified this experts-in-the-loop chatbot approach by developing *CataractBot*, which answers queries related to cataract surgery from patients and their attendants. Recognizing the limitations of generic LLMs in the medical domain, we utilized Retrieval-Augmented Generation [53] with an LLM over a custom knowledge base curated by doctors and hospital staff to provide hospital-specific and culturally sensitive responses. *CataractBot* has several features to enhance adoption and sustained use. Capitalizing on the widespread adoption of smartphones, we chose a mobile-first design paradigm, deploying the chatbot entirely on the ubiquitous WhatsApp platform. To cater to diverse hospital visitors with varying literacy levels and technological proficiency, we made *CataractBot* multilingual (supporting five languages, including English, Hindi, and Kannada) and multimodal (accepting both speech and text inputs). We use an LLM to process complex and ill-formed queries, generating instant responses from the curated knowledge base. Crucially, *CataractBot* incorporates a system where the LLM's responses are asynchronously verified by experts and corrected if needed. Ophthalmologists verify answers to medical queries and patient coordinators to logistical ones. These expert-provided edits are used to update the knowledge base, to minimize future expert intervention.

Our work aims to answer the following research questions: **(RQ1)** What is the role of *CataractBot* in meeting the information needs of patients (and attendants) undergoing surgery? **(RQ2)** How do the different features of *CataractBot*, including LLM-generated answers, experts-in-the-loop and multimodality, contribute to its usage? **(RQ3)** How can *CataractBot* be integrated into the current doctor-patient workflow? To answer these questions, we conducted an in-the-wild deployment study involving multiple stakeholders (See Table 1): 49 information seekers (19 patients, 30 attendants) and 6 experts (4 doctors and 2 patient coordinators). Our findings indicated that patients and attendants appreciated *CataractBot* because it alleviated hesitations associated with power differences in asking questions, and accommodated individuals with low literacy and tech proficiency through multilingual and multimodal support. Patients and attendants reported that their trust in the system was established through the verification performed by doctors and coordinators. Experts praised the bot as a facilitator, introducing a privacy layer between them and patients, and providing flexibility through asynchronous communication.

Our main contributions are: (1) The design and development of a novel LLM-powered, experts-in-the-loop, multilingual, multimodal chatbot, created in collaboration with doctors and hospital staff to assist cataract surgery patients with their information needs. (2) The deployment of our bot in a real-world multi-stakeholder setting, among 49 cataract surgery patients/attendants, four doctors, and two patient coordinators at a tertiary eye hospital in India. (3) Drawing from a comprehensive mixed-methods analysis of gathered data, the paper offers pivotal insights specific to *CataractBot* and broader implications of LLM-powered experts-in-the-loop chatbots.

2 RELATED WORK

Our work is primarily informed by two areas of relevant research: solutions addressing patients' information needs and the use of conversational agents, particularly those powered by LLMs, in healthcare settings.

2.1 Patients' Information Needs

Patients and their attendants seek information throughout the various stages of their medical journey, such as preparing for doctor consultations [52, 74], verifying diagnoses and treatment plans [10, 18, 29], assessing the need for clinical interventions [88, 97], and aiding in the recovery process [20, 30, 92]. Studies emphasize that satisfying these diverse information needs improves patient understanding, motivates adherence to treatment regimens, and contributes to overall satisfaction with healthcare services [83]. Patients primarily trust doctors with their medical queries [26]. However, traditional in-person consultations pose difficulties, including high costs [11], long wait times, and a strong hierarchy in the patient-physician relationship [32]. Patients often miss opportunities to ask additional (clarification) questions during these brief encounters, as they face information overload [31, 46]. Doctors also struggle with time due to staff shortages straining healthcare ecosystems [4, 11]. Thus, patients and caregivers are increasingly turning to the internet for easier access to medical information [17, 95]. However, the challenge lies in identifying relevant and appropriate information. Individuals struggle to understand complex health information or may even consume inaccurate information, making the process frustrating [5, 22, 101].

Information seekers may also consult experienced patients within their social networks who can share insights related to managing similar health conditions [20, 78]. In the absence of such known individuals, people may resort to social media, which is not only ineffective for quality health support due to widespread misinformation and poor management [59], but also risky in terms of personal privacy [24]. Chandwani and Kulkarni [17] found that although most doctors regard the practice of seeking health information online as an affront to their authority, some see it as inevitable—and even as an opportunity to elevate patients' health literacy through technology.

Recent work in the field of Computer-Human Interaction has explored novel avenues to address these medical information gaps. For instance, Leong et al. [52] propose interactive 'science museum' exhibits in clinical waiting rooms to educate children and their parents about sickle cell anemia, Wilcox et al. [94] and Bickmore et al. [10] propose information displays in hospital rooms to provide patients with real-time status updates and treatment

details, and Pfeifer Vardoulakis et al. [72] propose using mobile phones to present dynamic, interactive reports on patients' progress and care plans throughout their emergency department stay. Although these are specific solutions for certain diseases and infrastructures, discussions have been ongoing around making Electronic Health Record (EHR) data accessible to patients [16, 19, 82]. Patient-facing EHR services have received mixed reviews, with healthcare professionals expressing concerns about adding strain to already understaffed healthcare systems, while patients generally welcome the idea [16].

Prior work has also explored "chat" as an interface to address patients' information needs, by connecting them with doctors in both synchronous [55] and asynchronous [40, 41] manners. Synchronous messaging requires both patient and doctor to be simultaneously active, allowing instant feedback and coherent dialogue. On the other hand, asynchronous messaging platforms like WhatsApp [96], WeChat [90], SMS [71], email [58], or web portals [60, 89] enable communication at the convenience of both parties. Research indicates that both doctors and patients prefer asynchronous texts over instant audio/video calls [40, 79], despite lacking instant feedback [55].

In summary, patients need reliable, relevant, and timely information. However, obtaining this from busy doctors and hospital staff is challenging. This dichotomy results in patients either receiving generic, unverified, and hard-to-understand information from the internet and social circles, or, if lucky, obtaining personalized, verified answers from healthcare experts—at a higher cost. Our solution aims to strike a balance by providing a generic, high-quality LLM answer, later verified by a medical professional with minimal increase to their workload.

2.2 LLM Chatbots in Healthcare

Chatbots enable natural language conversations, offering a minimal learning curve and a personalized experience [86]. AI-powered chatbots surpass rule-based platforms in language understanding, allowing users to build complex queries message by message while retaining context [95]. With the advent of LLMs, building complex chatbots has become significantly easier, enabling researchers to apply LLM-powered bots across diverse healthcare settings [86], including pre-consultation data collection [54], chronic disease management [35, 64], and mental health support [39, 45, 50]. Commercial solutions, such as Babylon Health [28], now provide individuals with instant access to any health-related information. However, LLM-powered healthcare tools have generally been criticized for hallucinations, errors, and a lack of transparency in their reasoning [7, 25]. As the involvement of doctors in these bots is—at most—limited to the data curation step [95], they are susceptible to generating inaccurate information and are "*not ready for clinical use*" [7].

Several methods have been proposed to address these issues, including chain-of-thought prompting [93], retrieval-augmented generation [53], and few-shot learning. Seitz et al. [76] identified that patients' trust in healthcare chatbots is primarily based on a rational evaluation of their capabilities, while trust in human medical professionals relies more on emotional connections and personal rapport. Integrating doctors into the loop with automated LLM chatbots could combine the best of both worlds, a concept we propose in *CataractBot*. In our work, we leverage LLM-based responses in real-time, while human medical experts verify and respond asynchronously. This hybrid approach is a novel and reliable method for addressing healthcare information needs.

We chose cataract surgery—one of the world's most common procedures [61]—as the use case for developing and testing our system. This decision was motivated by our longstanding collaboration with Sankara Eye Hospital, an eye hospital in India. Also, cataract surgery involves well-defined protocols and common patient queries, making it an ideal starting point for testing the bot's ability to deliver accurate, standardized information from a curated knowledge base. However, our solution is versatile, and we believe it can effectively support a range of health conditions, including mental health and chronic care. Given our specific focus on cataract surgery, we examined literature at the intersection of ophthalmology and LLMs. Bernstein et al. [9] reported that responses generated by LLMs to patient eye-care-related questions are comparable to those written by ophthalmologists. Based on similar findings, Ittarat et al. [37] proposed integrating chatbots into ophthalmology practices to provide

24/7 support for common inquiries on eye conditions. Additionally, Yilmaz and Dogan [99] reported that chatbots are an effective tool for educating cataract surgery patients. Our work contributes to this body of literature.

Finally, real-world healthcare deployments of AI are limited due to socio-technical uncertainties in the last mile [70]. To avoid such shortcomings, as recommended by Thieme et al. [84, 85], we designed and studied our system within a tertiary eye hospital in India and examined its integration into clinical workflows.

3 FORMATIVE STUDY

The development of *CataractBot* was informed by our literature survey, our understanding of the Indian healthcare ecosystem, and insights from a formative study conducted at Sankara Eye Hospital. To identify requirements, we conducted semi-structured interviews with 4 ophthalmologists and 2 patient coordinators between July-Aug 2023. We solely interviewed experts, as we aimed to understand the status quo, including their workflows, the types of questions patients ask, and experts' concerns regarding patients' knowledge gaps. We then conducted pilot tests with all end users—patients, attendants, doctors, and coordinators—which iteratively informed the design of *CataractBot*. Interview participants had an average age of 38.2 ± 8.3 years, with 11.3 ± 7.3 years of work experience (Table 4 in Appendix A.2). They participated in the study voluntarily without compensation. The interviews were conducted in English and lasted around 45 minutes each. All sessions were audio-recorded and transcribed by the interviewer. The interviewer open coded the interview transcripts, following inductive thematic analysis [13] to identify key design requirements for the chatbot. Two researchers met regularly to review the emerging codes and finalized the high-level themes.

3.1 Background

Our design requirements were shaped by the workflows and information needs specific to cataract surgery at Sankara Eye Hospital. In the current workflow, patients (with their attendants) first consult a doctor who diagnoses the cataract's maturity and type, recommending the urgency of surgery and potential complications. After deciding on surgery, the patient and attendant meet with a patient coordinator for guidance on pre- and post-operative measures, lens options across different budget ranges, and logistical tasks such as insurance and scheduling. Patients also undergo pre-operative medical tests. For further questions, they can either visit the hospital to meet the doctor or contact the coordinator by phone. On the day of surgery, patients arrive with their attendants in the morning, undergo surgery by mid-morning, and are discharged by the afternoon or evening. The doctor and coordinator provide post-surgery care instructions, covering medication, exercise, bathing, travel, and screen time. One week post-surgery, patients return for a follow-up appointment with the operating doctor.

3.2 Design Requirements

Accuracy. The chatbot's ability to answer both medical and logistical questions with accuracy, precision and contextual relevance was found to be essential given the medical context. While a standard LLM like GPT-4 [69] might offer accurate information, it may not be tailored to India- or hospital-specific nuances. For instance, doctors informed us that although the phacoemulsification method is prevalent globally for cataract surgeries, they often favor manual small incision-based procedures [77] due to their cost-effectiveness. Similarly, cultural considerations significantly influence pre- and post-surgery questions. As a doctor stated, “*People have unique questions... like ‘when can I do nasal yoga post-op?’*” Further, queries related to diet require awareness of Indian food options for precise guidance. Thus, our bot relies on a custom knowledge base curated by hospital staff.

Trustworthiness. Access to trustworthy information is crucial during medical treatment. As one doctor noted, “*The success of treatment is not only in medicine, but the trust they place on us.*” Patients and their attendants typically place their utmost trust in the operating doctor and hospital staff. Even when patients learn something from the

internet or other sources, they often confirm that information with their doctor. To ensure trustworthiness, we decided that (a) each response would be verified and, if necessary, edited by a doctor or patient coordinator, and (b) patients (or their attendants) would be explicitly notified upon expert verification. Note: We intentionally designed *CataractBot* to be non-specific to individual patients by not connecting it with the hospital’s scheduling system or EHR data. This decision was made because LLMs can memorize training data, including personally identifiable information (PII) like emails and phone numbers, and leak it during inference [12], raising privacy concerns and potentially breaking users’ trust.

Timeliness. Discussions with hospital staff highlighted the crucial need for real-time responses. Patient coordinators noted that patients frequently reach out to them over multiple phone calls to ask logistical questions and relay medical queries to their consulting doctors. Due to experts’ busy schedules, immediate responses are not always feasible. This often results in patient seeking information from other sources (including the internet), which can propagate misinformation, or fail to completely address their information needs. To counter this, we decided that the chatbot should provide an instantaneous response to every patient query using the LLM and a custom knowledge base. Subsequently, these responses get reviewed and refined by doctors/coordinators. Moreover, doctors mentioned their unavailability for several hours while performing surgeries. To ensure swift verification, *CataractBot* forwards queries to an ‘escalation’ expert if they remain unverified for 3 hours.

Usability. After cataract surgery, patients can reference their discharge summaries for any queries, but this is challenging due to their temporarily reduced vision. Prior research [33] has identified the crucial role family members play in monitoring the patient’s health. Coupled with doctors’ input (“*Patients ask hardly 10% of the question... attendants are the ones who ask.*”), this underscored the necessity for the bot to be usable by both the patients and their attendants. To cater to diverse backgrounds, we made the chatbot multilingual and multimodal. We chose WhatsApp as the messaging platform due to its widespread use in India [80] and therefore minimal learning curve and quick onboarding without additional installations, despite the richer feature set of other platforms like Telegram or the development flexibility of a custom chatbot. Considering doctors’ demanding schedules, often involving consulting ~50 patients and performing ~10 cataract surgeries a day, we implemented features to minimize their bot-related workload: (1) one-click interaction to verify an answer, (2) allow experts to provide corrections using informal messages [8], (3) handle spelling and grammar errors, and (4) use expert edits to enhance the bot’s knowledge base, minimizing similar edits in future.

4 CATARACTBOT SYSTEM DESIGN

Here, we describe the key components of the *CataractBot* system (Figure 2)—input language and modality, response generation and verification, escalation and reminders—and provide a detailed account of its implementation.

4.1 Components

4.1.1 Input Language and Modality. Sankara Eye Hospital in Bangalore (in the state of Karnataka), referred to as the Silicon Valley of India, caters to patients from various linguistic, educational, and technical backgrounds, including those from the Information Technology sector and neighboring states. Analysis of a 2011 census highlights Bangalore as one of India’s most linguistically diverse cities [66]. To accommodate this varying patient demography, *CataractBot* is designed to support five languages: English, Hindi (the local language of nine states in India), Kannada (the local language of Karnataka), and Tamil and Telugu (the local languages of two neighboring states of Karnataka). As the expert may not be proficient in all five languages, their interactions with the bot are exclusively in English. Upon onboarding, the language preference of the patient and their attendant is collected through an online form, and *CataractBot* initiates the conversation by sending a set of ‘welcome messages’ in the chosen language. Among these messages is an option allowing users to modify their preferred language in future.

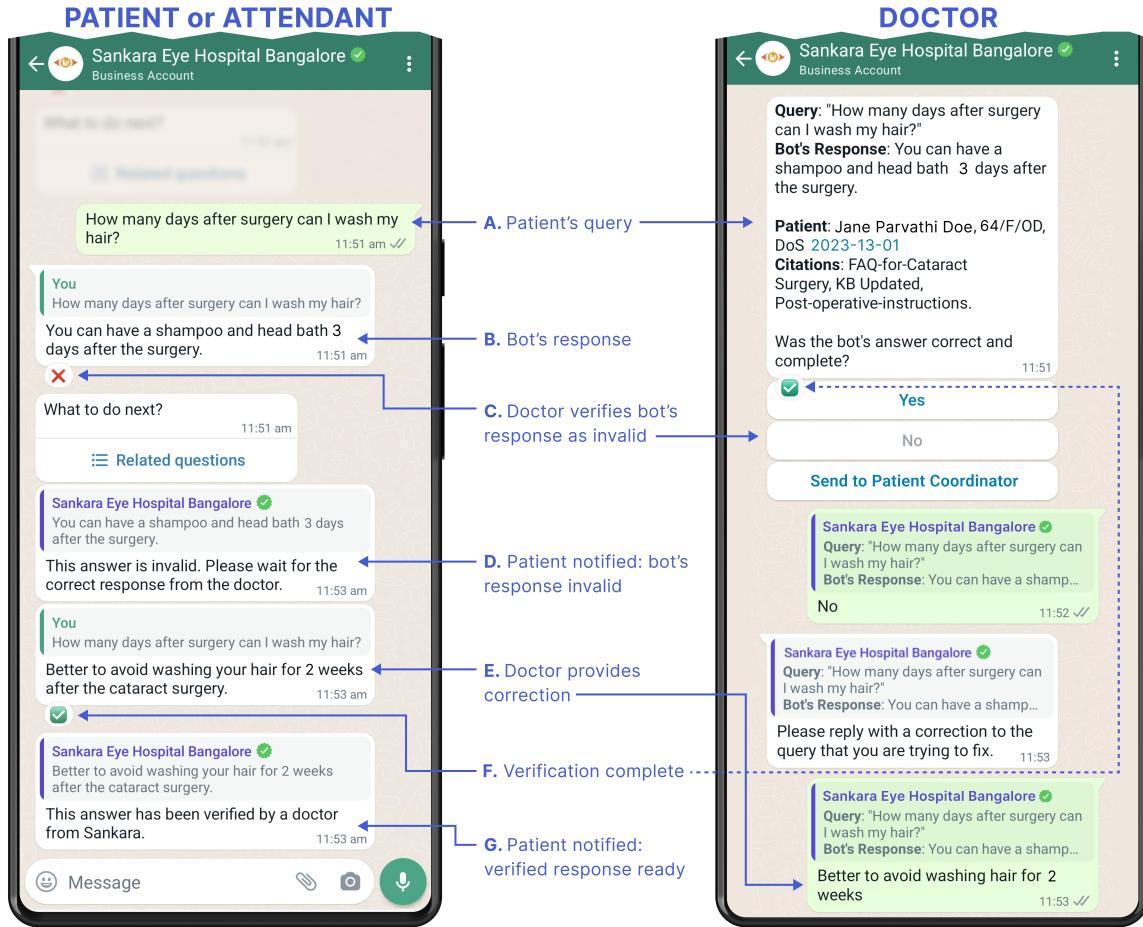


Fig. 1. *CataractBot* provides an initial response to the patient/attendant by querying the knowledge base. The doctor (or coordinator, if the question is logistical) verifies and corrects this response, and the patient/attendant is notified.

To engage with users with different levels of literacy *CataractBot* supports both text and speech inputs. For every voice message, *CataractBot* sends both a text response and a spoken version of the text as an audio message.

Recognizing the challenge users face in initiating a conversation, ('Blank Page Syndrome' [62, 100]), *CataractBot* provides a set of three frequently asked questions along with the welcome message. Additionally, every response includes a "What to do next?" prompt, offering three related questions based on the preceding query (Figure 4F in Appendix A.1.2). Users can select one of the suggested questions or pose a different query to continue the conversation. This feature lowers the barrier to entry (as formulating questions has been found to be non-trivial for low-literate individuals [88]), enhances the natural flow of conversation, and saves time.

4.1.2 Response Generation, Verification and Icons. Upon receiving a message, *CataractBot* classifies it as a medical question (e.g., dos and donts before and after surgery), logistical question (e.g., scheduling or insurance related), small talk (e.g., greetings or expressions of gratitude), or 'other', and provides a response in real-time. For medical and logistical questions, the bot strictly employs the knowledge base curated by the hospital's medical team

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

Stakeholder		Role	
Information seeker	Patient	Person scheduled for cataract surgery. Asks <i>CataractBot</i> surgery-related questions.	
	Attendant	Person accompanying the patient (e.g., child of the patient). Asks <i>CataractBot</i> surgery-related questions.	
Expert	Doctor	Operating doctor	Surgeon scheduled to operate on the patient. Verifies <i>CataractBot</i> 's answers to users' medical questions.
		Escalation doctor	Senior surgeon. Verifies <i>CataractBot</i> 's answers to medical questions that the operating doctor is unable to address in time.
		Knowledge base expert	Senior surgeon. Selects and edits verified answers for addition to <i>CataractBot</i> 's knowledge base.
	Patient coordinator	Operating coordinator	Liaison between patients/attendants and operating doctors. Verifies <i>CataractBot</i> 's answers to users' logistical questions.
		Escalation coordinator	Senior coordinator. Verifies <i>CataractBot</i> 's answers to logistical questions that the operating coordinator is unable to address in time.

to generate an appropriate response. This custom knowledge base comprises of various cataract surgery and hospital-specific documents curated by hospital staff (doctors, patient coordinators, and members of Quality Control and the Patient Safety team). These documents, comprising approximately 30 pages in total, include the Consent Form, Standard Operating Procedures, FAQs, Pre- and Post-Operative Guidelines, and Hospital Information. Some documents, like Post-Operative Guidelines, were included in patients' discharge summaries. Others, such as SOPs and FAQs, existed in the hospital ecosystem but were not directly accessible to patients. The generated response includes a question mark icon  as a 'reaction' indicating the unverified status (Figure 4E in Appendix A.1.2). In instances where the custom knowledge base lacks an answer, the bot responds with a template "*I don't know*" response. For small talk messages (such as "Hello" or "Thank you for the information"), the chatbot provides corresponding small talk responses.

For medical questions, the patient's operating doctor (Table 1) receives a message comprising of the question asked, the bot's response, and patient demographics (Figure 1A). As Rajashekhar et al. [75] found that citations improve trust in LLM-generated responses, this message also includes citations of the documents used to generate the response. The doctor is prompted with the question, "*Is the answer accurate and complete?*" offering three response options: "Yes", "No", and "Send to Patient Coordinator". Tapping 'Yes' replaces the question mark icon  in the information seeker's received answer with a green tick icon , confirming verification (Figure 1F). Additionally, the bot notifies the information seeker that the answer has been verified, tagging that particular response (Figure 1G). Tapping 'No' replaces the question mark  with a red cross icon , indicating an incorrect answer (Figure 1C), and asks the information seeker to await a corrected response from the doctor (Figure 1D). The doctor is asked to provide a correction; they are not required to edit the bot's message, instead, they can offer a correction in free form text. The system combines the bot's initial answer with the doctor-provided correction to frame a new response, which is delivered to the user with a green tick icon  (Figure 1E). This new response is not sent to the doctor, to minimize their workload by limiting verification to a single round. If a question is misclassified, i.e., a logistical question is sent to the doctor, they have the option to 'Send to Patient Coordinator'. Similar to the doctors' workflow, patient coordinators verify and provide corrections to the bot's responses for logistical questions. We deliberated between displaying unverified responses in real-time or only verified responses post-expert verification and correction. Through our formative study and discussions with hospital

staff, it became clear that providing real-time responses was crucial. Delayed responses might lead to patients making repeated calls to the hospital or resorting to online sources, failing to meet their information needs.

4.1.3 Escalation Mechanism. The *CataractBot* system employs an intricate escalation and reminder mechanism to ensure swift verification. In the doctor's WhatsApp interface, unanswered questions are marked with a question mark icon ? (Figure 4E in Appendix A.1.2) and answered questions with a green tick enabling them to easily identify pending queries. If an answer is not verified by the operating doctor within three hours, it is automatically sent to the designated escalation doctor (Table 1). Both the operating and escalation doctors can then verify/correct the response. Selecting 'Yes' or 'Send to Patient Coordinator' immediately marks the query with a green tick (Figure 1F) for both the doctors, indicating task completion. However, if either doctor selects 'No', the green tick appears only after that doctor provides a correction. If neither the operating nor escalation doctor verifies a question within six hours, both receive a reminder notification indicating the pending status. Additionally, every four hours during working hours (at 8 am, 12 pm, and 4 pm), a list of all questions pending verification for more than six hours is sent to both the operating and escalation doctors. This reminder was added based on the doctors' feedback. This workflow is mirrored for the operating and escalation patient coordinators.

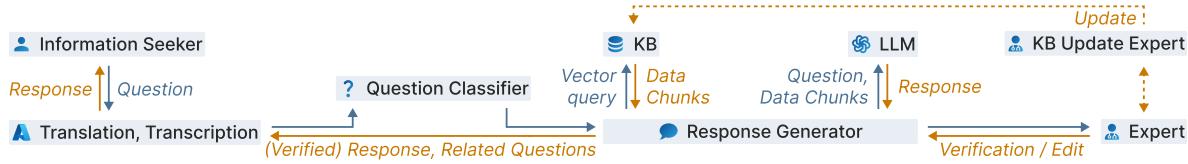
4.1.4 Knowledge Base Update Process. To minimize experts' labor, we use expert-provided edits to update the knowledge base, which increases the likelihood of 'Yes' responses from experts to similar questions in the future. However, certain responses, such as those specific to individual patients (e.g., "When to reach the hospital for my surgery?") must not be added to the knowledge base as they are not generalizable. We appointed a senior cataract surgeon as the 'knowledge base expert' (Table 1), who received a spreadsheet via email at 8pm daily. Their task involves reviewing each row, which consists of question-answer pairs, and determining whether the information should be added to the knowledge base (by responding with a 'Yes'/ 'No' in the 'Should Update Knowledge Base?' column), and if so, modifying as needed the 'Final Answer for Knowledge Base' column, containing the bot's updated answer based on the expert's correction. At 3 am daily, the system extracts 'Question' and 'Final Answer for Knowledge Base' data from all rows marked 'Yes', and append these to the knowledge base in the 'expert-FAQ' document. Additionally, we prompt the LLM to prioritize this 'expert-FAQ' document within the knowledge base. This ensures that *CataractBot* improves with these updates, gradually enhancing its accuracy over time.

4.2 Implementation Details

The *CataractBot* system relies on these five components (Figure 2): (a) LLM: for response generation, (b) Vector Database: for storing and retrieving the custom knowledge base, (c) Language Technologies: for translation and transcription, (d) WhatsApp Services: for message exchange, and (e) Cloud Storage: for storing interaction logs.

We opted for GPT-4, the leading LLM at the time of our system development in May 2023 [69]. The documents for the custom knowledge base are ingested as data chunks into a Chroma vector database, using the OpenAI model 'text-embedding-ada-002' to generate embeddings. Upon receiving a question, GPT-4 classifies the query type. For a medical/logistical question, *CataractBot* employs a retrieval-augmented generation [53] approach. This involves performing a vector search on the knowledge base to extract the three most relevant data chunks. GPT-4 is then prompted (full prompt available in Appendix A.1.1) to generate an answer for the query from these data chunks. If the answer is not present, the bot responds with an "*I don't know*" message.

The prompt underwent several iterations. To validate its effectiveness, a doctor and a patient coordinator recorded 153 questions posed by patients/attendants scheduled for cataract surgery over a week. We used our GPT-4 prompt to obtain answers to these questions from the custom knowledge base, and subsequently the generated responses were evaluated by the same doctor and coordinator. This not only helped refine the LLM prompt, but also led the medical staff to improve existing documents and contribute additional documents to enrich the knowledge base. This process was iterated thrice.

Fig. 2. Overview of *CataractBot* System

Additionally, we utilize GPT-4 to incorporate corrections provided by doctors/coordinators to generate the final expert corrected response. This involved prompting GPT-4 with the patient’s question, the initial response from the bot, and the expert-provided correction, and asking it to generate a revised answer considering the expert’s input. Our experience showed that GPT-4 executed this task well. Regarding language integration, GPT-4 primarily comprehends and responds in English [3]. To bridge the language gap, our system adhered to the standard approach [1] of translating Indic languages (Hindi, Kannada, Tamil, and Telugu) into English for input using Azure AI Translator. It then translated GPT-4’s English responses back into the respective Indic languages for output. To facilitate speech as input and audio as output, *CataractBot* uses Azure speech-to-text and text-to-speech models. (Note: GPT-4 lacks support for speech input.) Finally, all text and audio interactions between the *CataractBot* system and users (including patients, attendants, operating and escalation doctors, coordinators, and knowledge base expert) get logged in a cloud storage for further analysis.

5 STUDY DESIGN

To evaluate the effectiveness of this LLM-powered experts-in-the-loop chatbot in addressing the informational needs of patients undergoing cataract surgery, we conducted a mixed-method user study during Nov 2023-Jan 2024 at our collaborators’ institute, Sankara Eye Hospital. This is one of the leading tertiary eye care and teaching institutions in Bangalore, India. It typically attends to more than 500 patients every day, and conducts over 50 cataract surgeries daily. This study was approved by both the Scientific and Ethics Committees of Sankara Eye Hospital. None of the study participants—patients, attendants, doctors, or coordinators—received any financial incentives for their participation, in accordance with hospital norms.

5.1 Procedure for Patients and Attendants

The operating patient coordinator was tasked to assess the patient’s eligibility for the study based on specific criteria: (a) age of 18 or older, (b) fluent in one of the five languages supported by *CataractBot*, (c) scheduled for cataract surgery within a week, (d) no history of cataract surgery in the last 6 months (as recent patients would likely have minimal informational needs), and (e) having one of the three participating operating doctors as their surgeon. If these criteria were met, the patient and their attendant were directed by the coordinator to meet a researcher (the first author) stationed at the hospital.

The researcher introduced them to *CataractBot* and outlined the study’s protocol which involved using the bot for around two weeks and engaging in an interview pre- and post-surgery. If they agreed to participate, participants were requested to sign a consent form, after which the researcher filled a web-based onboarding form. This form included details such as the patient and attendant’s phone numbers linked to WhatsApp, preferred languages, consulting doctor and coordinator, surgery date, and basic demographic information (age, gender, and education). Upon form submission, participants received ‘welcome messages’ from *CataractBot*. The researcher instructed them to ask a trial question, either by choosing from suggested questions or by typing/speaking in their preferred language. Upon receiving a response, participants were briefed on the icons and expert verification system. Throughout this process, the researcher encouraged them to ask any questions regarding the bot’s usage.

After onboarding, participants received two daily reminder messages (at 7:30 am and 4 pm) prompting them to “ask any cataract surgery related questions” until one week post surgery, when their access was revoked.

We conducted two semi-structured interviews: one on the surgery day and another a week post-surgery. These specific days were chosen because they coincided with the patient’s required hospital visits. The interviews explored their overall experience, specific features they liked or disliked, suggestions for improvement, and questions regarding the trustworthiness, timeliness, and accuracy of responses. Both interviews followed a similar structure, differing only in how the bot supported the participant before versus after surgery. After each interview, participants were requested to fill out a chatbot usability form. Those who agreed, rated their chatbot experience through eight questions, using a five-point Likert scale, covering aspects like ease of use, understandability, and timeliness (Appendix A.3). These questions were adapted from metrics outlined in prior usability evaluations of healthcare bots [2, 36]. Patient-attendant pairs were interviewed together but filled separate usability forms.

All interviews with patients and attendants were conducted by the first author in English, Hindi, Kannada, or Tamil, as the researcher is fluent in these languages. (Note: Participants who selected Telugu as their preferred language during onboarding were interviewed in one of the other four languages.) Interviews on the day of surgery took place in person at the hospital, while post-surgery interviews were conducted either in person at the hospital or through pre-scheduled telephone calls. This decision was influenced by the substantial waiting time on the surgery day—patients typically spend about five hours in the hospital, with the procedure itself taking ~45 minutes. Conversely, during the post-surgery visit, patients receive priority and experience minimal waiting time. Each interview typically spanned 20 to 40 minutes and was audio-recorded with the participant’s consent.

5.2 Procedure for Doctors and Coordinators

We recruited four doctors and two patient coordinators. Three doctors served as operating doctors, and one as both an escalation doctor and knowledge base expert. Both patient coordinators specialize in handling cataract surgery patients, with one serving as a coordinator and the other as an escalation coordinator. The escalation doctor and escalation coordinator held senior positions in the hospital and carried administrative responsibilities. All participating experts attended one of two one-hour training sessions conducted by three researchers. During these sessions, *CataractBot* was demonstrated, and experts were onboarded. The researchers assumed patient roles, asking medical and logistical questions for the doctors and coordinators to verify and edit. The first session revealed a few bugs, and feedback was collected to enhance the bot. For example, a doctor suggested displaying the question first (rather than the patient’s demographics) in verification messages, as only the top portion is visible in ‘tagged’ WhatsApp messages (Figure 1E). After using *CataractBot* for two weeks or longer, the participating experts were interviewed by the first author. These interviews followed a semi-structured format, focusing on the bot’s usability, its integration into their workflow, and its impact on interactions with patients. The interviews were audio-recorded, conducted in English, in-person, and typically lasted 45–60 minutes.

5.3 Participants

A total of 31 patient-attendant pairs (Table 3 in Appendix A.2), consisting of 19 patients and 30 attendants, took part in our study. Due to the predominantly elderly demographic undergoing cataract surgery, 12 patients lacked access to WhatsApp/smartphones and were unable to participate but their attendants used the bot. One patient was not accompanied by an attendant. The average age of the 19 patients was 58.8 ± 8.01 years. Although females comprised 18 (32.7%) of all participants, only 3 (15.8%) were patients. This aligns with prior work showing a gender disparity in accessing cataract surgery in India [73, 98]. Patients had diverse preferred languages (5 English, 5 Hindi, 5 Kannada, 3 Tamil, and 1 Telugu) and education levels ($6 \leq$ Grade 10, 4 Grade 12, 6 Bachelors, and 3 Masters). In contrast, the attendants were younger, predominantly fluent in English, and well-educated. The average age of the 30 attendants (11 female) was 37.2 ± 10.3 years. A majority (23) preferred English, with

2 each favoring Hindi, Kannada, and Tamil, and 1 preferring Telugu. Education levels were mostly high, with 10 having Masters and 13 having Bachelors degrees. Among the 31 patient-attendant pairs, 22 participated in the surgery-day interview, and 10 took part in the one-week post-surgery interview. Various factors influenced participation, including time constraints (resulting in a few declined interviews) and data reaching saturation. Additionally, four doctors and two patient coordinators participated in the study (Table 4 in Appendix A.2). The roles of escalation doctor and knowledge base expert were performed by the same doctor. Note: there was a partial overlap with two doctors and one coordinator participating in both this study and the formative study.

5.4 Data collection and analysis

We performed a mixed-method analysis to systematically analyze the collected data comprising of usage logs, interview transcripts, and usability survey responses. *CataractBot* interaction logs and usability survey data were quantitatively analyzed. Descriptive statistics and statistical tests (such as t-tests and ANOVAs) were used to analyze the count and type of questions asked, verification styles and response times of doctors and coordinators, edits performed by the knowledge base update expert, and responses to the usability survey. Note: For ANOVAs, the sphericity assumption was tested using Mauchly's test, and in case of sphericity violations, the Greenhouse-Geisser correction was applied. The first author translated and transcribed all interviews (totalling 11.5 hours) into English soon after they were conducted. Throughout the study, all authors regularly engaged in discussions to review observations and interviews. For interview transcripts, we conducted an inductive thematic analysis [13] approach, with the first author open-coding all interviews line-by-line. Subsequently, two authors collaboratively reviewed these codes to identify interesting themes in the data (similar to [67]). We iterated on these themes to distill higher-level themes (like "Usage" and "Features"), that we present in our findings below. Note: While there were a total of 49 patients and attendants using *CataractBot*, at times, we treat them as 31 participants (patient-attendant pairs). This choice stems from instances highlighted during interviews, where patients instructed attendants to query the bot using either of their phones. Additionally, both were interviewed together. Only when comparing their distinct roles, we analyze their usage log data separately.

5.5 Positionality

All authors are of Indian origin, and currently reside in Bangalore, India. One author is a practising ophthalmologist, performing over 10 cataract surgeries weekly, and another author is part of the Quality Control and Patient Safety team at the same hospital. Both were involved in designing the study, shaping research questions, providing feedback on initial bot versions, curating the knowledge base, and ensuring that any risk of harm to the patients was mitigated. The ophthalmologist also served as the knowledge base expert. Two authors specialize in HCI and design, three possess healthcare research experience, and four have expertise in software development. We approached this research drawing on our individual experiences and learning from working at the intersection of computing and healthcare in India. Our interest in making healthcare more accessible to the masses has informed the design of *CataractBot* and guided our study design.

6 FINDINGS

Overall, 31 patient-attendant pairs sent 343 messages to *CataractBot* (11.06 ± 8.44 messages/pair). 225 messages were medical questions, 87 logistical, 27 small talk, and 4 others. A researcher manually classified all questions into 12 categories, identified in a bottom-up manner (similar to [22, 55]). Several questions fell into multiple categories. Figure 3A illustrates the relative significance of each category before, on, and after the day of surgery. Before the surgery, questions focused on the procedure (like "anaesthesia", "lens", "safety") and admission/discharge logistics, while post-surgery questions centered on medication and recovery (e.g., "washing hair", "doing yoga"). Doctors directly approved (said 'Yes' to) 69.8% (157) of the bot-generated answers while patient coordinators

directly approved 58.6% (51). 17 of the 22 (77.3%) usability survey respondents found *CataractBot*'s responses to be useful, appropriate, and informative, while 20 (91.0%) indicated willingness to use it in the future.

Below, Section 6.1 presents the reasons behind patients' and attendants' engagement with the bot (answering RQ1). Section 6.2 examines the impact of key features of the bot, including the LLM, experts-in-the-loop and support for multiple stakeholders, languages and modalities (RQ2). Finally, Section 6.3 describes the integration of *CataractBot* into the doctor-patient workflow and discusses issues of personalization and accountability (RQ3). Table 5 in Appendix A.4 presents a summary of key findings across sections.

6.1 Bot's Role in Addressing Information Needs

6.1.1 Reasons for (Not) Using CataractBot. Information seekers used *CataractBot* to address questions they had forgotten or were uncomfortable to ask the doctor or patient coordinator, clarifying answers, and seeking updates. D1 stated that most information seekers come prepared for face-to-face conversations, optimizing their limited time and avoiding multiple visits. *CataractBot* can alleviate the burden of having to remember every question. Ten information seekers highlighted that the bot helped recall information provided during in-person consultations, which they had either forgotten or not completely understood. Patients found comfort in “written” information, available as a reference (similar to [38]), “*I need not remember everything.*” (A20). Six participants also mentioned using the bot to “*double-check*” (A7) and seek “*reassurance*” (A17), despite already knowing certain answers.

Due to existing power differences between doctors and patients [42] and the judgmental attitudes of some doctors, individuals often hesitate to ask questions. Experts acknowledged that educated individuals are comfortable posing questions. In contrast, “*less educated patients who don't speak English*” are the ones for whom *CataractBot* “*can do wonders*” (D3). Five participants, all with a Grade-12 education or below, relied heavily on *CataractBot* asking 57 questions in total, as they reported being unaccustomed to seeking information online. One of them expressed that checking with the bot was more comfortable than approaching a doctor, citing the bot’s non-judgmental nature: “(*The patient*) has cholesterol issues and heart surgery... the doctor would make faces when I ask about food intake... but not the bot.” (A4).

Despite these reasons, 8 out of 49 information seekers used the bot minimally, sending only one or two messages. Two key reasons emerged. First, six participants mentioned that their in-person interactions with the doctor and patient coordinator sufficed to address their queries. This suggests that *CataractBot* could play a more pivotal role in primary healthcare centers, where the human infrastructure may be less capable of addressing information needs or where power imbalances make individuals less likely to seek answers. Second, two participants considered cataract surgery to be a “*simple*” procedure, leading them to have no questions. This indicates that a similar bot could be valuable for more complex (surgical) procedures.

6.1.2 “I don’t know” Responses. Participants complained about receiving “I don’t know” responses (57 total) A total of 36 (10.5%) messages requested status updates—seeking information about the surgery time and discharge time. Due to privacy concerns, *CataractBot* was not integrated with the hospital’s patient management system, resulting in “I don’t know” responses. Patient coordinators later provided answers. While participants acknowledged the appropriateness of “I don’t know” responses, they suggested providing a “*tentative response*” (A10) if possible. For instance, to the question “Mine is general anesthesia or local anesthesia?” A10 expected an answer such as “*99% of cataract surgeries at Sankara Eye Hospital are done under local anesthesia*”. From an expert’s perspective, responding to certain “I don’t know” questions was challenging as they lacked the full context. E.g., A21 asked: ‘I want to postpone the surgery to Feb... is this okay or do I need to get this operated immediately?’, receiving an “I don’t know” response. To correct this, the doctor said: “Please come and check”. This correction frustrated the attendant, who stated: “*But (the patient) already went through the complete eye check; the doctor know her condition... I don't want to come and visit the doctor [again]... it will waste 2-3 hours.*” (A21).

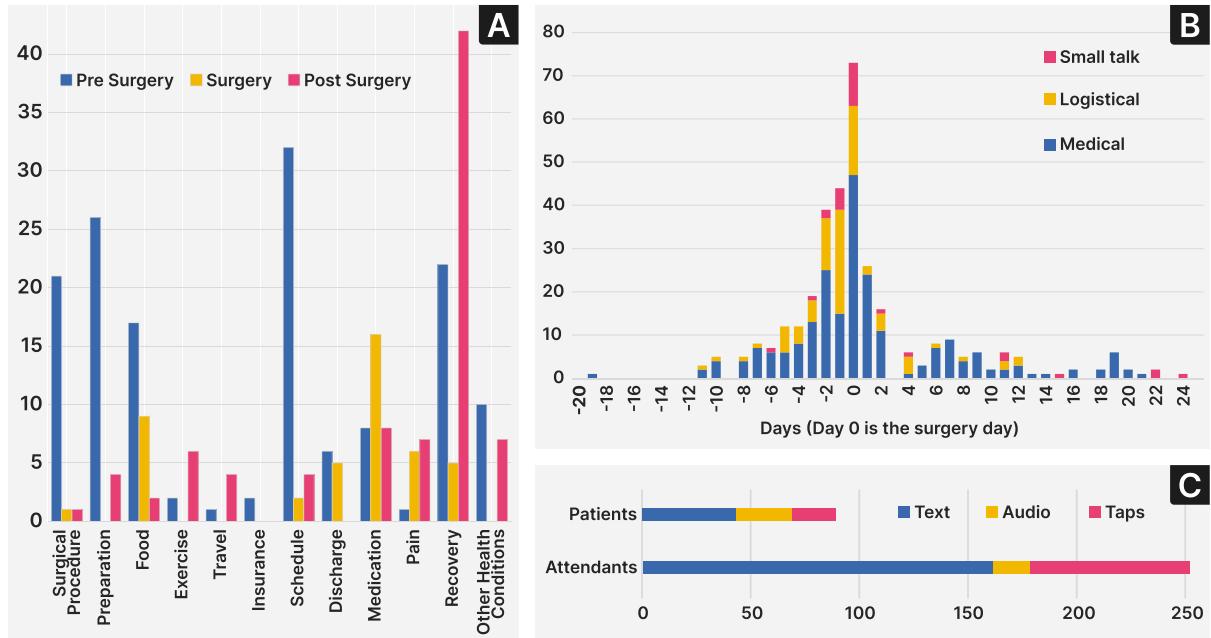


Fig. 3. A. (Left) Messages sent per day pre- and post-surgery. B. (Top right) Messages sent before surgery, on the day of surgery, and after surgery. C. (Bottom right) Message input modality.

When participants received an “I don’t know” response, they sought information from other sources, particularly the internet. For time-sensitive questions, participants reported using the internet to “double-check” the initial LLM-generated, unverified response due to delays in receiving an expert verified response. Four participants, accustomed to relying on the internet for their informational needs, reserved *CataractBot* exclusively for questions requiring expert opinions, such as “When can i do exercises as anulom vilom?” (P27). According to D4, without *CataractBot*, patients might have resorted to the internet and risked following inaccurate advice, given the unavailability of doctors for immediate consultations. *CataractBot* improves the likelihood of receiving accurate responses, as it uses a custom knowledge base.

6.2 Effect of Key Features on the Bot’s Usage

6.2.1 LLM-powered. Patients and attendants found the LLM-generated responses from the curated knowledge base were useful, and experts described them as “*highly accurate*”, noting that they minimized their workload for corrections. These instant responses saved time as information seekers did not have to constantly wait, either physically or over a call. Seven participants emphasized that traveling is time-consuming, and contacting busy hospitals over the phone is challenging due to high patient loads, as “*the phone is always busy*” (A4). Further, the robustness of the GPT-4 model enabled *CataractBot* to appropriately handle complex queries from participants with limited literacy, who struggled to formulate questions, consistent with prior research [38]. For example, P18 asked “Tomorrow operation When should come No phone call received yet?”. *CataractBot* was able to address 45/65 such questions, of which 26 were asked using audio in an Indic language. Typing in Indic languages using the Latin script (*e.g.*, instead of Devanagari script for Hindi) was rare, occurring in only three messages. In two instances, the messages mixed Hindi and English, and the bot responded accurately:

A8: agar opration k baad pain ho raha to kya karna hai?

Bot: If the patient is experiencing pain post-surgery, they should report to the doctor immediately...

For experts, the LLM enabled them to swiftly correct the bot's responses using natural language, without worrying about grammatical or spelling errors. For example, in response to a query, D2 added the correction: "Btr to book an appt". In other instances, experts also instructed the LLM to remove specific sentences, relying on it to generate appropriate responses from their feedback.

6.2.2 Experts-in-the-loop. Information seekers appreciated the novel feature of having human experts verify and update the auto-generated responses (Figure 1F), serving as the key reason for their trust in and engagement with the bot. For instance, "*That's why I gave my WhatsApp (number to onboard the bot), so I can talk directly to the doctor.*" (A2). A12 added: "*The trust only came when I saw the green tick mark... Before that, I was also in the question mark zone... thinking it's a machine... But then somebody is checking it, confirming it... It's actually trustworthy then.*" Despite our concerns that the icons might be complex, participants found them self-explanatory:

"It's simple... Even my dad understood. I didn't tell him [about] the tick or question mark... He told me, "*Doctor verified... the tick mark came*"... If MY dad was able to use it, I think anyone can." – A13.

Although the experts-in-the-loop feature increased trust in the final responses, it had a negative impact on perceived bot intelligence. Four participants noted that when an expert corrected *CataractBot*'s initial answer, their trust in the bot's unverified answers reduced. For instance, "*After some time, when the doctor said it is invalid, I was like, okay, should I even trust the bot?*" (P19). In total, 12.8% (44) of the bot's generated answers (excluding "I don't know") for medical and logistical queries were marked incorrect by the expert. This could be due to the limited custom knowledge base, the experts' high standards for tone and structure of answers, or our prompting strategy favoring caution. Also, since participants knew that these messages were being reviewed by experts, four used the bot to send direct messages to them instead of asking questions (e.g., "Dr. Give me a call?" (P19)).

Experts appreciated the intermediary role of the bot, introducing a layer of separation from patients: "*From a doctor's standpoint, it is very good because it puts in a curtain between us.*" (D4). Further analysis revealed that their suggested edits fell into five categories: adding missing information (76.8%), asking clarification questions (14.6%), making factual corrections (11.0%), asking the patient to visit the hospital (8.8%), and removing unnecessary information (3.7%). The length of expert corrections in characters varied significantly ($t(91)=5.2, p<0.05$) with 152.4 ± 104.9 for medical questions and 50.9 ± 42.6 for logistical questions.

Finally, experts did not have access to the final answer sent to information seekers after their suggested edits. This limited verification to a single round, and was designed to minimize their workload. However, this lack of transparency sometimes left experts uncertain if *CataractBot* had understood their corrections. Only the knowledge base update expert had access to the updated answer sent to the information seeker.

6.2.3 Multimodality. The 19 patients sent 89 messages (43 text, 26 audio, and 20 taps), while the 30 attendants sent 254 messages (163 text, 17 audio, and 74 taps) (Figure 3C). One-way ANOVAs on message types found statistical significance for attendants ($F(2, 87)=13.74, p<.001$), but not for patients ($F(2, 54)=0.77, p=0.46$). Post-hoc pairwise comparisons, corrected with Holm's sequential Bonferroni procedure, indicated statistically significant differences between 'text' and 'audio' ($t(29)=4.8, p<.01$) and 'text' and 'taps' ($t(29)=3.0, p<0.01$) for attendants. The minimal usage of audio by attendants could be due to demographic differences, with the majority being under 45 years and well-educated. Audio messages were predominantly used by older semi-literate patients. For instance, "*I can't read (or) write Hindi that well, so asking and listening to audio, I did that.*" (P12).

Post-surgery limitations on screen use also indicated that speech input and audio output the preferred modality for patients, with 18 out of 26 audio questions by patients asked post-surgery. 31 participants used the related questions at least once. As P7 stated: "*I used the audio message once. After that there was no need for me [to type or use audio]... I just kept selecting from the [related questions] options.*" While they found the related questions

feature convenient and frequently used it, there were instances of dissatisfaction when (in eight cases) the answers to those resulted in “I don’t know” responses. This issue arose because the *CataractBot* system uses an LLM call to suggest three questions based on the users’ last question, independent of the knowledge base. This approach aimed to uncover gaps in the knowledge base by not restricting questions to the curated knowledge base.

On the expert end, they could only interact with *CataractBot* using text. This design choice facilitated a “*hygiene check*” (D4) for thoroughness and precision in experts’ edits, aligning with previous findings [55].

Patients mentioned utilizing the persistent information provided by *CataractBot* as a reference and sharing it with others undergoing cataract surgery. Also, as per D1, this information can help when discussing their surgery with experts outside of the Sankara Eye Hospital ecosystem. However, it also increased the sense of accountability among experts, making them “*somewhat nervous*”. They felt the need to be “*very lawyer-like in our conversation because we don’t want it to come back and bite us tomorrow.*” (D4).

6.2.4 Multilinguality. The majority of information seekers (29 out of 49) selected English, while 7 opted for Hindi, 6 Kannada, 5 Tamil, and 2 Telugu. None of the participants changed their initially selected language. Given the challenges of typing in Indic languages on smartphones [43], our non-English participants predominantly relied on speech input and selecting from related questions. In total, patients who chose English typed 35 questions, asked 0 questions in speech, and tapped 6 related questions. In contrast, patients opting for Indic languages typed 8 questions, sent 26 audio questions and tapped 14 related questions in total. A similar trend was observed among attendants. Patients found value in local language interaction as input, helping them use the bot independently: “*Without even asking me for help, he (patient) was able to use that bot as it supports Tamil.*” (A13).

The audio queries in Indic languages were transcribed and translated to English using Azure AI services. However, this was not highly accurate due to the variability in Indic language dialect. This is a known issue and prior work [38] relied on human wizards to correct such errors. Experts faced challenges in “*deciphering*” such messages since they only received the English transcription. In the future, there should be a way to share the original audio query with the expert on demand.

6.2.5 Support for Multiple Stakeholders. Patients sent a total of 89 messages (4.7 ± 5.4 message/patient), while attendants sent 254 messages (8.5 ± 6.8 message/attendant), significantly more than patients, with $t(47)=2.04$, $p<0.05$ (Figure 3C). During our interviews, we identified a few reasons for the patients to minimally use the bot, particularly related to demographics. Patients were typically older with lower levels of literacy and tech-literacy compared to their attendants. These patients trusted the hospital to provide them with all the required information and did not actively seek information themselves. For instance:

“Dad didn’t try... I was the one thinking about it [the surgery]. He was like, “*It’s OK! Not everyone needs to bother in the family... (to know) about what will be this, what will be that.*” – A3.

Additionally, patients were advised to minimize screen use post-surgery, contributing to reduced bot interaction. Interestingly, patients sometimes offloaded the knowledge-seeking ‘work’ to their active attendants—e.g., “*I have been using it. I told him [the patient] what answers I got. So maybe he did not feel the need (to use it).*” (A20).

Moving on to experts, both the doctors and patient coordinators highlighted the significant role of escalation experts. Among the 309 verified medical and logistical answers, 62.5% were verified by the operating doctor/coordinator and the rest 37.5% by the escalation expert. Doctors mentioned having 2-3 days per week reserved for surgeries, and on these “*operating days*”, they rarely checked their phones during work hours, leading to more escalations. Despite their crucial role, escalation experts interacted differently from operating doctors and coordinators. Unlike the latter, who had face-to-face meetings with patients, escalation experts lacked contextual knowledge, limiting them to providing generic responses.

6.2.6 Knowledge Base. Out of the 91 responses edited by experts, 72 (79.1%) received approval from the knowledge base expert (D4) to be added to the knowledge base. For 49.3% of these, the bot’s final response (generated by

merging the initial response and expert edit suggestion) underwent additional editing by the knowledge base expert. These changes were minimal with an average relative Hamming distance of 28.2%, and mainly addressed aspects of “*tone and structure*”, such as preferring active voice and making responses more generally applicable across patients. This hints that the bot effectively incorporated experts’ edits. Between the first and sixth weeks of the study, the proportion of LLM-generated answers that experts marked as “*accurate and complete*” increased by 9.9% for medical questions and 22.9% for logistical questions. This suggests that as the knowledge base was updated, the experts’ verification workload decreased.

6.3 Bot Integration into Doctor-Patient Workflows

6.3.1 Workflows. The majority of messages (71.4%) were sent in the 8-day period from 5 days prior to surgery to 2 days post-surgery with the highest number of messages (76) sent on the surgery day (Figure 3B). Messages were sent throughout the day (7am to 11pm), as A18 mentioned: “*Whenever I am idle at work or at home... whatever (question) comes to my mind, I just ask.*” The workflow employed by information seekers varied. Most participants would ask a question, leave their phones, and check the answer later when free. They hardly noticed the wait due to being occupied with their day-to-day activities. In such cases, notifications played a key role. The message stating that the expert had provided a verified response (Figure 1G), made the otherwise unnoticeable green tick (Figure 1F) prominent. Participants, when in need of a quick response, either relied on the initial (LLM-generated) answer or re-verified it on the internet.

19/22 survey respondents (86.4%) agreed that the bot responded quickly. On average, the bot took 11.8 ± 10.2 seconds to respond. Expert verification took 162.9 ± 172.3 minutes. We found verification without edits (151.1 ± 147.1 minutes) to be significantly faster than verification with edits (191.3 ± 172.4 minutes) with $t(265)=2.7$, $p<0.05$. Participants were “*okay to wait*” for the expert’s verification, since the response was being verified by “*busy doctors*”. This demonstrated their clear understanding of the bot’s workflow.

“I know that it will take its own time... based on their [doctors’] busy schedule, their availability... It’s not like there’s some dedicated doctor monitoring it all the time, right? If that is the case, then (the) bot is not required. They [doctor] can directly reply.” – A17.

Notifications played a key role. The message stating that the expert had provided a verified response (Figure 1G), made the otherwise unnoticeable green tick (Figure 1F) prominent and the bot’s process self-explanatory.

For experts, the asynchronous nature of the bot provided them the flexibility to offer verifications at their convenience. The majority of responses (31) were provided by experts at 4-5 pm, post their busy workday, as most OPD and surgeries finish by 4 pm. The second-highest responses (28) occurred at 1-2 pm, during their lunch break. This usage pattern was described as: “*I use my mobile phone only when I go for breaks... lunch break, coffee break, or in the morning before I start OPD. The rest of the time, my (cellular) data is off.*” (D1). Experts refrained from using their phones in front of patients due to the negative perception associated with using phones for personal reasons. We found that the experts were “*fine*” with the additional task of verifying the bot’s responses, as articulated by D4: “*It takes only about 15-20 seconds to answer a question... It was not really eating into my time... It’s much easier than replying to my mailbox.*”

As five experts used *CataractBot* on their personal WhatsApp accounts, it blurred the boundary between their home and work, similar to previous findings [63, 90]. Our experts also noted that WhatsApp had a low signal-to-noise ratio, resulting in instances where they missed *CataractBot* messages: “*My WhatsApp has 5000+ [unread] messages... college groups, family groups... I just see messages in the important groups.*” (D2). Addressing this issue, the escalation doctor D4 suggested that experts should be recommended to “*pin*” *CataractBot* on WhatsApp, ensuring it remains consistently at the top of their chat list.

6.3.2 Personalization vs Privacy. Three participants valued that the bot could not access their medical records.

“Using personal health information [by the bot] is... a violation of patients’ rights. If their data is added... external agencies can extract that information and might use it for advertising.” – A17.

These participants were educated (holding Master’s degrees) and considered personalization as “*an unnecessary add-on*”. They raised concerns about complications associated with sharing PII with LLMs, both in terms of legislation and “*privacy threats*”. A1, for example, remarked “*I would definitely not want my mom’s data to be out there, all over the internet.*”. Moreover, lack of personalization enabled usage of the chatbot for others undergoing cataract surgery: “*Two people I know are going to have their cataract surgery soon... I’ll use this bot for them.*” (A11).

On the other hand, fueled by the recent hype around AI and LLMs, eight participants expected “*personalized*” answers to their queries with little concern for privacy, and were slightly disappointed with generic responses. This aligns with prior findings indicating minimal privacy concerns related to digital data in the global south, specifically in India [67]. Interestingly, three participants did not even consider their medical data to be “*private*”, as they stated, “*It’s nothing personal to be hidden. All (of us) have common health issues.*” (A14). Moreover, two participants questioned why others would be interested in their medical data, as “*it is not useful for others*” (P18).

Two participants wanted personalization, along with the sharing of patient data, to be an “*opt-in*” feature where individuals conduct a cost-benefit analysis and make an informed choice based on their preferences. They believed that a personalized bot would offer more relevant, and fewer “I don’t know”, responses.

“If you don’t want to share your information, it (the bot) should say “*If you ask personal questions, I won’t have access, so you may have to visit the hospital.*”... We have to compromise.” – A21.

Five experts voted to integrate the chatbot with patient data, including schedules and consultation history, to offer personalized responses. D3 suggested integrating patient medical history to facilitate customized responses from the bot and doctors: “*We don’t even know whether they’re diabetic, they’re hypertensive. So, when they ask “What should I eat and come?”... I am not sure what to say.*” (D3). However, D1 and D2 also voiced skepticism towards the LLMs capability to understand complex patient records accurately. “*There are a lot of variables... the surgeon, the patient, their tests... It’s too much information to process.*” (D1). In addition, the knowledge base expert cautioned that highly personalized responses would result in minimal updates to the knowledge base, potentially increasing *CataractBot*-related workload for experts and negatively impacting chatbot adoption among experts.

7 DISCUSSION

In this paper, we present *CataractBot*, an LLM-powered experts-in-the-loop WhatsApp chatbot to address the information needs of patients undergoing cataract surgery. Building on limited prior research in multi-stakeholder settings [39, 97], we provide valuable insights into LLM-mediated interactions among patients, attendants, doctors, and coordinators. Our field study revealed positive evidence of *CataractBot*’s usefulness and usability across stakeholders. Doctor-verified responses were key to patients’ and attendants’ trust in and engagement with the bot. They appreciated that *CataractBot* instantly answered questions anytime, saving time by reducing the need for hospital calls or visits. Those with limited (tech-)literacy found it easy to use because of the multilingual and multimodal support. Doctors and coordinators commended *CataractBot* for acting as a facilitator, creating a layer of privacy between them and patients, and providing them the flexibility to verify responses at their convenience.

Below, we identify design implications, raising critical insights and open questions that warrant further attention for such experts-in-the-loop chatbots to succeed on a larger scale across domains.

7.1 Design Implications

7.1.1 Generalization and Scaling. We designed and evaluated our experts-in-the-loop chatbot system in the context of cataract surgery in India. Expanding it beyond India may require further research. In the United States, for example, insurance plays a crucial role in healthcare. Patients often require prior authorization for treatments, and they would value a bot providing information in navigating these processes [81]. This necessitates involving

insurance-specific experts in knowledge base creation and response verification. Also, our study raised concerns about expert accountability due to the persistence of written communication in the bot. This issue could be more pronounced in the Global North, where doctors are increasingly wary of lawsuit risks [81]. Balancing these needs and tensions will be critical for *CataractBot*'s adoption across geographies. Finally, health communication is most effective when users perceive the source as demographically and attitudinally similar to themselves, not just contextually relevant [49]. This highlights the need to adapt *CataractBot* for different settings by (1) parametrizing context-specific details, such as payment methods and hospital protocols, and (2) tailoring language to ensure cultural relevance through familiar terms and metaphors.

Our open-sourced framework can be adapted beyond cataract surgery, to support more complex, atypical, or long-term treatments—e.g., glaucoma surgery, cancer treatment, or pregnancy—where easy access to verified information would be valuable for patients. The applicability can be extended further, to fields beyond the medical domain, such as law, finance and education. In these scenarios, end-users (e.g., defendants, taxpayers, or students) could receive synchronous responses from an LLM using a custom knowledge base, and subsequent asynchronous verified responses from experts (e.g., lawyers, accountants, or teachers). Expert-mediated bots can even be used to upskill less-trained workforce members. For instance, these bots could serve as valuable learning resources for community health workers, who typically have minimal medical training [96].

7.1.2 Expert-in-the-loop Approaches. Prior studies have raised concerns about the application of LLMs in healthcare scenarios, due to their inconsistency, potential errors, and bias [7, 25]. Moreover, users who are less (technologically) literate face a heightened risk of harm [44]. Our experts-in-the-loop framework not only addresses these concerns but also adheres to OpenAI's usage policies [68], which state that “*tailored medical/health advice cannot be provided without review by a qualified professional*”. By using our framework, chatbots can operate at the “*sweet spot of patient-LLM-clinician collaboration*” [35], where the capabilities of technology and human judgment are integrated. In our study, this ensured reliability and cultivated trust, as information seekers knew that their operating doctor was verifying responses. However, trust in the initial automated answers reduced when these answers were subsequently marked as incorrect/incomplete by the doctor. To address this in future work, one design approach could be to eliminate the initial automated response, providing only expert-verified answers. We caution that this may lead to long wait times, potentially discouraging use. For high-risk use cases like cancer treatment, verified accuracy may be prioritized over immediacy. Future research could compare the two paradigms: a hybrid system with synchronous automated answers followed by asynchronous verification versus a fully asynchronous system delivering only verified answers. Such a study would provide valuable insights into balancing immediate access with the accuracy required for critical medical information in chatbot design.

Doctors are a scarce resource in healthcare ecosystems. Our chatbot system can help scale their impact, by reducing the time spent answering routine questions at length during consultations and calls, thus freeing up time for additional patient care. However, as doctors had to verify responses during deployment, *CataractBot* also added to their workload. To minimize this, one approach is to deliver pre-verified answers to specific questions that the expert has repeatedly marked as correct. Given the potential for LLM hallucinations [25], cached responses must be carefully matched to similar patient demographics and question phrasing before reuse. Another approach is to use a second LLM for quality control while generating each response, as GPT-4’s evaluations of AI-generated responses align well with human expert assessments [34]. This maker-checker verification module would enable tracking of the bot’s response quality, especially as the knowledge base expands. The module could also evaluate the confidence level in the generated response, or the perceived urgency of the question, which could be used to prioritize sending nudges to experts for verification.

7.1.3 Multimodality and Personalization. Our system faced challenges in accurately interpreting questions from voice messages due to limitations in transcription and translation technologies. The same inaccurate transcriptions were shared with doctors leading to confusion. Such issues would not arise in deploying a voice-based input

modality to a predominantly English-speaking population or a tier-1 language setting in the Global North, as seen in prior work [97]. Additionally, non-verbal cues in audio messages—such as signs of anxiety, impatience, or the presence of others—were lost in textual transcriptions. This hindered experts' ability to deliver appropriate care (also noted in [55]). To address this, we suggest making the original audio clips available to experts on-demand. Beyond audio, considering the visual capabilities of recent LLM models like GPT-4V, such bots could be enhanced to process image-based inputs, aligning with recommendations in prior studies [38]. For example, a patient could upload a photo of a specific medicine to ask questions about it, instead of providing a written description. While this could improve the user experience for information seekers, it may increase the workload for experts. Further, prior research [35] cautions that although LLMs are proficient in text processing, their ability to handle other forms of media, such as image and video, remains less reliable.

Moving on to personalization, while tailored responses can boost adoption and utility by providing more relevant information, they also raise privacy concerns, potentially reducing usage among certain users. This issue may be more pronounced in the developing world, where there is more awareness around privacy. Additionally, experts may face increased workloads as they reference user-specific information for verification, since many personalized edits may not be suitable for inclusion in the generic knowledge base. Low (tech-)literate users might require assistance in making informed decisions about the use of their data [67]. A binary approach of either sharing all records or none may not address individual preferences. Instead, users should have the agency to selectively share records they deem relevant and acceptable for the bot's access. This aligns with the philosophy of India's National Digital Health Mission [65]. One solution could be to allow users to upload relevant documents or manually enter specific details they want the bot to access. This can empower users to control shared information while benefiting from a personalized experience. Further, when developing *CataractBot* in May 2023, we chose to use GPT-4, as it was the leading LLM [69]. However, recent open-source models like LLaMA have shown comparable performance to GPT in medical applications [56]. Incorporating these models into the *CataractBot* system would enable local processing of patient data, helping address privacy concerns.

7.2 Limitations

We acknowledge several limitations of this work. First, results should be interpreted with caution, given the specificities of the study, including a relatively small sample size, users' first encounter with a medical chatbot, and the cataract treatment scenario. Positive user responses, minimal increase in expert workload, and other findings should be validated in future studies, preferably with longitudinal designs. Second, system-level analyses, such as scalability testing and power measurement, are needed before deploying the proposed solution more broadly. Finally, although participants reported high trust, repeated "I don't know" responses or inaccurate information during regular use may erode this trust. We note the potential risk of disseminating inaccurate medical information at scale with chatbot and LLM technologies. Therefore, stronger methods are needed to develop, review, and update the knowledge base, as well as to routinely evaluate the accuracy of bot's responses.

8 CONCLUSION

Building a system that addresses patients' information needs by providing expert-verified information is an open problem in healthcare. We propose a novel solution—an LLM-powered experts-in-the-loop chatbot framework, that utilizes retrieval-augmented generation over a custom knowledge base to provide synchronous responses, and utilizes experts to provide verified responses asynchronously. Our in-the-wild study involving 49 information seekers (patients and attendants) and 6 experts (4 doctors and 2 coordinators) demonstrated the positive impact of this technological intervention. Patients not only trusted the chatbot's response, but were also willing to wait, understanding that their busy doctors were verifying them. Simultaneously, doctors and coordinators, despite their hectic schedules, made time during breaks and at home to help patients by verifying LLM-generated

responses. The favorable feedback from various stakeholders in the healthcare ecosystem indicates that a chatbot, delivered through ubiquitous smartphones and WhatsApp, can effectively enhance information access in critical healthcare settings, even for individuals with limited literacy and technology experience.

ACKNOWLEDGMENTS

Many thanks to all the participants for their time and patience.

REFERENCES

- [1] [n. d.]. An AI service for real-time document and text translation. <https://www.microsoft.com/en-us/ai/ai-translator> Accessed: 2024-02-01.
- [2] Alaa Abd-Alrazaq, Zeineb Safi, Mohannad Alajlani, Jim Warren, Mowafa Househ, and Kerstin Denecke. 2020. Technical metrics used to evaluate health care chatbots: scoping review. *Journal of medical Internet research* 22, 6 (2020), e18301.
- [3] Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. arXiv:2303.12528 [cs.CL]
- [4] Shashank M Akerkar and Lata S Bichile. 2004. Doctor patient relationship: changing dynamics in the information age. *Journal of postgraduate medicine* 50, 2 (2004), 120–122.
- [5] Neeraj K Arora, Bradford W Hesse, Barbara K Rimer, Kasisomayajula Viswanath, Marla L Clayman, and Robert T Croyle. 2008. Frustrated and confused: the American public rates its cancer-related information-seeking experiences. *Journal of general internal medicine* 23 (2008), 223–228.
- [6] Simon J. Atfield, Anne Adams, and Ann Blandford. 2006. Patient information needs: pre- and post-consultation. *Health Informatics Journal* 12, 2 (2006), 165–177. <https://doi.org/10.1177/1460458206063811> arXiv:<https://doi.org/10.1177/1460458206063811> PMID: 17023406.
- [7] Joshua Au Yeung, Zeljko Kraljevic, Akish Luinetel, Alfred Balston, Esther Idowu, Richard J Dobson, and James T Teo. 2023. AI chatbots not yet ready for clinical use. *Frontiers in Digital Health* 5 (2023), 60.
- [8] Robert E Beasley. 2009. Short message service (SMS) texting symbols: A functional analysis of 10,000 cellular phone text messages. *The Reading Matrix* 9, 2 (2009).
- [9] Isaac A Bernstein, Youchen Victor Zhang, Devendra Govil, Iyad Majid, Robert T Chang, Yang Sun, Ann Shue, Jonathan C Chou, Emily Schehlein, Karen L Christopher, et al. 2023. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Network Open* 6, 8 (2023), e2330320–e2330320.
- [10] Timothy W Bickmore, Laura M Pfeifer, and Brian W Jack. 2009. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1265–1274.
- [11] Thomas S Bodenheimer and Mark D Smith. 2013. Primary care: proposed solutions to the physician shortage without training more physicians. *Health Affairs* 32, 11 (2013), 1881–1886.
- [12] Jaydeep Borkar. 2023. What can we learn from Data Leakage and Unlearning for Law? *arXiv preprint arXiv:2307.10476* (2023).
- [13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. <https://doi.org/10.1191/1478088706qp063oa> arXiv:<https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa>
- [14] G Byrne and R Heyman. 1997. Patient anxiety in the accident and emergency department. *J. Clin. Nurs.* 6, 4 (July 1997), 289–295.
- [15] Åsa Cajander and Christiane Grünloh. 2019. Electronic Health Records Are More Than a Work Tool: Conflicting Needs of Direct and Indirect Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300865>
- [16] Åsa Cajander and Christiane Grünloh. 2019. Electronic Health Records Are More Than a Work Tool: Conflicting Needs of Direct and Indirect Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300865>
- [17] Rajesh Chandwani and Vaibhavi Kulkarni. 2016. Who's the Doctor? Physicians' Perception of Internet Informed Patients in India. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (*CHI '16*). Association for Computing Machinery, New York, NY, USA, 3091–3102. <https://doi.org/10.1145/2858036.2858500>
- [18] Eva Christalle, Jördis Zill, Wiebke Frerichs, Martin Härtter, Yvonne Nestorius, Jörg Dirmaier, and Isabelle Scholl. 2019. Assessment of patient information needs: A systematic review of measures. *PLOS ONE* 14 (01 2019), e0209165. <https://doi.org/10.1371/journal.pone.0209165>
- [19] Chia-Fang Chung. 2017. Supporting patient-provider communication and engagement with personal informatics data. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers* (Maui, Hawaii) (*UbiComp '17*). Association for Computing Machinery, New York, NY, USA, 335–338. <https://doi.org/10.1145/3123024.3123197>

- [20] Andrea Civan-Hartzler, David W. McDonald, Chris Powell, Meredith M. Skeels, Marlee Mukai, and Wanda Pratt. 2010. Bringing the field into focus: user-centered design of a patient expertise locator. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) (*CHI '10*). Association for Computing Machinery, New York, NY, USA, 1675–1684. <https://doi.org/10.1145/1753326.1753577>
- [21] Martina A Clarke, Joi L Moore, Linsey M Steege, Richelle J Koopman, Jeffery L Belden, Shannon M Canfield, Susan E Meadows, Susan G Elliott, and Min Soon Kim. 2016. Health information needs, sources, and barriers of primary care patients to achieve patient-centered care: A literature review. *Health Informatics Journal* 22, 4 (2016), 992–1016. <https://doi.org/10.1177/1460458215602939> PMID: 26377952 arXiv:<https://doi.org/10.1177/1460458215602939> PMID: 26377952
- [22] Martina A Clarke, Joi L Moore, Linsey M Steege, Richelle J Koopman, Jeffery L Belden, Shannon M Canfield, Susan E Meadows, Susan G Elliott, and Min Soon Kim. 2016. Health information needs, sources, and barriers of primary care patients to achieve patient-centered care: A literature review. *Health Informatics Journal* 22, 4 (2016), 992–1016. <https://doi.org/10.1177/1460458215602939> arXiv:<https://doi.org/10.1177/1460458215602939> PMID: 26377952 arXiv:<https://doi.org/10.1177/1460458215602939> PMID: 26377952
- [23] Inger Dahlén and Christer Janson. 2002. Anxiety and depression are related to the outcome of emergency treatment in patients with obstructive pulmonary disease. *Chest* 122, 5 (Nov. 2002), 1633–1637.
- [24] Munmun De Choudhury, Meredith Ringel Morris, and Ryen W. White. 2014. Seeking and sharing health information online: comparing search engines and social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) (*CHI '14*). Association for Computing Machinery, New York, NY, USA, 1365–1376. <https://doi.org/10.1145/2556288.2557214>
- [25] Kerstin Denecke, Richard May, and Octavio Rivera Romero. 2024. Potential of Large Language Models in Health Care: Delphi Study. *J Med Internet Res* 26 (13 May 2024), e52399. <https://doi.org/10.2196/52399>
- [26] Xianghua Ding, Xinning Gui, Xiaojuan Ma, Zhaofei Ding, and Yunan Chen. 2020. Getting the Healthcare We Want: The Use of Online "Ask the Doctor" Platforms in Practice. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI '20*). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376699>
- [27] F Duffy, Geoffrey Gordon, Gerald Whelan, Kathy Cole-Kelly, Richard Frankel, Natalie Buffone, Stephanie Lofton, MaryAnne Wallace, Leslie Goode, and Lynn Langdon. 2004. Assessing Competence in Communication and Interpersonal Skills: The Kalamazoo II Report. *Academic medicine : journal of the Association of American Medical Colleges* 79 (07 2004), 495–507. <https://doi.org/10.1097/00001888-200406000-00002>
- [28] eMed. 2024. *Babylon Health*. Retrieved Sept 1, 2024 from <https://www.emed.com/uk>
- [29] Magda Eriksson-Liebon, Susanne Roos, and Ingrid Hellström. 2021. Patients' expectations and experiences of being involved in their own care in the emergency department: A qualitative interview study. *Journal of clinical nursing* 30, 13-14 (2021), 1942–1952.
- [30] Ahmed Fadhil. 2018. A conversational interface to improve medication adherence: towards AI support in patient's treatment. *arXiv preprint arXiv:1803.09844* (2018).
- [31] Fabrice Ferré, Nicolas Boeschlin, Bruno Bastiani, Adeline Castel, Anne Ferrier, Laetitia Bosch, Fabrice Muscari, Matt Kurrek, Olivier Fourcade, Antoine Piau, and Vincent Minville. 2020. Improving Provision of Preanesthetic Information Through Use of the Digital Conversational Agent "MyAnesth": Prospective Observational Trial. *J Med Internet Res* 22, 12 (4 Dec 2020), e20455. <https://doi.org/10.2196/20455>
- [32] Grethe Fochsen, Kirti Deshpande, and Anna Thorson. 2006. Power Imbalance and Consumerism in the Doctor-Patient Relationship: Health Care Providers' Experiences of Patient Encounters in a Rural District in India. *Qualitative Health Research* 16, 9 (2006), 1236–1251. <https://doi.org/10.1177/1049732306293776> arXiv:<https://doi.org/10.1177/1049732306293776> PMID: 17038755
- [33] Erik Grönvall and Nervo Verdezoto. 2013. Beyond self-monitoring: understanding non-functional aspects of home-based healthcare technology. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Zurich, Switzerland) (*UbiComp '13*). Association for Computing Machinery, New York, NY, USA, 587–596. <https://doi.org/10.1145/2493432.2493495>
- [34] Varun Gumma, Anandhita Raghunath, Mohit Jain, and Sunayana Sitaram. 2024. HEALTH-PARIKSHA: Assessing RAG Models for Health Chatbots in Real-World Multilingual Settings. *arXiv:2410.13671 [cs.CL]* <https://arxiv.org/abs/2410.13671>
- [35] Yuexing Hao, Jason Holmes, Mark Waddle, Nathan Yu, Kirstin Vickers, Heather Preston, Drew Margolin, Corinna E Löckenhoff, Aditya Vashistha, Marzyeh Ghassemi, et al. 2024. Outlining the Borders for LLM Applications in Patient Education: Developing an Expert-in-the-Loop LLM-Powered Chatbot for Prostate Cancer Patient Education. *arXiv preprint arXiv:2409.19100* (2024).
- [36] Samuel Holmes, Raymond Bond, Anne Moorhead, Jane Zheng, Vivien Coates, and Michael McTear. 2023. Towards Validating a Chatbot Usability Scale. In *Design, User Experience, and Usability*, Aaron Marcus, Elizabeth Rosenzweig, and Marcelo M. Soares (Eds.). Springer Nature Switzerland, Cham, 321–339.
- [37] Mantapond Ittarat, Wisit Cheungpasitporn, and Suneet Chansangpetch. 2023. Personalized Care in Eye Health: Exploring Opportunities, Challenges, and the Road Ahead for Chatbots. *Journal of Personalized Medicine* 13, 12 (2023), 1679.
- [38] Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q. Vera Liao, Khai Truong, and Shwetak Patel. 2018. FarmChat: A Conversational Agent to Answer Farmer Queries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 170 (dec 2018), 22 pages. <https://doi.org/10.1145/3287048>

- [39] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [40] Anders Johansson, Monica Larsson, and Bodil Ivarsson. 2020. General practitioners' experiences of digital written patient dialogues: a pilot study using a mixed method. *Journal of Primary Care & Community Health* 11 (2020), 2150132720909656.
- [41] Anders Johansson, Monica Larsson, and Bodil Ivarsson. 2020. Patients' experiences with a digital primary health care concept using written dialogues: a pilot study. *Journal of Primary Care & Community Health* 11 (2020), 2150132720910564.
- [42] Natalie Joseph-Williams, Glyn Elwyn, and Adrian Edwards. 2014. Knowledge is not power for patients: A systematic review and thematic synthesis of patient-reported barriers and facilitators to shared decision making. *Patient Education and Counseling* 94, 3 (2014), 291–309. <https://doi.org/10.1016/j.pec.2013.10.031>
- [43] Anirudha Joshi, Girish Dalvi, Manjiri Joshi, Prasad Rashinkar, and Aniket Sarangdhar. 2011. Design and evaluation of Devanagari virtual keyboards for touch screen mobile phones. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. 323–332.
- [44] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 158, 18 pages. <https://doi.org/10.1145/3491102.3517533>
- [45] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-Woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 701, 20 pages. <https://doi.org/10.1145/3613904.3642937>
- [46] P Kinnersley, A Edwards, K Hood, N Cadbury, R Ryan, H Prout, D Owen, F Macbeth, P Butow, and C Butler. 2007. Interventions before consultations for helping patients address their information needs. *Cochrane Database Syst. Rev.* 2010, 3 (July 2007), CD004565.
- [47] Predrag Klasnja, Andrea Civan Hartzler, Kent T. Unruh, and Wanda Pratt. 2010. Blowing in the Wind: Unanchored Patient Information Work during Cancer Care. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) (CHI '10). Association for Computing Machinery, New York, NY, USA, 193–202. <https://doi.org/10.1145/1753326.1753355>
- [48] Predrag Klasnja, Andrea Civan Hartzler, Kent T. Unruh, and Wanda Pratt. 2010. Blowing in the wind: unanchored patient information work during cancer care. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) (CHI '10). Association for Computing Machinery, New York, NY, USA, 193–202. <https://doi.org/10.1145/1753326.1753355>
- [49] Matthew W Kreuter and Stephanie M McClure. 2004. The role of culture in health communication. *Annu. Rev. Public Health* 25, 1 (2004), 439–455.
- [50] Harsh Kumar, Yiyi Wang, Jiakai Shi, Ilya Musabirov, Norman AS Farb, and Joseph Jay Williams. 2023. Exploring the Use of Large Language Models for Improving the Awareness of Mindfulness. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.
- [51] Suzanne M. Kurtz. 2002. Doctor-Patient Communication: Principles and Practices. *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques* 29, S2 (2002), S23–S29. <https://doi.org/10.1017/S0317167100001906>
- [52] Zeina Atrash Leong, Michael S. Horn, Lisa Thaniel, and Emily Meier. 2018. Inspiring AWE: Transforming Clinic Waiting Rooms into Informal Learning Environments with Active Waiting Education. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3173672>
- [53] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper_2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf
- [54] Brenna Li, Ofek Gross, Noah Crampton, Mamta Kapoor, Saba Tauseef, Mohit Jain, Khai N. Truong, and Alex Mariakakis. 2024. Beyond the Waiting Room: Patient's Perspectives on the Conversational Nuances of Pre-Consultation Chatbots. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 438, 24 pages. <https://doi.org/10.1145/3613904.3641913>
- [55] Brenna Li, Tetyana Skoropad, Puneet Seth, Mohit Jain, Khai Truong, and Alex Mariakakis. 2023. Constraints and Workarounds to Support Clinical Consultations in Synchronous Text-based Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 342, 17 pages. <https://doi.org/10.1145/3544548.3581014>
- [56] David Li, Kartik Gupta, Mousumi Bhaduri, Paul Sathiadoss, Sahir Bhatnagar, and Jaron Chong. 2025. Comparative diagnostic accuracy of GPT-4o and LLaMA 3-70b: Proprietary vs. open-source large language models in radiology. *Clinical Imaging* 118 (2025), 110382. <https://doi.org/10.1016/j.clinimag.2024.110382>

- [57] Mojgan Lotfi, Vahid Zamanzadeh, Leila Valizadeh, and Mohammad Khajehgoodari. 2019. Assessment of nurse–patient communication and patient satisfaction from nursing care. *Nursing open* 6, 3 (2019), 1189–1196.
- [58] Nisrine N. Makarem and Jumana Antoun. 2016. Email communication in a developing country: different family physician and patient perspectives. *Libyan Journal of Medicine* 11, 1 (2016), 32679. <https://doi.org/10.3402/ljm.v11.32679> arXiv:<https://doi.org/10.3402/ljm.v11.32679> PMID: 28349842.
- [59] Lisa Mekioussa Malki, Dilisha Patel, and Aneesha Singh. 2024. "The Headline Was So Wild That I Had To Check": An Exploration of Women's Encounters With Health Misinformation on Social Media. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 128 (apr 2024), 26 pages. <https://doi.org/10.1145/3637405>
- [60] Lena Manykina, Elizabeth Mynatt, Patricia Davidson, and Daniel Greenblatt. 2008. MAHI: investigation of social scaffolding for reflective thinking in diabetes management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) (CHI '08). Association for Computing Machinery, New York, NY, USA, 477–486. <https://doi.org/10.1145/1357054.1357131>
- [61] Charles NJ McGhee, Jie Zhang, and Dipika V Patel. 2020. A perspective of contemporary cataract surgery: the most common surgical procedure in the world. *Journal of the Royal Society of New Zealand* 50, 2 (2020), 245–262.
- [62] Rebecca Moden. 2022. *Blank Page Syndrome And How To Beat It*. Retrieved Sept 1, 2023 from <https://www.ool.co.uk/blog/blank-page-syndrome-and-how-to-beat-it/>
- [63] Anouk Mols and Jason Pridmore. 2021. Always available via WhatsApp: Mapping everyday boundary work practices and privacy negotiations. *Mobile Media & Communication* 9, 3 (2021), 422–440. <https://doi.org/10.1177/2050157920970582>
- [64] Sara Montagna, Stefano Ferretti, Lorenz Cuno Klopfenstein, Antonio Florio, and Martino Francesco Pengo. 2023. Data decentralisation of LLM-based chatbot systems in chronic disease self-management. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*. 205–212.
- [65] Ministry of Health National Health Authority and Government of India Family Welfare. 2020. National Digital Health Mission: Strategy Overview. https://www.niti.gov.in/sites/default/files/2023-02/ndhm_strategy_overview.pdf
- [66] Government of India Office of the Registrar General & Census Commissioner, India; Ministry of Home Affairs. 2011. Census of India 2011: C-16: Population by mother tongue, Karnataka - 2011. <https://censusindia.gov.in/nada/index.php/catalog/10208/study-description> Accessed: 2024-07-24.
- [67] Chinasa T. Okolo, Srujana Kamath, Nicola Dell, and Aditya Vashistha. 2021. "It Cannot Do All of My Work": Community Health Worker Perceptions of AI-Enabled Mobile Health Applications in Rural India. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 701, 20 pages. <https://doi.org/10.1145/3411764.3445420>
- [68] OpenAI. 2024. Usage Policies. <https://openai.com/policies/usage-policies/> Accessed: 2024-06-18.
- [69] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emry Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman,

- Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillett, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 Technical Report. arXiv:arXiv:2303.08774
- [70] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Elizabeth Kazuiunas, Stina Matthiesen, and Farah Magrabi. 2021. Realizing AI in Healthcare: Challenges Appearing in the Wild. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (, Yokohama, Japan,) (CHI EA '21). Association for Computing Machinery, New York, NY, USA, Article 108, 5 pages. <https://doi.org/10.1145/3411763.3441347>
- [71] Trevor Perrier, Nicola Dell, Brian DeRenzi, Richard Anderson, John Kinuthia, Jennifer Unger, and Grace John-Stewart. 2015. Engaging Pregnant Women in Kenya with a Hybrid Computer-Human SMS Communication System. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) (CHI '15). Association for Computing Machinery, New York, NY, USA, 1429–1438. <https://doi.org/10.1145/2702123.2702124>
- [72] Laura Pfeifer Vardoulakis, Amy Karlson, Dan Morris, Greg Smith, Justin Gatewood, and Desney Tan. 2012. Using mobile phones to present medical information to hospital patients. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) (CHI '12). Association for Computing Machinery, New York, NY, USA, 1411–1420. <https://doi.org/10.1145/2207676.2208601>
- [73] Manya Prasad, Sumit Malhotra, Mani Kalaivani, Praveen Vashist, and Sanjeev K Gupta. 2020. Gender differences in blindness, cataract blindness and cataract surgical coverage in India: a systematic review and meta-analysis. *British Journal of Ophthalmology* 104, 2 (2020), 220–224.
- [74] Natalia Radionova, Eylem Ög, Anna-Jasmin Wetzel, Monika A Rieger, and Christine Preiser. 2023. Impacts of Symptom Checkers for Laypersons' Self-diagnosis on Physicians in Primary Care: Scoping Review. *Journal of Medical Internet Research* 25 (2023), e39219.
- [75] Niroop Channa Rajashekhar, Yeo Eun Shin, Yuan Pu, Sunny Chung, Kisung You, Mauro Giuffre, Colleen E Chan, Theo Saarinen, Allen Hsiao, Jasjeet Sekhon, Ambrose H Wong, Leigh V Evans, Rene F. Kizilcec, Loren Laine, Terika McCall, and Dennis Shung. 2024. Human-Algorithmic Interaction Using a Large Language Model-Augmented Artificial Intelligence Clinical Decision Support System. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 442, 20 pages. <https://doi.org/10.1145/3613904.3642024>
- [76] Lennart Seitz, Sigrid Bekmeier-Feuerhahn, and Krutika Gohil. 2022. Can we trust a chatbot like a physician? A qualitative study on understanding the emergence of trust toward diagnostic chatbots. *International Journal of Human-Computer Studies* 165 (2022), 102848. <https://doi.org/10.1016/j.ijhcs.2022.102848>
- [77] Kamaljeet Singh, Arshi Misbah, Pranav Saluja, and Arun Kumar Singh. 2017. Review of manual small-incision cataract surgery. *Indian journal of ophthalmology* 65, 12 (2017), 1281.
- [78] Meredith M. Skeels, Kenton T. Unruh, Christopher Powell, and Wanda Pratt. 2010. Catalyzing Social Support for Breast Cancer Patients. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) (CHI '10). Association for Computing Machinery, New York, NY, USA, 173–182. <https://doi.org/10.1145/1753326.1753353>
- [79] Vess Stamenova, Payal Agarwal, Leah Kelley, Jamie Fujioka, Megan Nguyen, Michelle Phung, Ivy Wong, Nike Onabajo, R Sacha Bhatia, and Onil Bhattacharyya. 2020. Uptake and patient and provider communication modality preferences of virtual visits in primary care: a retrospective cohort study in Canada. *BMJ open* 10, 7 (2020), e037064.
- [80] Statista. 2022. Most used social media and messaging apps in India 2022. <https://www.statista.com/statistics/1235799/india-most-used-social-media-and-messaging-apps/> Accessed: 2024-02-01.
- [81] Si Sun, Xiaomu Zhou, Joshua C. Denny, Trent S. Rosenbloom, and Hua Xu. 2013. Messaging to your doctors: understanding patient-provider communications via a portal system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) (CHI '13). Association for Computing Machinery, New York, NY, USA, 1739–1748. <https://doi.org/10.1145/2470654.2466230>
- [82] Paul C Tang and David Lansky. 2005. The missing link: bridging the patient-provider health information gap. *Health Aff. (Millwood)* 24, 5 (Sept. 2005), 1290–1295.
- [83] P C Tang and C Newcomb. 1998. Informing patients: a guide for providing patient health information. *J. Am. Med. Inform. Assoc.* 5, 6 (Nov. 1998), 563–570.
- [84] Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge Palacios, Rita Faia Marques, Cecily Morrison, and Gavin Doherty. 2023. Designing Human-centered AI for Mental Health: Developing Clinically Relevant Applications for Online CBT Treatment. *ACM Trans. Comput.-Hum. Interact.* 30, 2, Article 27 (mar 2023), 50 pages. <https://doi.org/10.1145/3564752>
- [85] Anja Thieme, Aditya Nori, Marzyeh Ghassemi, Rishi Bommasani, Tariq Osman Andersen, and Ewa Luger. 2023. Foundation Models in Healthcare: Opportunities, Risks & Strategies Forward. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (, Hamburg, Germany,) (CHI EA '23). Association for Computing Machinery, New York, NY, USA, Article 512,

- 4 pages. <https://doi.org/10.1145/3544549.3583177>
- [86] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.
- [87] John R. Tongue, Howard R. Epps, and Laura L. Forese. 2005. Communication skills for patient-centered care : Research-based, easily learned techniques for medical interviews that benefit orthopaedic surgeons and their patients. *Journal of Bone and Joint Surgery, American Volume* 87 (2005), 652–658. <https://api.semanticscholar.org/CorpusID:71035586>
- [88] Yuan-Chi Tseng, Weerachaya Jarupreechachan, and Tuan-He Lee. 2023. Understanding the Benefits and Design of Chatbots to Meet the Healthcare Needs of Migrant Workers. 7, CSCW2, Article 315 (oct 2023), 34 pages. <https://doi.org/10.1145/3610106>
- [89] Anupriya Tuli, Shaan Chopra, Neha Kumar, and Pushpendra Singh. 2018. Learning from and with Menstrupedia: Towards Menstrual Health Education in India. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 174 (nov 2018), 20 pages. <https://doi.org/10.1145/3274443>
- [90] Ding Wang, Santosh D. Kale, and Jacki O'Neill. 2020. Please Call the Specialist: Using WeChat to Support Patient Care in China. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376274>
- [91] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. “Brilliant AI Doctor” in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 697, 18 pages. <https://doi.org/10.1145/3411764.3445432>
- [92] Liuping Wang, Dakuo Wang, Feng Tian, Zhenhui Peng, Xiangmin Fan, Zhan Zhang, Mo Yu, Xiaojuan Ma, and Hongan Wang. 2021. CASS: Towards Building a Social-Support Chatbot for Online Health Community. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 9 (apr 2021), 31 pages. <https://doi.org/10.1145/3449083>
- [93] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.
- [94] Lauren Wilcox, Dan Morris, Desney Tan, and Justin Gatewood. 2010. Designing Patient-Centric Information Displays for Hospitals. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) (CHI '10). Association for Computing Machinery, New York, NY, USA, 2123–2132. <https://doi.org/10.1145/1753326.1753650>
- [95] Ziang Xiao, Q. Vera Liao, Michelle Zhou, Tyrone Grandison, and Yunyao Li. 2023. Powering an AI Chatbot with Expert Sourcing to Support Credible Health Information Access. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) (IUI '23). Association for Computing Machinery, New York, NY, USA, 2–18. <https://doi.org/10.1145/3581641.3584031>
- [96] Deepika Yadav, Preerna Malik, Kirti Dabas, and Pushpendra Singh. 2019. Feedpal: Understanding Opportunities for Chatbots in Breastfeeding Education of Women in India. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 170 (nov 2019), 30 pages. <https://doi.org/10.1145/3359272>
- [97] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 73 (may 2024), 35 pages. <https://doi.org/10.1145/3659625>
- [98] Qunru Ye, Yanxian Chen, William Yan, Wei Wang, Jingxian Zhong, Cong Tang, Andreas Müller, Bo Qiu, et al. 2020. Female gender remains a significant barrier to access cataract surgery in South Asia: A systematic review and meta-analysis. *Journal of ophthalmology* 2020 (2020).
- [99] Ibrahim Edhem Yilmaz and Levent Dogan. 2024. Talking technology: exploring chatbots as a tool for cataract patient education. *Clinical and Experimental Optometry* (2024), 1–9.
- [100] Pradyumna YM, Vinod Ganesan, Dinesh Kumar Arumugam, Meghna Gupta, Nischith Shadagopan, Tanay Dixit, Sameer Segal, Pratyush Kumar, Mohit Jain, and Sriram Rajamani. 2023. Pwr: Exploring the Role of Representations in Conversational Programming. *arXiv:2309.09495 [cs.HC]*
- [101] Yan Zhang. 2010. Contextualizing consumer health information searching: an analysis of questions in a social Q&A community. In *Proceedings of the 1st ACM international health informatics symposium*. 210–219.

A APPENDIX

A.1 CataractBot Implementation Details

A.1.1 CataractBot LLM Prompts. The *CataractBot* system leverages LLM (GPT-4 in our case) for these four tasks (Table 2):

- (1) **Response Generation:** For every medical and logistical question asked by the patient/attendant, the system performs a vector search on the Knowledge Base (KB) to extract the three most relevant data chunks

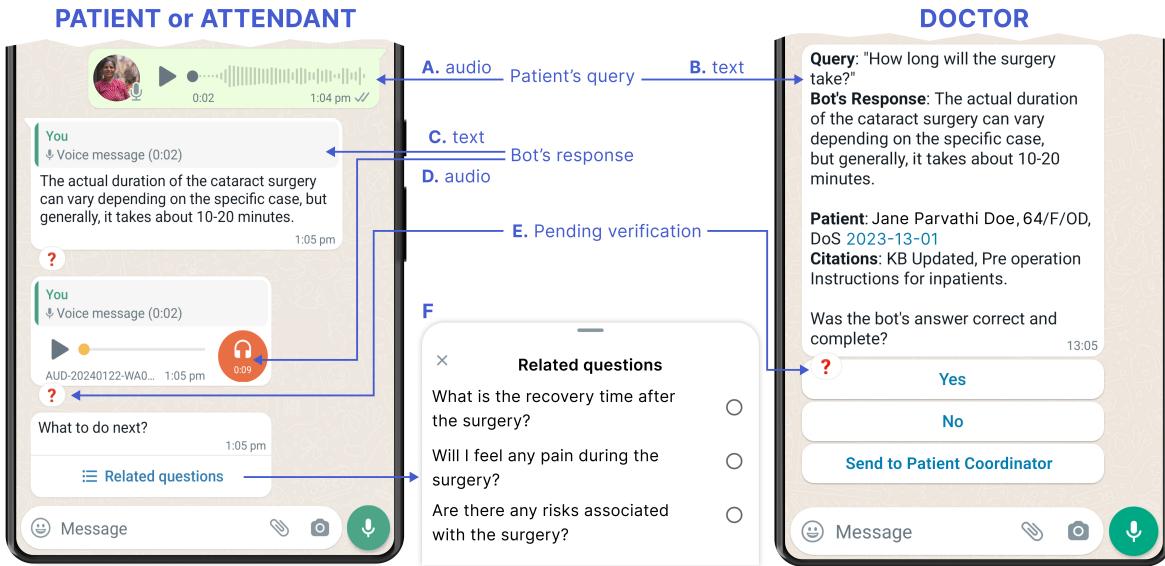


Fig. 4. A question asked (using audio), receiving an unverified response and Related Questions from *CataractBot*.

related to the query. The LLM is then prompted (Table 2) to extract an answer for the query from these data chunks. Note: Question Classification is part of the same LLM call to improve efficiency.

- (2) **Related Questions Generation:** For every medical and logistical question, the system prompts LLM to generate three related questions based on the preceding query. Note: The 72 character limit is due to the WhatsApp's message limit in interactive suggestions.
- (3) **Final Response Generation:** If the expert marks the initial LLM-generated response as incorrect or incomplete, the system prompts LLM to generate the final response by merging both the initial response and the expert's edit suggestion.
- (4) **Shorten Response:** If the generated response exceeds WhatsApp's message limit of 700 characters, the system prompts LLM to summarize it within the specified character limit.

A.1.2 *CataractBot's Support for Multimodal Communication*. See Figure 4.

A.2 Participant Demography

See Tables 3 and 4.

A.3 Chatbot Usability Evaluation Form

Responses were given on a 5-point Likert scale, ranging from *Strongly Disagree* to *Strongly Agree*.

- (1) CataractBot understands me well.
- (2) CataractBot's responses are easy to understand.
- (3) CataractBot's responses were useful, appropriate, and informative.
- (4) CataractBot responds quickly.
- (5) CataractBot seems to have a good grasp of medical knowledge.

Table 2. LLM Prompts used in CataractBot.

	System Prompt	Query Prompt
Response Generation	<p>You are a Cataract chatbot whose primary goal is to help patients undergoing or undergone a cataract surgery. If the query can be truthfully and factually answered using the knowledge base only, answer it concisely in a polite and professional way. If not, then just say "I do not know the answer to your question. If this needs to be answered by a doctor, please schedule a consultation."</p> <p>In case of a conflict between raw knowledge base and new knowledge base, prefer the new knowledge base. One exception to the above is if the query is a greeting or an acknowledgement or gratitude. If the query is a greeting, then respond with a greeting. If the query is an acknowledgement or gratitude to the bot's response, then respond with an acknowledgement of the same. Some examples of acknowledgement or gratitude to the bot's response are "Thank You", "Got it" and "I understand". In addition to the above, indicate like a 3-class classifier if the query is "medical", "logistical" or "small-talk". Here, "small-talk" is defined as a query which is a greeting or an acknowledgement or gratitude. Answer it in the following json format:</p>	<p>The following knowledge base have been provided to you as reference: Raw documents are as follows: <relevant chunks string> New documents are as follows: <relevant updated chunks string> The most recent conversations are here: <conversation string> You are asked the following query: <user query></p> <p>Ensure that the query type belongs to only the above mentioned three categories. When not sure, choose one of "medical" or "logistical".</p>
Related Questions Generation	What are three possible follow-up questions the patient might ask? Respond with the questions only in a python list of strings. Each question should not exceed 72 characters.	<p>A patient asked the following query: <query> A chatbot answered the following: <response></p>
Final Response Generation	You are a Cataract chatbot whose primary goal is to help patients undergoing or undergone a cataract surgery. A cataract patient asks a query and a cataract chatbot answers it. But, the doctor gives a correction to the chatbot's response. Update the cataract chatbot's response by taking the doctor's correction into account. Respond only with the final updated response.	<p>A cataract patient asked the following query: <query> A cataract chatbot answered the following: <response> A doctor corrected the response as follows: <correction></p>
Shorten Response	You are a Cataract chatbot, and you have to summarize the answer provided by a bot. Please summarise the answer in 700 characters or less. Only return the summarized answer and nothing else.	<p>You are given the following response: <response></p>

Table 3. Demographic details for 19 patients and 30 attendees, with their interview participation, the duration of *CataractBot* usage (calculated as the difference between the first and last day of messages sent), and the total number of messages sent.

PId	Age	Sex	Language	Highest Education	AId	Age	Sex	Language	Highest Education	Surgery-day Interview	Post-surgery Interview	Days	# of Messages
					A1	43	M	English	Masters	Yes	Yes	7	9
					A2	31	F	Kannada	Grade 12	Yes	Yes	3	13
P3	56	M	Tamil	≤Grade 10	A3	35	M	English	Bachelors	Yes	No	4	6
P4	52	M	English	Grade 12	A4	39	F	English	Grade 12	Yes	No	4	8
P5	67	M	Kannada	≤Grade 10	A5	40	M	Kannada	Bachelors	Yes	No	5	4
					A6	35	M	Telugu	≤Grade 10	Yes	No	1	2
P7	58	M	Telugu	≤Grade 10	A7	25	M	English	Masters	Yes	Yes	23	17
					A8	39	M	Hindi	Bachelors	Yes	No	6	6
P9	69	M	English	Masters	A9	38	F	English	Masters	Yes	Yes	6	14
					A10	71	M	English	Bachelors	Yes	Yes	26	15
					A11	32	M	Hindi	≤Grade 10	Yes	No	14	6
P12	65	M	Hindi	Bachelors	A12	31	F	English	Masters	Yes	Yes	2	6
P13	69	M	Tamil	Grade 12	A13	35	M	English	Masters	Yes	Yes	25	18
P14	56	F	Kannada	≤Grade 10	A14	33	F	English	Bachelors	Yes	Yes	22	30
					A15	57	M	English	Grade 12	Yes	No	4	4
P16	57	F	Hindi	≤Grade 10	A16	25	M	English	Masters	Yes	No	1	1
					A17	39	M	English	Bachelors	Yes	Yes	2	8
P18	56	M	English	Grade 12	A18	23	M	English	Masters	Yes	Yes	27	24
P19	41	M	English	Masters	A19	36	F	English	Bachelors	Yes	No	19	15
P20	64	M	Tamil	Grade 12	A20	60	F	Tamil	Grade 12	Yes	No	4	16
					A21	43	M	English	Masters	Yes	No	15	19
					A22	35	F	English	Bachelors	Yes	No	7	11
P23	54	M	English	Bachelors								7	8
P24	66	M	Kannada	Bachelors	A24	37	M	English	Masters	No	No	1	1
					A25	40	F	Tamil	≤Grade 10	No	No	7	3
P26	65	M	Hindi	Bachelors	A26	34	M	English	Bachelors	No	No	19	15
P27	64	M	Hindi	Bachelors	A27	35	M	English	Bachelors	No	No	18	14
					A28	40	M	English	Masters	No	No	1	1
P29	45	F	Kannada	≤Grade 10	A29	29	F	English	Bachelors	No	No	15	10
P30	50	M	Hindi	Masters	A30	25	F	English	Bachelors	No	No	5	3
P31	63	M	Kannada	Bachelors	A31	32	M	English	Bachelors	No	No	15	36

Table 4. Demography details of doctors and patient coordinators who participated in the formative and/or deployment study

Id	Role(s)	Age	Sex	Highest Education	Experience	Surgeries per week	Formative study?	Deployment study?
D1	Operating doctor	42	F	Masters	10+ years	10-20	No	Yes
D2	Operating doctor	44	F	Masters	15+ years	10-20	No	Yes
D3	Operating doctor	46	M	Masters	20+ years	30-40	Yes	Yes
D4	Escalation doctor, Knowledge base expert	46	M	Masters	20+ years	15-20	Yes	Yes
D5	Doctor	30	M	Masters	6+ years	10	Yes	No
D6	Doctor	29	F	Masters	2+ years	1-2	Yes	No
C1	Operating patient coordinator	33	F	Bachelors	10+ years	N/A	Yes	Yes
C2	Escalation patient coordinator	36	F	Bachelors	15+ years	N/A	No	Yes
C3	Patient coordinator	45	F	Bachelors	10+ years	N/A	Yes	No

- (6) CataractBot is kind and helpful.
- (7) CataractBot is easy to use.
- (8) I would be willing to use CataractBot (or a similar bot before/after a major surgical treatment) in future.

Table 5. Key findings from CataractBot deployment study

Research Q	Theme	Finding (Information Seekers)	Finding (Experts)
RQ1: Information Needs	Reasons for usage	<ul style="list-style-type: none"> Asking forgotten or uncomfortable questions Clarifying and verifying information Seeking updates 	<ul style="list-style-type: none"> Commitment to patients
	Reasons for lack of usage	<ul style="list-style-type: none"> In-person expert interactions suffice Lack of questions “I don’t know” responses 	<ul style="list-style-type: none"> Persistent medium gives rise to accountability concerns
RQ2: Features	LLM-powered	<ul style="list-style-type: none"> Instant answers. Saves time. Understands complex/ill formed questions 	<ul style="list-style-type: none"> Accurate answers (Informal) corrections are quick and easy
	Experts-in-the-loop	<ul style="list-style-type: none"> Drives trust and engagement Lowers perceived intelligence of bot 	<ul style="list-style-type: none"> Adds a layer of privacy from patients Corrections expand bot’s answers No access to final answer reduces both workload and transparency
	Multimodality and multilinguality	<ul style="list-style-type: none"> Independent usage by older, less educated, visually impaired patients 	<ul style="list-style-type: none"> Hard to decipher English transcriptions
	Support for multiple stakeholders	<ul style="list-style-type: none"> Attendants (younger, educated) message more than patients 	<ul style="list-style-type: none"> Escalation experts are crucial, but lack patient context
	Knowledge base		<ul style="list-style-type: none"> Minimal edits after first expert verification Bot’s performance improves over time Experts’ workload eases
RQ3: Workflow	Workflows	<ul style="list-style-type: none"> Messages spread throughout hours of day Most messages on surgery day Waiting for verification is fine 	<ul style="list-style-type: none"> Verify at their convenience WhatsApp blurs work-home boundaries
	Personalization vs. privacy	<ul style="list-style-type: none"> No personalization implies limited usefulness Medical history not seen as private by some Personalization wanted on opt-in basis 	<ul style="list-style-type: none"> Lack of medical history makes verification challenging Personalization would add complexity and workload

A.4 Summary of Key Findings

See Table 5