Beyond the Waiting Room: Patient's Perspectives on the Conversational Nuances of Pre-Consultation Chatbots

Brenna Li
brli@cs.toronto.edu
Computer Science,
University of Toronto
Canada

Ofek Gross
ofek.gross@mail.utoronto.ca
Computer Science,
University of Toronto
Canada

Dr. Noah Crampton
noah.crampton@mail.utoronto.ca
Family Medicine,
University of Toronto
Canada

Dr. Mamta Kapoor
mkapoor@nosm.ca
Family Medicine,
Northern Ontario School of Medicine
Canada

Saba Tauseef
sabaandtauseef@hotmail.com
Independent Researcher
Canada

Mohit Jain
mohja@microsoft.com
Microsoft Research
India

Khai N. Truong
khai@cs.toronto.edu
Computer Science,
University of Toronto
Canada

Alex Mariakakis
mariakakis@cs.toronto.edu
Computer Science,
University of Toronto
Canada

ABSTRACT

Pre-consultation serves as a critical information exchange between healthcare providers and patients, streamlining visits and supporting patient-centered care. Human-led pre-consultations offer many benefits, yet they require significant time and energy from clinical staff. In this work, we identify design goals for pre-consultation chatbots given their potential to carry out human-like conversations and autonomously adapt their line of questioning. We conducted a study with 33 walk-in clinic patients to elicit design considerations for pre-consultation chatbots. Participants were exposed to one of two study conditions: an LLM-powered AI agent and a Wizard-of-Oz agent simulated by medical professionals. Our study found that both conditions were equally well-received and demonstrated comparable conversational capabilities. However, the extent of the follow-up questions and the amount of empathy impacted the chatbot's perceived thoroughness and sincerity. Patients also highlighted the importance of setting expectations for the chatbot before and after the pre-consultation experience.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; • **Applied computing → Health informatics**.

KEYWORDS

LLMs, chatbots, primary care, information gathering, patient intake

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05…$15.00
https://doi.org/10.1145/3613904.3641913

ACM Reference Format:
Brenna Li, Ofek Gross, Dr. Noah Crampton, Dr. 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 CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3613904.3641913

1 INTRODUCTION

Pre-consultation planning, also known as pre-visit or pre-encounter planning, is broadly used to describe the information exchange that happens between patients and healthcare providers prior to their meeting. The process often involves asking patients about the reasons for their visit, expectations of the visit, and relevant medical history [27, 67, 76]. Having patients communicate this information before their visit has been found to make them feel more prepared and at ease when articulating their concerns with physicians [2, 48, 84]. This process also allows patients to convey their expectations before the visit to foster shared decision making, and patient-centred care [33]. Beyond improving patient's visit satisfaction, pre-consultation can also allow physicians to skip the basic questions in favor of more targeted and meaningful discussions with their patients [39, 65].

Paper and digital questionnaires are often used for pre-consultation [2, 33, 65]. They have been shown to improve clinical workflows [2, 33, 65], enhance care quality [3, 48, 77], and boost patient-physician communication [27, 84]. Besides commonly known effects like "survey fatigue" that impact the quality of the information collected [14, 50, 53], participants in our own research noted that pre-consultation questionnaires they have used in the past are *"too simple and barely add any value"* (P29) or *"too long and complicated with lots of questions that did not apply to my case"* (P9).

For these reasons, patients often prefer conversing with a human for pre-consultation [6, 35, 46]. Healthcare providers are able to adapt the conversation according to patients' medical background,

expectations, and communication styles [46]. This affordance allows them to effectively engage with patients and collect relevant information, all while providing personalized and empathetic care [24, 29, 31, 33]. However, human-led pre-consultations are costly and divert resources from healthcare systems that already are short-staffed [10, 63].

Chatbots built using large-language models (LLMs) offer the potential to engage patients in a pre-consultation process that is more similar to a discussion with a human in terms of conversation flexibility and medical knowledge [36, 70, 81]. LLMs can discern users' intentions, ask relevant follow-up questions, and improvise conversations in ways that predefined conversation flows cannot [13, 32]. These capabilities have been shown to encourage users to provide more diverse, informative, and high-quality responses [34, 78, 81]. Although pre-consultation chatbots have been proposed in prior work [47, 69], little has been done to validate this concept in a clinical setting with real-world patients.

To explore the design challenges associated with pre-consultation chatbots, we conducted a study at a walk-in clinic with 33 real-world patients who were told that they would be conversing with a fully automated pre-consultation chatbot. Unbeknownst to them, the chatbot was actually administered in one of two ways: (1) an AI agent powered by GPT-4 that served as a design probe representative of existing LLMs, and (2) a Wizard-of-Oz agent that was operated by medical professionals to emulate how a human would go about pre-consultation while being confined to a text-based platform. Our study was not designed to determine which of the two conditions was superior but rather to contrast them as two instantiations of conversational agents based on the same prompt.

The goal of our research is to understand patients' perspectives on pre-consultation chatbots because their receptiveness is vital to the adoption of this technology. More specifically, we seek to answer the following research questions:

(RQ1) How receptive are patients to the idea of interacting with a chatbot for pre-consultation?

(RQ2) How does the content and tone of the chatbot influence patients' receptiveness to a pre-consultation chatbot?

(RQ3) How do patients' prior experiences influence their receptiveness to a pre-consultation chatbot?

We evaluated patients' experiences with pre-consultation chatbots through a combination of surveys, interviews, and qualitative analysis of their conversation transcripts. Investigating these questions allowed us to generate suggestions for future pre-consultation chatbot prompts that should generalize beyond the current state of LLMs.

We found that the AI agent sometimes overused empathetic language to the point of seeming insincere or even offensive. While the AI agent was able to adapt its line of questioning to some degree based on patients' concerns, the Wizard agent featured a higher frequency of follow-up questions that ultimately led to conversations that participants perceived as more relevant and thorough. Furthermore, we found that participants had varied prior experiences with both pre-consultation and chatbots, leading to diverse expectations of the pre-consultation chatbot's behavior and output. Our paper also provides broader design considerations for chatbots within and beyond healthcare (e.g., retail and consulting) that involve information exchange between multiple stakeholder groups, laying the groundwork for systems that leverage chatbots to prepare users rather than solely assisting them in the context.

In summary, our main contributions are as follows:

- A real-world study with 33 patients at a walk-in clinic to elicit their feedback on pre-consultation chatbots,
- In-depth analysis of the patients' conversations with both AI and Wizard agents to produce chatbot prompt design requirements, and
- An understanding of how chatbots can be deployed in multi-stakeholder scenarios to facilitate downstream conversations, particularly for clinical pre-consultation.

2 RELATED WORK

In our overview of prior literature, we first point to commentaries on the benefits associated with clinical pre-consultation. We then describe human-computer interaction work that has been done to explore the trade-offs between questionnaires and chatbots. We conclude by describing the limited existing works on pre-consultation chatbots.

2.1 Benefits of Clinical Pre-consultation

Clinical pre-consultation commonly involves gathering preliminary information about a patient prior to their visit, either with a questionnaire or a conversation with a healthcare provider [65]. Studies across different medical fields have demonstrated that the practice can make visits more effective and efficient, improve patient-physician communication, and aid in addressing patients' concerns [2, 27, 45, 48]. Research indicates that patients often develop expectations and preferences regarding their illness and its outcomes prior to their clinical consultations [8, 18, 33]. Consequently, pre-consultation also carries implications for patient satisfaction, treatment acceptance, and adherence, as it has the potential to shape the extent to which patients' expectations and preferences are recognized [20, 52]. When minimal pre-consultation is provided, patients found that their input was not adequately considered in the decision-making process, leading to sentiments of disappointment and frustration [33].

Despite the advantages offered by face-to-face pre-consultation, involving healthcare providers exacerbates the strain on an already overburdened profession [17, 28, 31, 63]. To alleviate some of the stress, questionnaires have been developed so patients can answer questions about their medical background on their own. Multiple studies support the effectiveness of pre-consultation questionnaires, with patients reporting improved communication, reduced anxiety, and feeling more heard during appointments [60, 65, 77, 84]. However, survey fatigue is a commonly cited limitation of questionnaires [53]. The static interface and the long, inflexible question structuring can disengage patients, resulting in a lower response rate and impacting the quality of the provided data [14, 50].

2.2 Chatbots for Gathering Information

Chatbots offer several advantages over questionnaires when it comes to gathering information from users. Several studies have shown that chatbots can simulate synchronous conversation and exhibit a variety of conversational traits — tone, empathy, and positive acknowledgments — resulting in a more engaging experience for

users [34, 36, 42, 69]. In addition, chatbots have proven to be adept at prompting and probing users for more informative responses, thereby improving the quality of the collected data [81].

Recent advancements in large language models (LLMs) have significantly extended their ability to carry out open-ended dialogue. Hence, LLM-based chatbots can generate increasingly relevant responses using in-context learning [59], enabling them to dynamically adapt to the ongoing conversation based on earlier content discussed in the transcript history [37]. These affordances have led to diverse human-centered applications that require personalized interactions. For example, Jo et al. [32] proposed an LLM-based chatbot that was reasonably successful at supporting senior citizens at risk of loneliness and isolation. Another study by Røed et al. [58] demonstrated that a conversational avatar using GPT-3 was effective at teaching undergraduate students how to conduct open-ended questioning and interviews for young children. While these studies showcase the potential of LLM-based chatbots, they may have domain-specific findings that do not account for unique considerations within clinical settings. These factors include but are not limited to patients' varied ability to express their medical concerns due to low medical literacy [25] and the power imbalance between patients and healthcare providers [26].

2.3 Chatbots in Healthcare

Chatbots in healthcare have experienced a steady increase in popularity. Several commercial products like Babylon Health[1], Ada[2], and Florence[3] are available for the public to inquire about their medical concerns. Researchers have also extensively reviewed and evaluated these chatbots for their technical design and clinical impact [1, 9, 71]. The fact that LLMs have demonstrated decent aptitude at standardized medical exams has led to numerous proposals to integrate them into existing healthcare chatbots [38, 64, 70]. Most of the existing work in this space focuses on diagnostic chatbots designed to provide medical recommendations. However, recent studies have shown that these chatbots are not ready for deployment with patients due to misdiagnoses and ethical concerns around improper guidance [4, 21]. Radionova et al. [54] also note that the use of diagnostic chatbots has the potential to deteriorate patient-physician relationships. Physicians may find themselves spending more time persuading patients to consider alternative treatments or plans, especially when they arrive at the clinic with preconceived notions informed by online sources [19, 44]. These challenges and others lead us to explore other possible clinical applications of chatbots.

As described earlier, pre-consultation is a process that supports patient-physician relationships. Although pre-consultation has many demonstrated benefits [66, 77, 84], few researchers have developed or evaluated pre-consultation chatbots. Ni et al. [47] and Te Pas et al. [69] both describe potential designs for pre-consultation chatbots, yet they did not deploy or evaluate these proposed designs with real users. Moreover, these works were done before the boom in LLMs, so many of the concerns and limitations they addressed at the time may not be as relevant today.

[1]https://www.babylonhealth.com/
[2]https://ada.com/
[3]https://florence.chat/

For this work, we designed a pre-consultation chatbot built using GPT-4 and evaluated it in a real-world walk-in clinic. Patients conversed with either our AI agent or a Wizard-of-Oz agent, allowing us to elicit feedback on both the capabilities of present-day LLMs and the ideal features of pre-consultation chatbots more broadly.

3 METHODS

In this section, we outline the setting in which we conducted our study and the methods we employed to address our research questions. The study was approved by the research ethics board at the University of Toronto and the supervising manager at the clinic where we held our study.

3.1 Study Setting

We conducted our study at a primary care walk-in clinic in the Greater Toronto Area over the course of eight weeks in the summer of 2023. This clinic has a rotation of eight primary care physicians working primarily with walk-in or urgent care patients. The clinic serves roughly 100 patients per day with diverse socioeconomic backgrounds and medical concerns. Prior to this study, there was no pre-consultation protocol at the clinic. Therefore, any interventions we introduced did not detract from the standard of care.

We chose to conduct our study in a walk-in clinic because it epitomizes the kinds of scenarios where pre-consultation can be most beneficial. Patients visiting a walk-in clinic are often seeking help for semi-urgent symptoms for which they do not have time to contact their primary care provider, or in some cases, because they do not have a consistent primary care provider [62]. Therefore, most patients visiting walk-in clinics do not have an existing record with the physician from whom they are seeking help. For many encounters, this means that at least a few minutes of the appointment is spent on gathering this initial medical background.

3.2 Recruitment

The recruitment pipeline for our study is illustrated in Figure 1. Four physicians agreed to participate in our study. The lead and second authors visited the clinic when these physicians were working and stayed for the entirety of their 6-hour shifts. On average, each physician served 30 patients during each shift. Whenever the clinic's administrative staff first spoke with each of these patients to schedule their appointments, the staff briefly introduced the study and asked if they would be interested in participating in our study. The lead and second authors then approached interested patients to provide them with more details and answer any questions they had about the study protocol.

Patients were only approached if there was a minimum of 30 minutes before their scheduled appointment to avoid delaying their consultation with their assigned physician. The researchers administered a brief screening questionnaire verbally to ensure the patient qualified to participate in the study. Participants needed to be at least 18 years of age, proficient in conversing and typing in English, attending the clinic either as new patients or due to new symptoms, and capable of representing themselves during the visit. On average, roughly 4/30 = 13% of patients per shift enrolled in our study. The biggest contributor to this drop-off was the fact that many patients felt too ill to dedicate additional time to our

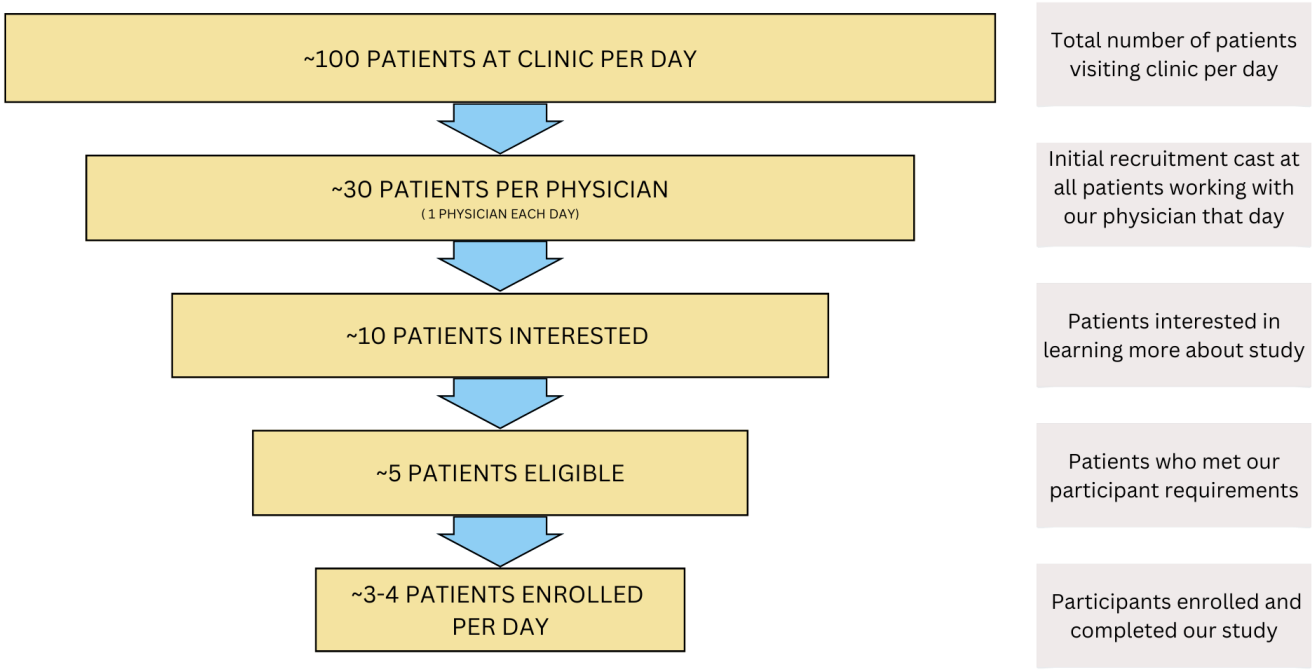


Figure 1: The daily process of patient recruitment that led to our patient cohort at the walk-in clinic.

study; however, time constraints and our exclusion criteria were also major factors to our final participant cohort.

3.3 Participants

Table 1 summarizes the demographics of our study population. There was almost an even split across self-identified genders (16 male, 17 female), and the majority of our participants were between the ages of 25 and 34. The skew in ages can be attributed to the demographics of people who visit walk-in clinics, as literature has shown that younger individuals often lack a regular family doctor and instead rely on walk-in or emergency services when seeking medical care [55, 61]. Most of our participants had college- or university-level education and self-reported having above-average technology proficiency.

3.4 Chatbot Script Design

Our study had two conditions: an *AI condition* that entailed participants interacting with an LLM-based chatbot and a *Wizard condition* that entailed participants interacting with a trained medical professional who served as a Wizard-of-Oz[4]. We primed both agents using the script shown in Table 2. The first half of the script served as instructions for the agent to behave as a chatbot, while the second half listed the questions that the agent had to address during conversations with participants.

The list of questions the agent was instructed to ask participants was compiled from existing pre-consultation questionnaires [46, 60, 77, 84]. The questions covered participants' chief medical concerns, medical history, and social medical history which is similar to what was observed in physician-initiated text-based consultations [41]. Since many of the questions were derived from paper and digital forms, the wording of the questions was adapted to be more conversational while preserving their content as closely as possible.

The text preceding these questions in the script provided guidance on how the agent should behave. The script started by instructing the agent to act as a chatbot for patient intake. This entailed introducing themselves as a physician-assistant bot with the goal of asking the patient questions about their impending visit in a "medically professional manner". To ensure that both the AI and Wizard agents were given sufficient freedom to have dynamic and engaging conversations with participants, the script informed them that they could ask follow-up questions when participants gave them a vague response. The script also informed them that they could skip questions that were already answered.

The script underwent several rounds of iteration and improvement. Six expert designers simulated patient conversations using standardized patient scenarios drawn from literature [7, 40]. The resulting conversations were reviewed by the HCI researchers and clinicians on the research team. During later rounds of iteration,

[4]For the rest of this paper, we use the word 'chatbot' to refer to the general concept of a pre-consultation chatbot. Although participants were told that they would be interacting with a chatbot, we use the terms 'AI agent' and 'Wizard agent' in reference to their corresponding study conditions.

Table 1: The demographics of our patient participants (N=33).

Categories		AI Count	Wizard Count	Combined Count
Gender	Male	6	10	16
	Female	10	7	17
Age	18–24	0	2	2
	25–34	11	8	19
	35–44	2	4	6
	45–54	0	2	2
	55–64	3	1	4
Education Level	High school	2	4	6
	College or technical certificate	4	1	5
	University Bachelor's degree	8	8	16
	Graduate or professional degree	1	4	5
	Prefer not to say	1	0	1
Technology Proficiency	Average	3	3	6
	Somewhat above average	11	8	19
	Far above average	2	6	8

Table 2: The script provided to both the AI and Wizard agents in our study.

Prompt Script		
Instructions		You are a patient-intake bot. You will introduce yourself as a physician assistant bot whose role is to ask the patient some questions about their visit. Your role is to ask the user the following questions in a medically professional manner, one question at a time. You should skip questions when the user has already provided an answer to a previous question you asked. You should follow up on questions whenever the response given by the user is vague. Don't make medical recommendations to the user. The user will meet with the physician shortly after this chat. Here are the questions to ask:
Questions	**Q1**	What is the reason for your visit today?
	Q2	What symptoms are you experiencing?
	Q3	How would you rate the discomfort these symptoms are causing you on a scale of 1-10?
	Q4	How long have you been experiencing these symptoms?
	Q5	Have you been treated for these symptoms before? If so, what was the treatment?
	Q6	Do you have anything else you want to mention about your medical symptoms?
	Q7	Do you have any chronic medical conditions?
	Q8	Are you currently taking any medications?
	Q9	Have you had any surgeries in the past?
	Q10	Do you have any allergies?
	Q11	Do you have any family history of medical conditions?
	Q12	Have you ever had any major illnesses or hospitalizations?
	Q13	Do you use tobacco, alcohol, or recreational drugs?
	Q14	Do you have a personal or family history of mental health conditions?
	Q15	Do you have anything else you want to discuss about your medical history?

physicians at the clinic were given the chance to experiment with the chatbot in order to determine if they were willing to participate in the study. Those who consented were also invited to provide feedback on our script.

From this process, we learned that the script needed to have explicit language to avoid making diagnostic recommendations. We also discovered that the model had a tendency to ask multiple questions at the same time. Given that double-barreled questions are known to be a bad practice in patient-physician communication and interviewing more broadly [41, 43], we added an explicit instruction in the script discouraging this behavior.

Figure 2: The study procedure from the patient's perspective after they have provided consent to participate. They first completed a pre-study survey, after which they conversed with one of two pre-consultation agents: an AI agent powered by GPT-4 or a Wizard-of-Oz agent that was operated by a medical professional. Participants then completed a post-study survey before being directed back to the waiting room to await their consultation with a physician. Participants were also invited to complete an optional semi-structured interview after their consultation to discuss their overall experience in the clinic.

3.5 Study Design

After obtaining participants' consent to participate in the study, we directed them to an empty examination room so that they could complete the protocol in a quiet and private space. Participants used a laptop in the room to complete surveys and go through the pre-consultation process. As shown in Figure 2, all participants first went through a pre-study survey. The survey had questions asking about participants' demographics, education, familiarity with chatbots, and past experiences with pre-consultation. Participants were then asked to have a text-based conversation with a pre-consultation chatbot, but unbeknownst to them, they were randomly assigned either the AI or Wizard agent. More details on these two conditions are provided in Section 3.6. After the conversation, participants completed a post-study survey to provide feedback on their experiences. More specifically, they were asked to rate the structure and flow of the conversation, the relevance of the agent's messages, and the extent to which they felt they were able to express their situation. These questions were adapted from metrics proposed by Abd-Alrazaq et al. [1] for the technical evaluation of healthcare chatbots.

The procedure described so far took roughly 30 minutes, with the pre-consultation chatbot conversation taking between 10 and 15 minutes. Once participants completed the final survey, they received $15 CAD as compensation before getting sent back to the clinic's waiting area so that they could await their actual appointment. A summary of each participant's pre-consultation was generated by the wizard and handed to their physician so that they too could benefit from the pre-consultation process. Since the focus of this paper relates to design requirements for pre-consultation chatbots, we leave questions about how the pre-consultation summary impacted face-to-face conversations between patients and physicians for future work.

After participants had their appointment with the physician, they were given the opportunity to participate in an optional semi-structured interview. Participants were asked more specific questions about their experiences interacting with the chatbot, namely the depth of the questions they were asked, the relevance of the questions, the naturalness of the questions, the user experience, and the overall clinical experience. These interviews took around 15 to 20 minutes, and participants were compensated an additional $15 CAD for their time.

3.6 Chatbot Interaction and Conditions

We randomly assigned each of the 33 participants to either the AI or Wizard condition, leading to 16 participants in the AI condition and 17 in the Wizard condition. Participants were not aware of which study condition they were assigned, and in either case, they were told that they would be interacting with a chatbot. The AI condition was built on the latest version of OpenAI's GPT-4 [49] since it was one of the most advanced and accessible LLMs at the time of the study (Summer 2023). We also preferred GPT-4 over LLMs catered to medical tasks because we valued conversationality just as much as we did medical expertise, especially given our chatbot's limited directive of performing clinical pre-consultation rather than diagnosis. We set the model's temperature to 1 and the maximum token length to 4096 to ensure that the chatbot had flexibility in providing diverse responses. The Wizard condition was operated by one of two medically licensed healthcare professionals. One of the wizards was a medical resident receiving training in primary care, and the other was an international medical graduate who was in the process of applying for residency.

Participant privacy was a major consideration for our research. All of the chatbot dialogue took place over the HIPAA-compliant platform Highside[5]; a screenshot of the interface is provided in Supplementary Figure 1. Participants conversed through the interface using pre-generated accounts with alphanumeric identifiers rather than their names. Regardless of whether participants were assigned to the AI or Wizard conditions, a human was in the loop in order to ensure that personally identifiable information was handled responsibly. In the Wizard condition, the wizard replied back to the participants as they would in any other synchronous text messaging exchange. In the AI condition, the wizard served as an intermediary — submitting participants' messages to GPT-4 and then copying its responses back into Highside. In the event that participants revealed sensitive data, the wizard was instructed to replace identifiable information with a generic placeholder; however, this precautionary measure was never required. There was no noticeable difference in response speed between the two conditions,

[5]https://highside.io/

as a similar amount of latency was introduced in both. In the AI condition, time was required to manually copy responses between the chat interface and OpenAI, and in the Wizard condition, time was required to manually type out each response.

3.7 Analysis

Our analyses were primarily qualitative in nature as we sought to elicit design recommendations rather than attempting to assert that one condition was strictly better than the other. However, we used some descriptive statistics and statistical tests to identify prominent differences across the conditions. We compared participants' post-study survey responses using Kruskal-Wallis tests since the ratings were not normally distributed. Meanwhile, we calculated the number of messages sent and words exchanged by all parties and compared these values across conditions using t-tests.

To initiate our qualitative analysis, we first conducted partial closed-ended coding to categorize the questions that were asked by the agents. This was a non-trivial procedure since both entities were allowed to rephrase questions, skip questions, and ask follow-up questions as they deemed fit. All conversation turns were coded individually by two researchers who reached an agreement score of 0.94 according to Cohen's κ. There were a total of 15 codes, with each one mapping to one of the 15 original questions provided in the script. Follow-up questions were given a code that reflected the original question that prompted it (e.g., 4.1 to indicate a follow-up to the fourth question), and skipped questions were noted. These codes are enumerated more comprehensively in Supplementary Table 2 of the Appendix. To better understand the contents of agents' messages, we categorized their utterances into the categories listed in Table 3. The same two researchers coded each utterance, achieving a Cohen's κ of 0.83.

Finally, we transcribed the interviews and processed them using thematic analysis [12]. We derived codes related to the questions that were asked by the chatbot, the language and wording that the chatbot used, and how the pre-consultation chatbot influenced participants' overall clinical experience. We also extracted participants' quotes from these transcripts as shown in Supplementary Table 3 of the Appendix .

3.8 Positionality

The study was conducted at a local walk-in clinic located in the city of Toronto, Canada. One of the authors is a practicing family physician with industry experience in building technology for healthcare. The authors who participated as wizards have international medical degrees with one of them currently in residency at a major hospital. The rest of the authors are human-computer interaction researchers who often work at the intersection of computer science and healthcare. With the exception of one author based in India, the rest of the authors and physicians who participated in our research are based in a single major metropolitan area in North America.

4 FINDINGS

In Section 4.1, we use participants' feedback to support the perceived value of a pre-consultation chatbot and the perceived quality of the agents in both study conditions. In Sections 4.2 and 4.3, we provide both quantitative and qualitative analyses of the conversation content, detailing strengths and weaknesses that were identified in the agents. We first examine how the agents combined, followed up, and skipped questions, as well as the extent to which participants responded to these questions. We then recount how participants felt about the agents' tone and clarity. Finally, Sections 4.4 and 4.5 explore how the pre-consultation chatbot interaction was situated in the clinical experience. We first comment on the range of preconceptions and expectations that participants had of chatbots prior to joining our study, which influenced their initial reactions to the experience. We then describe the features that participants would have liked to have seen as they concluded their conversation and awaited their face-to-face clinical consultations.

4.1 Receptiveness to Pre-Consultation Chatbots

4.1.1 Overall Experience. From our post-study survey, we found that participants in both conditions had an overall positive experience with what they believed to be a pre-consultation chatbot. As shown in Figure 3, a majority of the participants said that they would be willing to use the chatbot 'most of the time' or 'always' during their future clinic visits. Participants praised the chatbot's ability to comprehend and deliver English text regardless of their assigned study condition, showing that the chatbot exhibited conversational qualities rivaling those of humans.

When asked to explain the value they saw in the chatbot, most participants commented that they recognized the potential for the chatbot to help their doctors collect information before their appointments. Knowing that they would eventually be speaking with a physician, participants were reassured that conversing with the chatbot would be unlikely to negatively influence their clinical consultations. In fact, this perception influenced the way that some participants interacted with the agents:

> *I kept things brief and high-level. The chatbot isn't going to diagnose me; the doctor is. The chatbot needs to know enough about the topics we'll be discussing to prepare the doctor.* (P9, AI agent)

Participants also noted that conversing with the agents provided benefits to themselves. For example, they mentioned that the conversation helped prepare them for questions that could have been asked during their appointments:

> *I think it kind of made me think about how I was feeling before I went into my appointment so I have better answers for the doctor. Like, it was asking me how long my symptoms were, and I had to think about it. I wouldn't have had that answer if the doctor had asked me.* (P18, Wizard agent)

Participants appreciated that conversing with a chatbot gave them the chance to think and respond to questions at their own pace. This affordance was especially important for participants who typically felt anxious speaking face-to-face with others, noting that *"being able to type versus talk is much easier for [them] to express how [they were] feeling"*. Similarly, P28 also commented on how the conversation gave them time to respond to the chatbot's questions without feeling rushed. They stated that if they were speaking to a person, *"they're not going to push me away, but it affects me mentally when I know that someone is waiting and I have to kind of be quick"*.

Table 3: The coding that was used to categorize the utterances sent by agents during the chatbot conversations.

Categories	Utterance codes	Example
Questions	From Script (Q1-Q15)	"What is the reason for your visit today?" (Table 2)
	Follow-up	"Is the pain constant, or does it come and go?"
Empathy	Salutation	"Hello, I am a physician assistant bot." or "It was a pleasure chatting with you today!"
	Appreciation	"Thank you for your honesty"
	Compassion	"I am sorry"
	Acknowledgement	"I see" or "I understand you're experiencing pain in the left region of your jaw"
Explanation	Directing Conversation	"Lets start with your symptoms one at a time."
	Informing Context	"This information is important for your medical record and can help your physician provide the best care."
	Instructions	"Please make sure to have the names and dosages of your medications ready before your appointment."

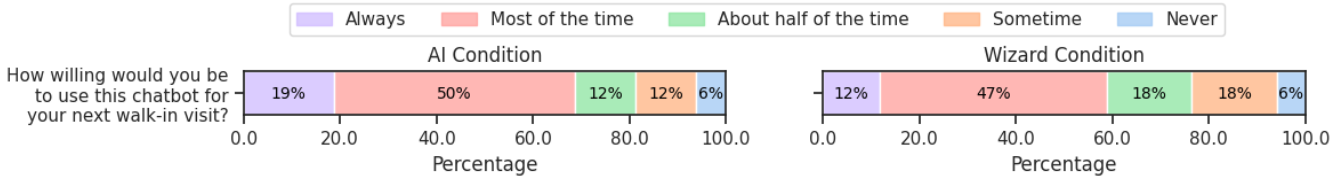


Figure 3: Participants' ratings regarding their willingness to use a pre-consultation chatbot in the future: (left) AI condition and (right) Wizard condition.

4.1.2 Engagement with the Agents. Figure 4 shows how participants rated the two conditions according to the conversation quality metrics we adopted from Abd-Alrazaq et al. [1]. We observed that participants who conversed with the AI agent gave comparable positive ratings relative to those who spoke to the Wizard agent. Kruskal-Wallis statistical tests showed no significant differences (p » 0.10) for any of these metrics between the conditions.

All respondents believed that the chatbot asked questions in a logical order and had a good grasp of the English language. More than 80% of the participants felt that the chatbot understood their responses and asked questions that were relevant to their health. Finally, more than 60% of the participants felt that the chatbot was engaging and had a good grasp of medical knowledge. We use these observations to conclude that the AI agent was reasonably successful in carrying out natural conversations with participants. Although the ratings for the Wizard agent were similar, we observed that they occasionally had typos that may have influenced how participants perceived their grasp of the English language. However, in either condition, the mistakes were not obvious or prevalent enough for participants to believe that it would not be suitable for real-world use.

4.2 Conversation Content: What Was Said

Although participants gave comparable ratings for the conversation experience across both conditions, we delved into the quantitative characteristics of their conversation messages as an objective proxy for engagement. We then examine the sequence of questions asked by the Wizard agent and the AI agent to identify any patterns that participants appreciated or disliked in either condition.

4.2.1 Conversations and Word Counts. Figure 5 illustrates the distributions of the number of messages exchanged as well as the number of words sent between the agents and the participants; the corresponding values can be found in Supplementary Table 1. Both the agents and the participants sent an average of nearly 3 more messages in the Wizard condition than they did in the AI condition. A pairwise t-test shows that the difference between the conditions was significant for both participants (t=-3.24, $p < .05$) and the agents (t=-3.78, $p < .05$). As we will show in Section 4.2.2, this is due in large part to the fact that the Wizard agent tended to ask more follow-up questions.

Regarding the total word counts, we observe that the AI agent used 50 more words on average compared to the Wizard agent, indicating that the AI agent was slightly more verbose than the Wizard agent. Despite this difference, participants in the Wizard condition typed an average of 10 more words compared to those in the AI condition. However, neither of these differences was statistically significant according to pairwise t-tests.

4.2.2 Pattern of Questioning. Table 4 summarizes how the original list of questions was modified by both agents. Despite being explicitly instructed to avoid double-barreled questions by the script, both agents occasionally combined two questions into a single message. The Wizard agent often combined Q2 and Q4 together in order to query both the symptoms that participants were experiencing and the duration of those symptoms, as shown below.

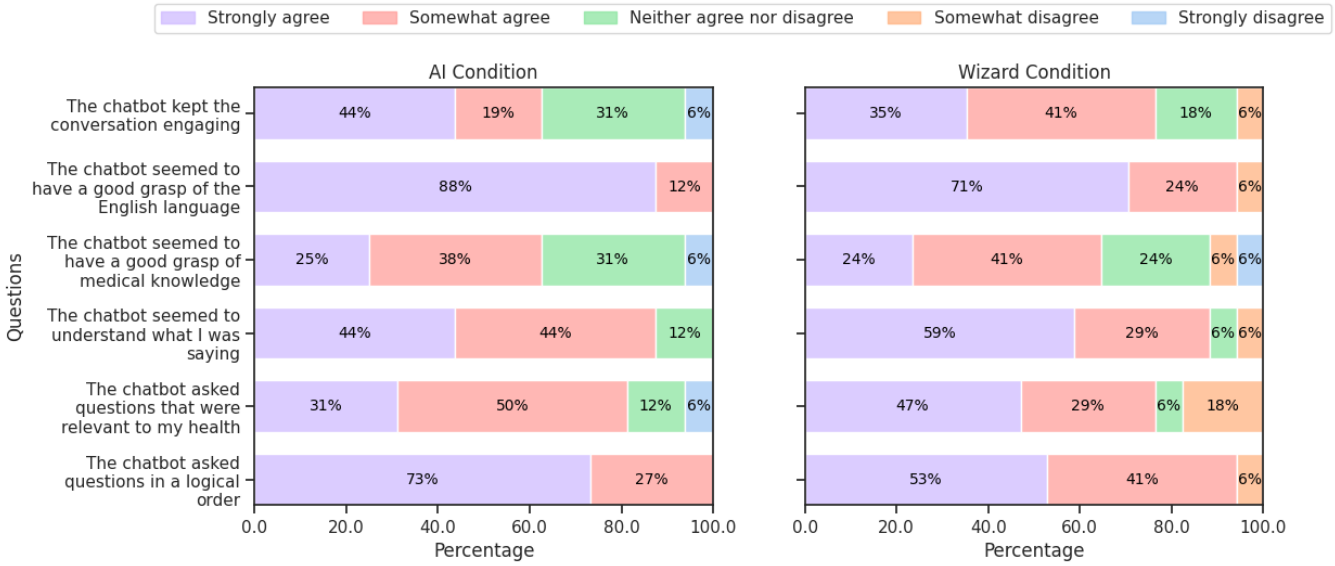
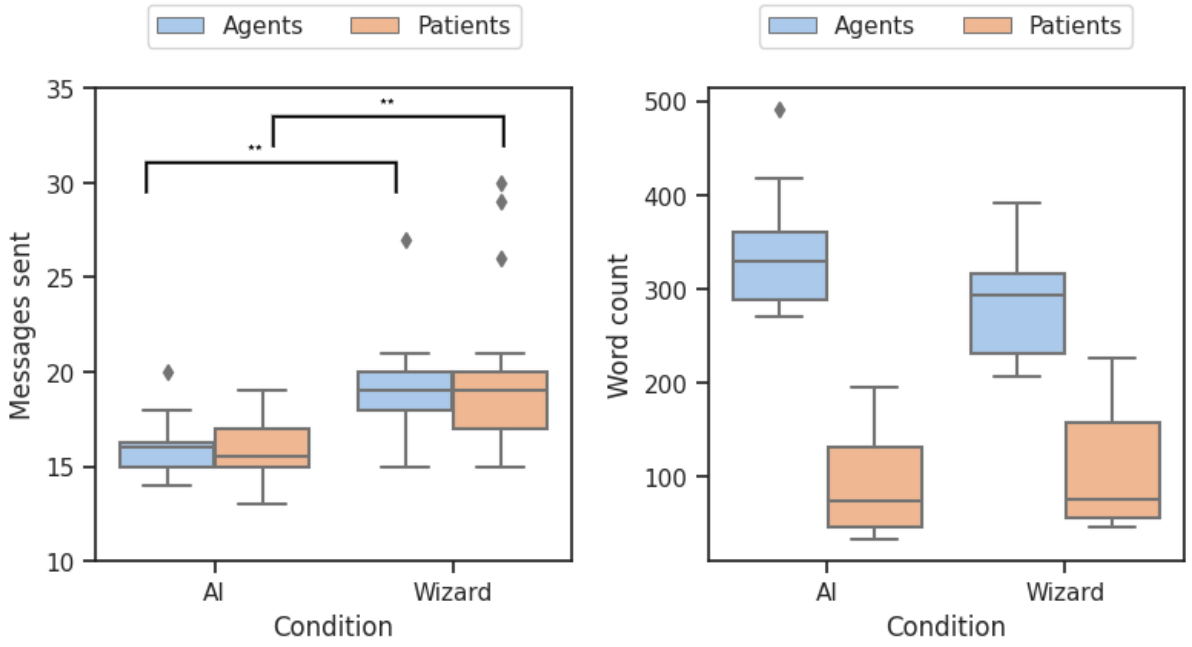


Figure 4: Participants' ratings regarding the perceived quality of the pre-consultation chatbot according to metrics by Abd-Alrazaq et al. [1]: (left) AI condition and (right) Wizard condition.



(a) The number of messages sent by agents and participants.

(b) The total word count for agents and participants.

Figure 5: Quantitative metrics indicating the length of conversations between agents and participants (*$p < .05$, **$p < .01$, *** $p < .001$).

> I am sorry to hear about that. Tell me about your symptoms and since how long have you been experiencing this? [Wizard agent in response to P14, conversation]

When asked about this habit, they noted that they would often ask these questions together during face-to-face consultations with patients.

Table 4: The question codes that were combined and their usage frequency across the AI and Wizard agents.

Combined Questions	Usage Frequency	
	AI Agent	Wizard Agent
1 + 2	0% (0/16)	5.9% (1/17)
2 + 4	6.3% (1/16)	2.9% (5/17)
3 + 5	6.3% (1/16)	0% (0/17)
4 + 5	6.3% (1/16)	0% (0/17)
6 + 7	1.3% (2/16)	0% (0/17)
9 + 12	0% (0/16)	5.9% (1/17)

An example of questions that were paired together by the AI agent was Q6 and Q7. These questions dealt with distinct topics: lingering details about symptoms that had not already been mentioned and the patient's medical history, respectively.

> `Understood. Do you have anything else you want to mention about your medical symptoms? Additionally, let's move on to discuss some details of your medical history: Do you have any chronic medical conditions?` [AI agent in response to P2, conversation]

On the one hand, grouping these questions together may have been a way of facilitating a transition between points of discussion. However, participants noted that groups of orthogonal questions like these made them feel like the topic of the discussion had changed, thereby discouraging them from addressing the first question in the pair.

> *When [the chatbot] asked the open-ended question, "Is there any more information that you would like to give?", I couldn't recall that information at that moment and it moved on quickly.* (P9, AI agent)

One of the more significant differences we observed between the AI agent and Wizard agent was the frequency and origins of follow-up questions. The Wizard agent asked an average of 3.0 additional questions per participant, while the AI agent only asked 0.4 additional questions per participant. The distribution of follow-up questions according to our original script is shown in Table 5, while specific examples asked by the AI and Wizard agents are shown in Figure 6. The most frequent questions added by the Wizard agent involved symptom presentation (Q2), medication usage (Q8), and the consumption of alcohol, drugs, and tobacco (Q13). The AI agent rarely asked additional questions on these topics, instead keeping follow-up questions limited to moments when participants gave brief and vague responses. Several participants in the AI condition were also caught off guard by the occasional lack of follow-up questioning that they would have expected from a health professional.

> *Maybe the follow-up questions that the chatbot can ask are not quite as specific as a nurse would . . . When it asked about my family history and I said my mom has several allergies, it didn't ask for what those were . . . and I was kind of expecting it to ask that . . . I thought it could have been relevant.* (P25, AI agent)

Because participants were not always asked the questions they anticipated, they sometimes felt that the thoroughness and relevance of the conversation content were lacking. For participants who were less familiar with pre-consultation in general, the shallow depth of information exchange caused them to question the point of the chatbot interaction entirely.

Since both agents merged and added questions during the conversations, they sporadically skipped questions that had already been addressed. The AI agent was more prone to skip questions, doing so an average of 1.0 times per participant compared to the 0.7 times per participant by the Wizard agent. The most frequently skipped question related to participants' symptoms (Q2) because that topic was often addressed in their response to the previous question on their reason for visiting the clinic (Q1).

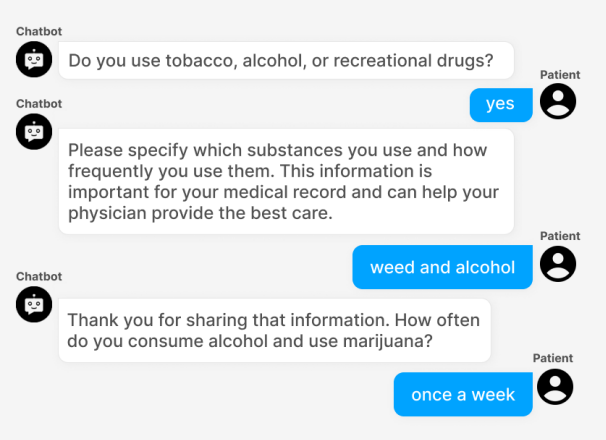
Supplementary Figure 2 and Supplementary Figure 3 show how the AI and Wizard agents adapted the order of the questions to suit the flow of the conversation. The Wizard agent made more adjustments to the question sequence, particularly when participants presented with multiple medical issues. Figure 7 illustrates an example of this behavior as the Wizard agent in that situation decided to cycle through questions one issue at a time to avoid significant context switching. This is in contrast to how the AI agent handled a similar situation illustrated in Figure 8. When the AI agent asked about the severity of all three symptoms at once, P30 responded with a single severity rating, thereby introducing ambiguity regarding the symptom being referenced.

4.3 Conversation Language: How It Was Said

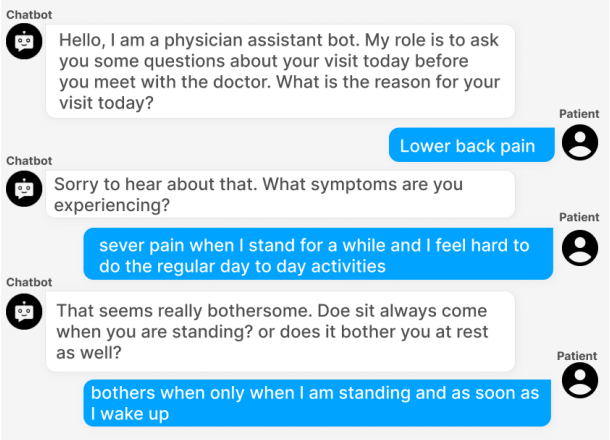
Among the 645 utterances by the AI agent, 275 (43%) were questions, 256 (39%) were expressions of empathy and 114 (18%) were explanations. Among the 630 utterances by the Wizard agent, 335 (53%) were questions, 214 (34%) were expressions of empathy and 81 (13%) were explanations. The breakdown of these utterances is further shown in Table 6. Below, we examine the notable differences between our two conditions and how they influenced participants' reactions to the corresponding agents.

4.3.1 Procedural Comments and Explanations. The Wizard agent often used language to direct the conversation and guide participants through their line of questioning, more so than the AI agent. These comments were particularly helpful for situations when participants reported having multiple medical issues, as shown earlier in Figure 7.

While the AI agent was not as proficient in this regard, it provided many short procedural utterances that indicated transitions between topics, including utterances like `"Moving onto the next question"` and `"Let's continue"`. In addition, the AI agent was more likely to explain to participants why certain questions were being asked. Such utterances involved informing participants about how their responses to questions would contribute to the clinical information-gathering process. For example, several participants mentioned that they did not share some components of their medical history because they did not perceive it to be relevant to their current visit. However, as depicted in Figure 9, the added explanations may have helped clarify some of the concerns they had when it came to sharing their personal medical information, leading them to disclose more.



(a) The AI agent following up responses by P17.



(b) The Wizard agent following up responses by P26.

Figure 6: Examples of follow-up questions asked by both the AI and Wizard agents. The conversation interface for this and future conversation examples have been altered for clarity. The original interface is shown in Supplementary Figure 1.

Table 5: The average rate of followed-up questions and skipped questions across the AI and Wizard agents.

Question #	Follow-up Rate		Skip Rate	
	AI Agent	Wizard Agent	AI Agent	Wizard Agent
1	0% (0/16)	5.9% (1/17)	0% (0/16)	0% (0/17)
2	0% (0/16)	100% (17/17)	31.3% (5/16)	17.6% (3/17)
3	0% (0/16)	11.8% (2/17)	6.3% (1/16)	0% (0/17)
4	0% (0/16)	5.9% (1/17)	18.8% (3/16)	17.6% (3/17)
5	0% (0/16)	17.6% (3/17)	6.3% (1/16)	5.9% (1/17)
6	0% (0/16)	0% (0/17)	12.5% (2/16)	5.9% (1/17)
7	0% (0/16)	5.9% (1/17)	6.3% (1/16)	0% (0/17)
8	6.3% (1/16)	70.6%(12/17)	0% (0/16)	0% (0/17)
9	12.5% (2/16)	5.9% (1/17)	0% (0/16)	5.9% (1/17)
10	0% (0/16)	11.8% (2/17)	0% (0/16)	0% (0/17)
11	12.5% (2/16)	23.5% (4/17)	0% (0/16)	0% (0/17)
12	0% (0/16)	17.6% (3/17)	6.3% (1/16)	5.9% (1/17)
13	18.8% (3/16)	23.5% (4/17)	0% (0/16)	0% (0/17)
14	6.3% (1/16)	5.9% (1/17)	0% (0/16)	5.9% (1/17)
15	0% (0/16)	0% (0/17)	6.3% (1/16)	0% (0/17)

Table 6: The categories and utterance codes from the AI and Wizard agent transcripts.

Categories	Utterence codes	AI Agent (Average per Conversation)	Wizard Agent (Average per Conversation)
Questions	From Script (Q1-Q15)	15.9 (254/16)	16.3 (277/17)
	Follow-up	1.3 (21/16)	3.4 (58/17)
Empathy	Salutation	1.4 (23/16)	1.2 (20/17)
	Appreciation	8.3 (132/16)	6.6 (112/17)
	Compassion	1.4 (22/16)	2.7 (45/17)
	Acknowledgement	4.9 (79/16)	2.2 (37/17)
Explanation	Directing Conversation	0.6 (9/16)	0.7 (11/17)
	Informing Context	5.8 (93/16)	3.6 (62/17)
	Instructions	0.8 (12/16)	0.5 (8/17)

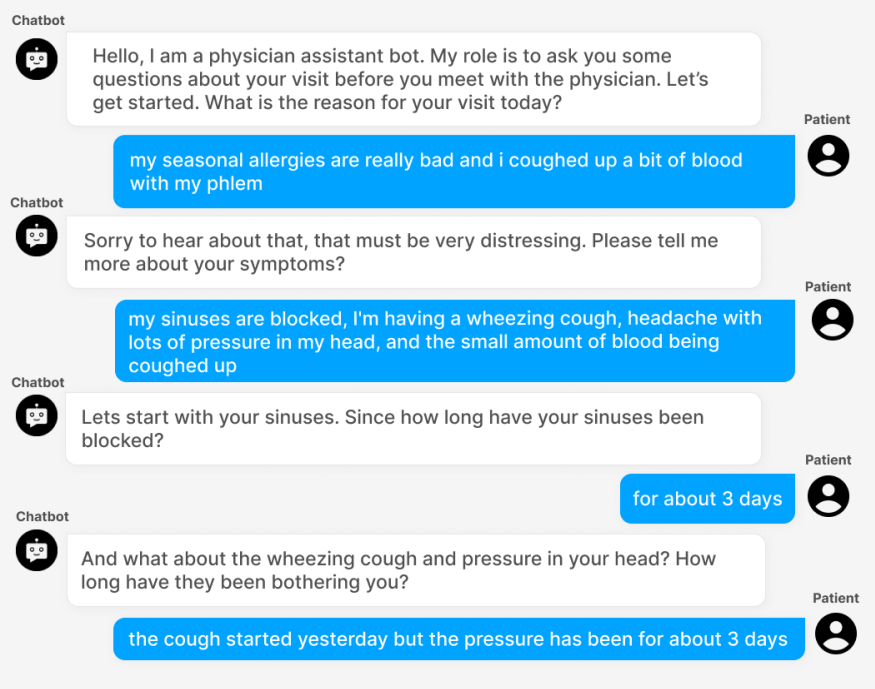


Figure 7: The Wizard agent adapting their line of questioning to multiple medical concerns mentioned by P18.

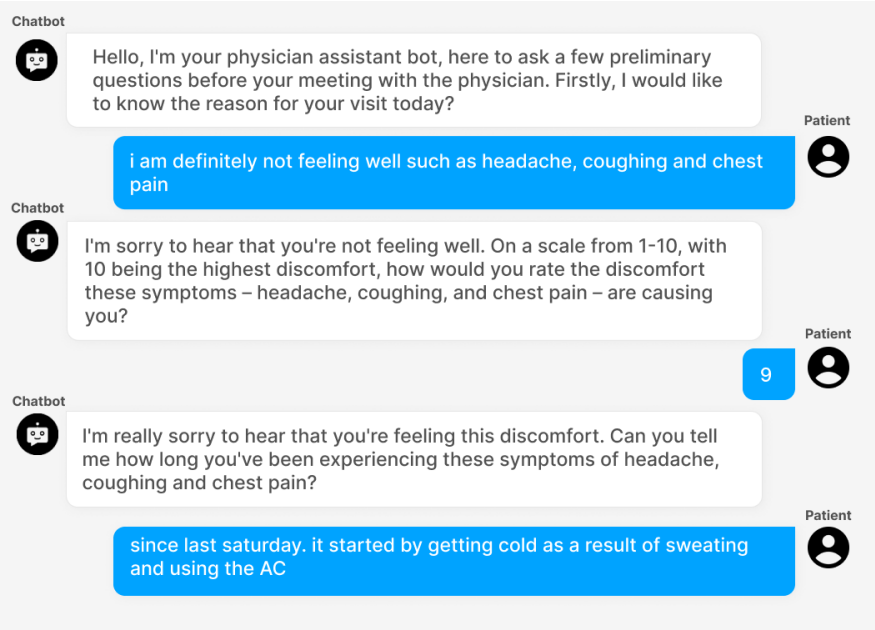


Figure 8: The AI agent asking about all of P30's symptoms at the same time, resulting in an ambiguity about which symptom is being discussed in later messages.

4.3.2 Excessive Appreciation, Compassion, and Acknowledgements. Politeness and professionalism are important traits for physicians to convey while talking with patients, and some participants appreciated these empathetic traits by the agents. However, some participants also found the appreciative tones to be counterintuitive coming from what they believed to be an AI agent because they expected algorithms to have fewer human-like mannerisms.

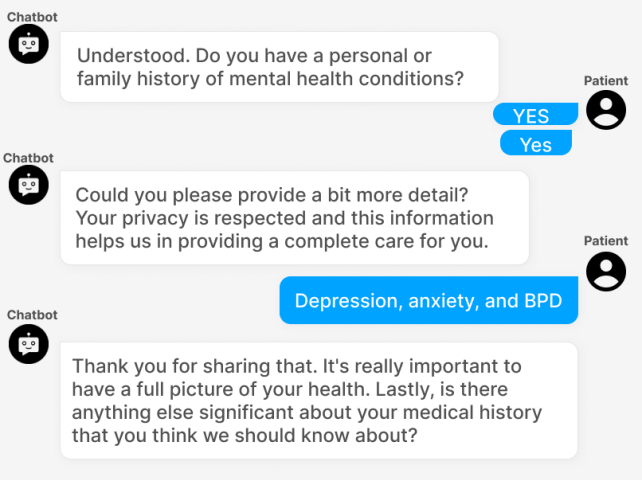


Figure 9: The AI agent explaining to P27 the importance of providing detailed information during pre-consultation.

> *I did find that to be almost strange. Like I know that I'm talking to like a robot, and the robot's talking like a person … To me, I found it strange because I'm like, "You can't feel sorry; you're literally a robot".* (P25, AI agent)

This comment was raised more frequently by participants in the AI condition, which could be related to the AI agent's propensity to include added remarks of appreciation and acknowledgment like "Thank you for sharing" and "Understood". In fact, some participants in this condition felt that the agent's empathy became monotonous, tiresome, and borderline disingenuous.

> *When you know it's a chatbot and it's an automatic answer that you're getting, like you know, "I'm sorry to hear that", it just seems very fake. Like unnecessary … You're really not sorry, you really don't care … Like "Thanks for my honesty"? What is it? Am I hiding that I'm an alcoholic? That part I thought was weird.* (P17, AI agent)

This comment emphasizes that while an empathetic tone can make a chatbot's messages feel more human-like, excessive displays of empathy can actually disrupt users' experiences and may sometimes even come across as insincere.

Another component of physician professionalism is the notion of active listening, which entails conveying to patients that their concerns are being heard. Physicians use verbal cues (e.g., affirmations, paraphrasing) and non-verbal cues (e.g., nodding, eye contact) during their conversations with patients to acknowledge what is being said. The AI agent in our study actually included more acknowledgments while conversing with participants compared to the Wizard agent. Again, participants occasionally felt that the AI agent overused these statements, distracting them from the task at hand and making the conversation unnecessarily verbose.

4.4 User Experience: Expectations Prior to the Conversation

Although participants were given some background about the study's purpose prior to conversing with the agents, we found that their prior experiences with pre-consultation and chatbots influenced how they approached the task.

4.4.1 Expectations of Pre-Consultation. We found that participants' presumptions about the pre-consultation process impacted how they viewed the chatbot conversation. Participants who had experienced some form of pre-consultation recognized that the purpose of the conversation was to help their physician get a better overview of their medical concerns. These participants opened up to the chatbot freely without feeling the need to hold back any information. They also rated the interaction highly and enjoyed this experience more than their prior interactions with intake nurses. They felt more comfortable providing information to the chatbot and trusted that the chatbot would be able to accurately relay their responses to their physician.

> *The chatbot seems easier than when you talk to a pre-screening person. They're usually not as thorough as a chatbot. Like, they miss questions. They also don't record your answers as well.* (P20, Wizard agent)

Participants who had no experience with pre-consultation often took longer to understand the purpose of some of the agents' questions, despite being told that a summary of the conversation would be given to their physician. This was particularly apparent when the agents asked questions about participants' medical history. Many participants did not understand the relevance of their family's medical history to their current visit to the clinic, especially when they considered their medical concerns to be straightforward.

> *I came in for a simple matter. I had to sit and fill out information about past surgeries, health … you know, yeah, health history, and all of that … And I understand that's important … But it wasn't important for why I was coming in.* (P12, AI agent)

Participants in these situations believed that fewer questions and less investigation from the chatbot were needed. They preferred for the chatbot to recognize the simplicity of their chief medical complaint and to stop the conversation early because the additional questions were not important for their physician. On the other hand, some participants who were previously unfamiliar with pre-consultation began to recognize the chatbot's goals as they progressed through the conversation. They initially found the chatbot's questions to be overly basic but later deduced that the information would serve as a good icebreaker for their physicians to ask for more information, although it would come at the slight cost of revisiting topics that were touched upon by the chatbot.

4.4.2 Expectations of Chatbots. We also discovered that participants' presumptions of chatbots significantly influenced their initial expectations for the conversation. Many participants did not expect the conversation to be similar to one they would have with a human. They expected chatbots to sound "robotic" and did not consider that to be a flaw. After conversing with the agent, P28 expressed the following opinion:

> *At the end of the day, I know it's an AI. It's a very different expectation when it comes to AI. When I see a human, I don't want it to be robotic. But when I see an AI, I don't care if it's human-like . . . At the end of the day, I want precise answers. That's the expectation there . . . So if I am talking to a chatbot, even if it's not asking me my name or something, that's fine . . . I mean it would be a very unreal expectation of me to have the AI talk to me as a human.* (P28, Wizard agent)

This view is particularly noteworthy because P28 was enrolled in the Wizard condition. Although the Wizard agent used less repetitive language and was more adept in guiding participants through non-linear conversational flows, P28 still believed that they were talking with an AI agent and prioritized its medical accuracy over its conversational skills.

Some participants also expected that chatbots would be less sophisticated in their ability to comprehend participants' messages. Regardless of whether they were conversing with an AI agent or a Wizard agent, participants purposefully altered their responses from how they would normally say things to make it easier for the agents to understand.

> *I was trying to keep it short and not too detailed. Because I know it might be a little tougher for it, but I don't know if that's accurate or not. Like, does it pick up all those little details? So yeah, maybe I was holding back.* (P14, Wizard agent)

The assumption that chatbots can only process simple sentences led some participants to shorten their messages and others to simplify their terminology. Participants may not have made these adjustments had they known that LLMs have made chatbots increasingly capable of comprehending complex messages, which would have streamlined the pre-consultation experience since they would not have felt compelled to reword their initial thoughts.

A limitation that participants were less willing to tolerate from chatbots was faulty lines of questioning. They noted that while it would be forgivable for a human with a finite memory capacity to forget previously discussed topics, chatbots should be able to keep an accurate and persistent memory of what had already been covered.

> *When you're doing it on the chatbot, it is better documented than human interactions . . . When you're talking to a person, there are usually some mistakes that they make when transcribing the information for sure.* (P20, Wizard agent)

Therefore, they believed that chatbots would also be able to use that information to skip irrelevant or redundant questions. Although both the Wizard and AI agents occasionally skipped questions, participants' presumptions about pre-consultation may have influenced their opinions on which questions were unnecessary.

4.4.3 Expectations and Concerns about Privacy. Participants were given assurances about data privacy before consenting to enroll in our study, and many of them were satisfied with these guarantees because of their existing trust in the healthcare system. This sentiment was particularly prominent among younger participants who were generally more accustomed to sharing information on the Internet. People like P15 recognized that their personal and medical information were already being stored digitally at various clinics, so having a record of their pre-consultation conversation was simply another entry in those systems.

In fact, some participants even considered conversing with an AI agent to be a way of enhancing their privacy since it allowed them to disclose sensitive information without being directly observed by a human who may judge their responses.

> *I'm not bothered by putting things out on the Internet. Or maybe it's just my age showing, but I would prefer this over talking openly in front of like a roomful of people.* (P27, AI agent)

Nevertheless, there were some participants who still expressed concerns about data privacy. They had questions about what data was being stored, how the data may be accessed, and who had access to the conversation record and summaries. Because their conversations were being typed out, these individuals recognized the potential permanence of their responses.

> *It feels different to have a conversation with a person where nothing's being written down necessarily versus typing something in . . . And that kind of makes you feel like it's like a permanent record, even if it isn't . . . something that like exists in the world though.* (P25, AI agent)

These opinions were often connected to participants' prior experiences with pre-consultation. Those who had never experienced any form of pre-consultation were more skeptical about whether their conversations with the chatbot would be kept private. On the other hand, those who had been seen by an intake nurse before had a better idea of the type and depth of information that was being recorded since they could observe the nurse typing on their computer or writing on their notepad.

4.5 User Experience: Guiding Patients Back to the Clinic

Although evaluating the influence of the different study conditions on the actual physician consultation was beyond the scope of this work, we highlight multiple themes that emerged regarding how participants situated the idea of a pre-consultation chatbot in their broader clinic visit.

4.5.1 Post-Conversation Summary. To ensure that participants were incentivized to converse with the agents, the Wizard agent summarized conversations from both conditions and handed those summaries to the physicians prior to their consultations. Many participants commented that they also would have wanted to see those summaries, with some expressing the concern that the agent might have misinterpreted the intent or subtext of their messages. Conversely, some participants were worried that they were the ones who misunderstood the topics being discussed, so they wanted to verify that their words were not being taken out of context.

> *I would like to see the summary because maybe I . . . maybe I misunderstood the question and provided the wrong information. And at the end of the summary, I would say like, "Oh, I didn't say that or I didn't mean to say that".* (P9, AI agent)

Regardless of whether participants believed they or the agents were at fault, they suggested that seeing the final outcome of the interaction would have served as a safety measure to mitigate downstream errors.

4.5.2 Instructions for Next Steps. After completing the study, several participants were confused about how to proceed with the rest of their clinic visit. As before, these expectations were largely informed by participants' familiarity with pre-consultation at other clinics. Some assumed that the agent would give them a diagnosis or medical advice, while others assumed that they would be seen by their assigned physician immediately after. Neither was true, as the script specifically prohibited the agents from giving advice to participants and participants were sent back to the waiting area to be called by the reception desk for their appointment.

In some cases, the AI agent actually alleviated this confusion. Without any prompting in the script, the AI agent occasionally gave participants instructions about how they should proceed.

> `Thank you for providing all the necessary information. The physician will review this information before your consultation. If they need any further details or clarification, they will ask during your appointment. Please wait for further instructions to meet with the physician. Have a great day!` [AI agent in response to P9, conversation]

Participants appreciated these kinds of messages because they explained how the pre-consultation was situated in their clinic visit. By telling patients that the physician would be reading a summary of their interaction, the AI agent reminded them that any and all health recommendations would be given by humans. Meanwhile, the fact that the AI agent told patients that they could clarify or revise their responses upon meeting with their physician provided them peace of mind in case they recalled new information later.

5 DISCUSSION

For our discussion, we first summarize some of the benefits participants perceived from using the pre-consultation chatbot. We then reflect on how our findings connect back to the initial prompt we gave the agents in order to suggest future design considerations in this space. To conclude, we propose ideas on how to better situate pre-consultation chatbots into patients' clinical experiences and discuss insights that may generalize to other information-gathering chatbots.

5.1 Perceived Benefits of Using Pre-Consultation Chatbots

Our participants saw value in having a chatbot collect preliminary information before the visit to help inform their doctor. They also appreciated the interaction because it was engaging and allowed them to reflect on their concerns at their own pace, which helped them feel more prepared for their appointment. These findings align with existing literature on the benefits of pre-consultation and the use of pre-consultation questionnaires [2, 33, 56, 65, 67, 77, 84].

The benefits of engaging in pre-consultation were noticed by participants in both study conditions. Whether they were interacting with the AI or Wizard agents, participants felt that the agent had moderate success at adapting its line of questioning using logic and medical knowledge. In fact, participants who conversed with the AI agent particularly appreciated the agent's proactive approach in explaining the pre-consultation process. While these findings show that general-purpose LLMs like GPT-4 already have many of the tools necessary for pre-consultation, the discussion that follows provides considerations for future work.

5.2 Improving Pre-Consultation Chatbot Prompts

As we developed our chatbot's prompt, we found that our AI agent adhered to our question sequence while adeptly skipping questions that had already been answered. Patients also perceived the conversation to be natural and intuitive. We added explicit instructions to the prompt so that the chatbot would align with some of our design objectives, namely language to discourage double-barreled questions and diagnostic recommendations. Although our chatbot was successful at avoiding diagnostic recommendations, double-barreled questions were still grouped together in several conversations. It is difficult to pinpoint a single reason why some of our instructions were more successful than others since an LLM's behavior is dictated by its prompting, its settings, and its underlying training data. As LLMs evolve, some features that currently have to be addressed with prompt engineering may become ingrained in future models [78, 82, 83]. Nevertheless, we use our findings to highlight challenges and provide prompt recommendations for future pre-consultation chatbots in Table 7. We elaborate on these goals and recommendations below.

5.2.1 Conversation Content. In our qualitative analysis of the conversation content, we found that the quantity and relevance of follow-up questions asked by the agent significantly influenced how participants perceived its thoroughness. The Wizard agent generally asked more follow-up questions than the AI agent, particularly questions on the symptoms and medications mentioned by participants. This could explain why some participants found the conversation with the AI agent to be less relevant for their medical visit, despite the fact that both conditions were primed with the same set of questions. The additional follow-up questions also help explain why there were significantly more messages exchanged between the Wizard agent and participants. Although this is a relatively simplistic metric for conversational depth, longer conversations give participants more opportunities to disclose information, which may improve the effectiveness of the pre-consultation.

These findings suggest that future pre-consultation chatbot prompts should place a greater emphasis on asking follow-up questions that uncover more depth into the reasons for a patient's visit. This could be achieved by assigning greater importance to the earlier questions in our chatbot's prompt (Q1–Q6). We also noticed that the Wizard agent was particularly adept at navigating situations when participants had multiple symptoms. When the AI agent was faced with these situations, it often asked participants each question in the prompt once to cover all of the symptoms. In contrast, the Wizard agent was able to go through the questions multiple times, ensuring a more comprehensive exploration of the participant's symptoms and a better understanding of their condition. Therefore,

Table 7: Prompt design goals and recommendations for pre-consultation chatbots.

Category	Prompt Goal	Recommendations
Content	Improve thoroughness of questions	Ask more follow-up questions, especially ones that relate to the symptoms the patient has mentioned.
	Improve structure of questions	Focus on one issue at a time unless it seems like the symptoms may be related.
Language	Convey more sincerity	Encourage appreciative language, but not at every conversation turn. Consider appreciation when it seems like the patient is sharing information they may otherwise not be comfortable providing.
	Improve clarity	Encourage acknowledgements, but not at every conversation turn. Provide a summary only when the chatbot needs to confirm information that they may have misunderstood.
Situating in the Clinical Experience	Set expectations for the chatbot	Either before the conversation or shortly after the initial greeting, describe the chatbot's conversation capabilities.
	Set expectations for the pre-consultation	Either before the conversation or shortly after the initial greeting, describe how pre-consultation will help inform the patient's consultation with the doctor.

a chatbot prompt could include specific instructions to encourage such logical looping if required.

Another way to foster targeted follow-up questions is by incorporating an LLM with more medical knowledge than what is currently available in a general-purpose model like GPT-4. Although intake nurses have significant medical knowledge, their ability to carry out personable and empathetic conversations is often just as important for delivering patient-centered care [57, 72, 75, 80]. We chose to use GPT-4 in our study to balance these traits. Using medically specialized LLMs such as Med-Palm2 [64] for pre-consultation may lead to beneficial follow-up questions, but future work would be needed to evaluate whether this would require sacrifices in other important dimensions of conversationality.

5.2.2 Conversational Language and Tone. Participants noticed that both the AI and Wizard agents used language to convey empathy, which prior literature has emphasized is important to patients when they share their concerns [15]. In fact, the AI agent used appreciative and acknowledging language far more frequently. Although this finding may go against the preconceived notion that many people have about chatbots being robotic and inexpressive [16], it actually aligns with recent work by Ayers et al. [5] who found a similar difference in how chatbots and physicians responded to patient questions on medical forums. However, we found that the AI agent in our study occasionally used such expressive language to the point of seeming insincere or even offensive.

In recognition of this potential pitfall, our findings suggest that chatbots should be prompted to be more tactful in how they convey empathy. Medical professionals undergo many years of medical training to communicate in a manner that balances compassion with precision and clarity, so chatbot prompts require substantial instruction to emphasize best practices in patient-physician communication. We found that it was helpful to define the chatbot's role as a "patient-intake bot" and to emphasize desirable behaviors like conversing in a "medically professional manner". At the same time, specific instructions might also be needed to discourage undesirable behaviors. Considering that one of the most egregious cases of perceived insincerity was a case when the AI agent repetitively added the phrase "thank you" after each participant response, an easy way of improving this aspect of the chatbot's behavior would be by including an instruction to avoid overly repetitive language.

5.3 Scaffolding the Pre-Consultation Experience

5.3.1 Before the Pre-Consultation. Using chatbots for pre-consultation was a new experience for all of our participants, but even the general concept of pre-consultation was unfamiliar to some. Those who had never gone through a pre-consultation before were confused about how some of the medical history questions related to their symptoms, and it was not immediately evident to them how this information would affect the rest of their appointment. Even though it was not explicitly instructed in our prompt, the AI agent occasionally provided background information about the goals of pre-consultation in order to alleviate these worries. The AI agent also sporadically provided participants with instructions on what to do after the conversation, which helped them position the pre-consultation process within the broader context of their clinic visit. Nevertheless, future prompts should ensure that this background is provided in every conversation.

Interacting with a chatbot was also a relatively novel experience. Most had anticipated a chatbot with a robotic demeanor and limited comprehension abilities but were instead met with a conversation that closely mimicked human interaction. Although people's expectations of chatbots will undoubtedly change as LLMs become more pervasive in healthcare and beyond, it is still important for future applications in these domains to prime users about the expected conversational dynamics so that they are not caught off guard or left dissatisfied with the experience. In other words, users should be told in advance if the chatbot was engineered to carry out human-like discourse. Users should also be told about the chatbot's comprehension limits so they do not withhold information that they assume the chatbot will not be able to understand. Having such a summary of a chatbot's capabilities and expectations can foster more effective and improved chatbot interactions in pre-consultation and any

other task that involves information exchange between multiple stakeholder groups.

5.3.2 After the Pre-Consultation. Several participants mentioned that they would have liked to have seen the conversation summary that was sent to their physician. We did not consider doing this when designing our study because it is not a common practice in pre-consultation. Nevertheless, medical transparency is a vital aspect of patient-centered care, so it is reasonable to assume that intake nurses will occasionally review everything that has been discussed to confirm that all of the patient's concerns have been understood [73].

Following this feedback, we suggest that future pre-consultation chatbots allow users to review the conversation and the subsequent automatically generated summary. This would afford patients the opportunity to not only confirm the accuracy of the summary but also amend incorrect statements, redact overly sensitive information, or augment the discussion with details that might have been initially overlooked. It is important to recognize, however, that clinicians often rely on notes with medical terminology, jargon, and abbreviations that may not be interpretable to the average person. Therefore, patients may need to be shown a different summary from the one given to their physicians, but it is important that patients are explained why these differences exist. Future research is needed to investigate how to optimally summarize the chatbot conversation for both patients and physicians to account for their needs and capabilities.

5.4 Sociotechnical Implications of Pre-Consultation Chatbots

We focused our research efforts on the idea of a pre-consultation chatbot because we felt that this application of LLMs would circumvent many ongoing concerns about diagnostic chatbots, namely the consequences of improper recommendations being given to users [9, 21, 30]. However, pre-consultation chatbots are not a foolproof clinical application of LLMs because they come with their own set of sociotechnical considerations.

Existing healthcare systems in industrialized countries rely on electronic health record (EHR) systems to document patient histories and to facilitate communication among healthcare providers [22, 51]. Our chatbot operated separately from the EHR system at our study site to avoid impacting patient care, so future efforts are needed to investigate the potential opportunities and pitfalls that emerge from this integration [68]. For example, EHR systems require healthcare providers to dedicate significant time documenting patient information [63]. A pre-consultation chatbot could alleviate this burden since it produces a written record of patients' history and medical concerns, but giving physicians the full chatbot transcript would require them to spend significant time reading and summarizing patient responses. On the other hand, automatically generating transcript summaries for physicians to read would require having strong guarantees that the summaries faithfully and comprehensively include all of the relevant information from the pre-consultation transcript. Issues around medical liability, data privacy, and confidential disclosure also become relevant once a pre-consultation chatbot is integrated into an EHR system.

Another consideration for pre-consultation chatbots is the characteristics of the patient population being served. Patients in our study tended to be young because walk-in clinic patients often lack a regular family doctor, as is the case with many young people [55, 61]. The fact that younger people are often more accepting of technology [11], combined with the novelty still attributed to LLMs across domains [23, 79], may have inflated our participants' receptivity to a pre-consultation chatbot. Older individuals often prefer face-to-face interactions over conversing with a chatbot [74], due in part to the inconvenience of typing for some individuals. Future systems could examine other interaction modalities like voice-to-text or voice-to-voice interactions to support these groups. Another demographic factor that should be considered is patients' language proficiency. Although it is likely safe to assume that patients would not be referred to a pre-consultation chatbot if they are not reasonably fluent in the language used to train the LLM, chatbots could still use medical jargon unfamiliar even to native speakers. In this regard, providing features that adjust chatbot prompting for specific audiences may be worth future investigation. These recommendations extend beyond healthcare to numerous other sectors, as tailoring options based on user demographics can enhance the preparation process and facilitate a more personalized pre-consultation experience.

6 CONCLUSION

By deploying an AI agent in a clinical setting and contrasting it with a human agent, we were able to examine how patients reacted to different instantiations of the same prompt. We found that our chatbot implementation had many shortcomings relative to a human agent, such as a lack of follow-up questions and excessive empathetic language. However, our chatbot also had its own strengths, namely its initiative in explaining the motivation behind specific questions and the pre-consultation process more broadly. Regardless of the study condition that highlighted these design considerations, our findings led to a series of design goals that we believe will improve the user experience and data collection efficacy of future pre-consultation chatbots.

REFERENCES

[1] 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.

[2] Akke Albada, Sandra van Dulmen, Margreet GEM Ausems, and Jozien M Bensing. 2012. A pre-visit website with question prompt sheet for counselees facilitates communication in the first consultation for breast cancer genetic counseling: findings from a randomized controlled trial. *Genetics in Medicine* 14, 5 (2012), 535–542.

[3] Anoop Anugraha, Rakesh Dalal, Marjan Raad, Neelam Patel, and Hari Sugathan. 2021. Preconsultation Questionnaires for Patients Attending Elective Foot and Ankle Clinics: Is This the Way Forward in Outpatient Clinics? *Foot & Ankle Specialist* (2021), 1938640020986644.

[4] Joshua Au Yeung, Zeljko Kraljevic, Akish Luintel, 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.

[5] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* (2023).

[6] Julian Barratt and Nicola Thomas. 2019. Nurse practitioner consultations in primary health care: a case study-based survey of patients' pre-consultation expectations, and post-consultation satisfaction and enablement. *Primary health care research & development* 20 (2019), e36.

[7] Howard S Barrows et al. 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *Academic Medicine – Philadelphia* 68 (1993), 443–443.

[8] Bruce Bartley and Peter Cameron. 2000. QUEST: Questionnaire relating to patients' Understanding and Expectations of their Symptoms and Treatment. *Emergency Medicine* 12, 2 (2000), 123–127.

[9] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. 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, 1–12. https://doi.org/10.1145/3313831.3376718

[10] 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.

[11] Petter Bae Brandtzaeg and Asbjørn Følstad. 2017. Why People Use Chatbots. In *Internet Science*, Ioannis Kompatsiaris, Jonathan Cave, Anna Satsiou, Georg Carle, Antonella Passani, Efstratios Kontopoulos, Sotiris Diplaris, and Donald McMillan (Eds.). Springer International Publishing, Cham, 377–392.

[12] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[13] Lang Cao. 2023. DiagGPT: An LLM-based Chatbot with Automatic Topic Management for Task-Oriented Dialogue. arXiv:2308.08043 [cs.CL]

[14] Rosaline De Koning, Abdullah Egiz, Jay Kotecha, Ana Catinca Ciuculete, Setthasorn Zhi Yang Ooi, Nourou Dine Adeniran Bankole, Joshua Erhabor, George Higginbotham, Mehdi Khan, David Ulrich Dalle, et al. 2021. Survey fatigue during the COVID-19 pandemic: an analysis of neurosurgery survey response rates. *Frontiers in Surgery* 8 (2021), 690680.

[15] Frans Derksen, Jozien Bensing, and Antoine Lagro-Janssen. 2013. Effectiveness of empathy in general practice: a systematic review. *British journal of general practice* 63, 606 (2013), e76–e84.

[16] Laury Donkelaar. 2018. *How human should a chatbot be?: The influence of avatar appearance and anthropomorphic characteristics in the conversational tone regarding chatbots in customer service field.* Master's thesis. University of Twente.

[17] Vari M Drennan and Fiona Ross. 2019. Global nurse shortages: the facts, the impact and action for change. *British medical bulletin* 130, 1 (2019), 25–37.

[18] Carlos El-Haddad, Iman Hegazi, and Wendy Hu. 2020. Understanding patient expectations of health care: a qualitative study. *Journal of patient experience* 7, 6 (2020), 1724–1731.

[19] Reem El Sherif, Pierre Pluye, Christine Thoër, and Charo Rodriguez. 2018. Reducing negative outcomes of online consumer health information: qualitative interpretive study with clinicians, librarians, and consumers. *Journal of medical Internet research* 20, 5 (2018), e169.

[20] 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.

[21] Xiangmin Fan, Daren Chao, Zhan Zhang, Dakuo Wang, Xiaohua Li, and Feng Tian. 2021. Utilization of self-diagnosis health chatbots in real-world settings: case study. *Journal of medical Internet research* 23, 1 (2021), e19928.

[22] Eric W Ford, Nir Menachemi, and M Thad Phillips. 2006. Predicting the adoption of electronic health records by physicians: when will health care be paperless? *Journal of the American Medical Informatics Association* 13, 1 (2006), 106–112.

[23] Luke K Fryer, Mary Ainley, Andrew Thompson, Aaron Gibson, and Zelinda Sherlock. 2017. Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners. *Computers in Human Behavior* 75 (2017), 461–468.

[24] Marsa Gholamzadeh, Hamidreza Abtahi, and Marjan Ghazisaeeidi. 2021. Applied techniques for putting pre-visit planning in clinical practice to empower patient-centered care in the pandemic era: a systematic review and framework suggestion. *BMC Health Services Research* 21, 1 (2021), 1–23.

[25] Suzanne Graham and John Brookey. 2008. Do patients understand? *The permanente journal* 12, 3 (2008), 67.

[26] Trisha Greenhalgh, Rosamund Snow, Sara Ryan, Sian Rees, and Helen Salisbury. 2015. Six 'biases' against patients and carers in evidence-based medicine. *BMC medicine* 13, 1 (2015), 1–11.

[27] Randall W Grout, Erika R Cheng, Matthew C Aalsma, and Stephen M Downs. 2019. Let them speak for themselves: improving adolescent self-report rate on pre-visit screening. *Academic pediatrics* 19, 5 (2019), 581–588.

[28] Pamela Herd and Donald Moynihan. 2021. Health care administrative burdens: Centering patient experiences. *Health Services Research* 56, 5 (2021), 751.

[29] Laura M Holdsworth, Chance Park, Steven M Asch, and Steven Lin. 2021. Technology-Enabled and artificial intelligence support for pre-visit planning in ambulatory care: findings from an environmental scan. *The Annals of Family Medicine* 19, 5 (2021), 419–426.

[30] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. 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 659, 14 pages. https://doi.org/10.1145/3411764.3445385

[31] Maria Jimènez Torres, Klara Beitl, Julia Hummel Jimènez, Hanna Mayer, Sonja Zehetmayer, Wolfgang Umek, and Nikolaus Veit-Rubin. 2021. Benefit of a nurse-led telephone-based intervention prior to the first urogynecology outpatient visit: a randomized-controlled trial. *International Urogynecology Journal* 32 (2021), 1489–1495.

[32] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 18, 16 pages. https://doi.org/10.1145/3544548.3581503

[33] Karin Kee, Reinie G Gerrits, Nelleke de Meij, Lieke HHM Boonen, and Paul Willems. 2023. 'What you suggest is not what I expected': How pre-consultation expectations affect shared decision-making in patients with low back pain. *Patient education and counseling* 106 (2023), 85–91.

[34] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. 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, 1–12. https://doi.org/10.1145/3290605.3300316

[35] Zalika Klemenc-Ketis, Andrej Kravos, Tonka Poplas-Susič, Igor Švab, and Janko Kersnik. 2014. New tool for patient evaluation of nurse practitioner in primary care settings. *Journal of clinical nursing* 23, 9-10 (2014), 1323–1331.

[36] Rafal Kocielnik, Elena Agapie, Alexander Argyle, Dennis T Hsieh, Kabir Yadav, Breena Taira, and Gary Hsieh. 2019. HarborBot: a chatbot for social needs screening. In *AMIA Annual Symposium Proceedings*, Vol. 2019. American Medical Informatics Association, AMIA, 552. https://pubmed.ncbi.nlm.nih.gov/32308849/

[37] Harsh Kumar, Kunzhi Yu, Andrew Chung, Jiakai Shi, and Joseph Jay Williams. 2023. Exploring The Potential of Chatbots to Provide Mental Well-being Support for Computer Science Students. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2* (Toronto ON, Canada) *(SIGCSE 2023)*. Association for Computing Machinery, New York, NY, USA, 1339. https://doi.org/10.1145/3545947.3576285

[38] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health* 2, 2 (2023), e0000198.

[39] Sharon Latimer, Wendy Chaboyer, and Brigid Gillespie. 2014. Patient participation in pressure injury prevention: giving patient's a voice. *Scandinavian Journal of Caring Sciences* 28, 4 (2014), 648–656.

[40] Brenna Li, Noah Crampton, Thomas Yeates, Yu Xia, Xirong Tian, and Khai Truong. 2021. Automating Clinical Documentation with Digital Scribes: Understanding the Impact on Physicians. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3411764.3445172

[41] 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

[42] Zhuoyang Li, Minhui Liang, Hai Trung Le, Ray Lc, and Yuhan Luo. 2023. Exploring Design Opportunities for Reflective Conversational Agents to Reduce Compulsive Smartphone Use. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) *(CUI '23)*. Association for Computing Machinery, New York, NY, USA, Article 37, 6 pages. https://doi.org/10.1145/3571884.3604305

[43] Peter R Lichstein. 1990. *The medical interview.* Butterworths, Boston, USA.

[44] Aijing Luo, Lu Qin, Yifeng Yuan, Zhengzijin Yang, Fei Liu, Panhao Huang, and Wenzhao Xie. 2022. The effect of online health information seeking on physician-patient relationships: systematic review. *Journal of Medical Internet Research* 24, 2 (2022), e23354.

[45] Elizabeth Magnan, Melissa Gosdin, Daniel Tancredi, and Anthony Jerant. 2021. Pilot randomized controlled trial Protocol: Life context-informed pre-visit planning to improve care plans for primary care patients with multiple chronic conditions including diabetes. *Journal of Multimorbidity and Comorbidity* 11 (2021), 26335565211062387.

[46] Mairead Murphy, Chris Salisbury, Anne Scott, Lucia Sollazzi-Davies, and Geoff Wong. 2022. The person-based development and realist evaluation of a pre-consultation form for GP consultations. *NIHR Open Research* 2 (2022).

[47] Lin Ni, Chenhao Lu, Niu Liu, and Jiamou Liu. 2017. MANDY: Towards a Smart Primary Care Chatbot Application. In *Knowledge and Systems Sciences*, Jian Chen, Thanaruk Theeramunkong, Thepachai Supnithi, and Xijin Tang (Eds.). Springer Singapore, Singapore, 38–52.

[48] Ai Nishida and Osamu Ogawa. 2022. The Effect of a Pre-consultation Tablet-Based Questionnaire on Changes in Consultation Time for First-Visit Patients With Diabetes: A Single-Case Design Preliminary Study. *Cureus* 14, 11 (2022).

[49] OpenAI. 2023. GPT-4. https://www.openai.com/research/gpt-4.

[50] Vikas N O'Reilly-Shah. 2017. Factors influencing healthcare provider respondent fatigue answering a globally administered in-app survey. *PeerJ* 5 (2017), e3785.

[51] Venkataraman Palabindala, Amaleswari Pamarthy, and Nageshwar Reddy Jonnalagadda. 2016. Adoption of electronic health records and barriers. *Journal of community hospital internal medicine perspectives* 6, 5 (2016), 32643.

[52] Kaya J Peerdeman, Chris Hinnen, Liesbeth M van Vliet, and Andrea WM Evers. 2021. Pre-consultation information about one's physician can affect trust and treatment outcome expectations. *Patient Education and Counseling* 104, 2 (2021), 427–431.

[53] Stephen R Porter, Michael E Whitcomb, and William H Weitzer. 2004. Multiple surveys of students and survey fatigue. *New directions for institutional research* 2004, 121 (2004), 63–73.

[54] 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.

[55] Bahram Rahman, Andrew P Costa, Anastasia Gayowsky, Ahmad Rahim, Tara Kiran, Noah Ivers, David Price, Aaron Jones, and Lauren Lapointe-Shaw. 2023. The association between patients' timely access to their usual primary care physician and use of walk-in clinics in Ontario, Canada: a cross-sectional study. *Canadian Medical Association Open Access Journal* 11, 5 (2023), E847–E858.

[56] Mark Rickenbach. 2019. Enhancing the medical consultation with prior questions including ideas, concerns and expectations. *Future Healthcare Journal* 6, Suppl 1 (2019), 181.

[57] Elizabeth A Rider and Constance H Keefer. 2006. Communication skills competencies: definitions and a teaching toolbox. *Medical education* 40, 7 (2006), 624–629.

[58] Ragnhild Klingenberg Røed, Gunn Astrid Baugerud, Syed Zohaib Hassan, Saeed S Sabet, Pegah Salehi, Martine B Powell, Michael A Riegler, Pål Halvorsen, and Miriam S Johnson. 2023. Enhancing questioning skills through child avatar chatbot training with feedback. *Frontiers in Psychology* 14 (2023).

[59] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 300–325. https://doi.org/10.18653/v1/2021.eacl-main.24

[60] Mindy K Ross, Sarah Friedman, Ilana Radparvar, and Gery Ryan. 2022. Partnered decision support: Parental perspectives of completing a pre-visit pediatric asthma questionnaire via the patient portal. *Pediatric Pulmonology* 57, 1 (2022), 100–108.

[61] Chris Salisbury, Terjinder Manku-Cott, Laurence Moore, Melanie Chalder, and Deborah Sharp. 2002. Questionnaire survey of users of NHS walk-in centres: observational study. *British Journal of General Practice* 52, 480 (2002), 554–560.

[62] Chris Salisbury and James Munro. 2003. Walk-in centres in primary care: a review of the international literature. *British Journal of General Practice* 53, 486 (2003), 53–59.

[63] Tait Shanafelt and Clair Kuriakose. 2023. Widespread Clinician Shortages Create a Crisis that Will Take Years to Resolve. *NEJM Catalyst Innovations in Care Delivery* 4, 3 (2023).

[64] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* (2023), 1–9.

[65] Christine A Sinsky, Thomas A Sinsky, and Ellie Rajcevich. 2015. Putting pre-visit planning into practice. *Family Practice Management* 22, 6 (2015), 30–38.

[66] Nadia Sourial, Janusz Kaczorowski, Amelie Quesnel-Vallee, Marie Therese Lussier, Vladimir Khanassov, Mylaine Breton, Elise Develay, Geraldine Layani, Claire Godard-Sebillotte, Alayne Adams, et al. 2023. Evaluation of a virtual pre-consultation tool for older adults in primary care: Results from a randomized trial.

[67] Trista J Stankowski-Drengler, Jennifer L Tucholka, Jordan G Bruce, Nicole M Steffens, Jessica R Schumacher, Caprice C Greenberg, Lee G Wilke, Bret Hanlon, Jennifer Steiman, and Heather B Neuman. 2019. A randomized controlled trial evaluating the impact of pre-consultation information on Patients' perception of information conveyed and satisfaction with the decision-making process. *Annals of surgical oncology* 26 (2019), 3275–3281.

[68] Zhaoyuan Su, Lu He, Sunit P Jariwala, Kai Zheng, and Yunan Chen. 2022. " What is Your Envisioned Future?": Toward Human-AI Enrichment in Data Work of Asthma Care. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[69] Mariska E Te Pas, Werner GMM Rutten, R Arthur Bouwman, and Marc P Buise. 2020. User experience of a chatbot questionnaire versus a regular computer questionnaire: prospective comparative study. *JMIR Medical Informatics* 8, 12 (2020), e21982.

[70] 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* (2023), 1–11.

[71] Andrew Reyner Wibowo Tjiptomongsoguno, Audrey Chen, Hubert Michael Sanyoto, Edy Irwansyah, and Bayu Kanigoro. 2020. Medical chatbot techniques: a review. *Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4* (2020), 346–356.

[72] John M Travaline, Robert Ruchinskas, and Gilbert E D'Alonzo. 2005. Patient-physician communication: why and how. *Journal of Osteopathic Medicine* 105, 1 (2005), 13–18.

[73] Shaghayegh Vahdat, Leila Hamzehgardeshi, Somayeh Hessam, and Zeinab Hamzehgardeshi. 2014. Patient involvement in health care decision making: a review. *Iranian Red Crescent Medical Journal* 16, 1 (2014).

[74] Margot J. van der Goot and Tyler Pilgrim. 2020. Exploring Age Differences in Motivations for and Acceptance of Chatbot Communication in a Customer Service Context. In *Chatbot Research and Design*, Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger, and Petter Bae Brandtzaeg (Eds.). Springer International Publishing, Cham, 173–186.

[75] Marta van Zanten, John R Boulet, and Danette McKinley. 2007. Using standardized patients to assess the interpersonal skills of physicians: six years' experience with a high-stakes certification examination. *Health communication* 22, 3 (2007), 195–205.

[76] Lidewij Eva Vat, Mike Warren, Susan Goold, Everard Davidge, Nicole Porter, Tjerk Jan Schuitmaker-Warnaar, Jacqueline EW Broerse, and Holly Etchegary. 2020. Giving patients a voice: a participatory evaluation of patient engagement in Newfoundland and Labrador Health Research. *Research Involvement and Engagement* 6 (2020), 1–14.

[77] Jonathan S Wald, Alexandra Businger, Tejal K Gandhi, Richard W Grant, Eric G Poon, Jeffrey L Schnipper, Lynn A Volk, and Blackford Middleton. 2010. Implementing practice-linked pre-visit electronic journals in primary care: patient and physician use and satisfaction. *Journal of the American Medical Informatics Association* 17, 5 (2010), 502–506.

[78] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2023. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data. arXiv:2301.05843 [cs.HC]

[79] John D Wells, Damon E Campbell, Joseph S Valacich, and Mauricio Featherman. 2010. The effect of perceived novelty on the adoption of information technology innovations: a risk/reward perspective. *Decision Sciences* 41, 4 (2010), 813–843.

[80] Yijin Wu. 2021. Empathy in nurse-patient interaction: a conversation analysis. *BMC nursing* 20, 1 (2021), 1–6.

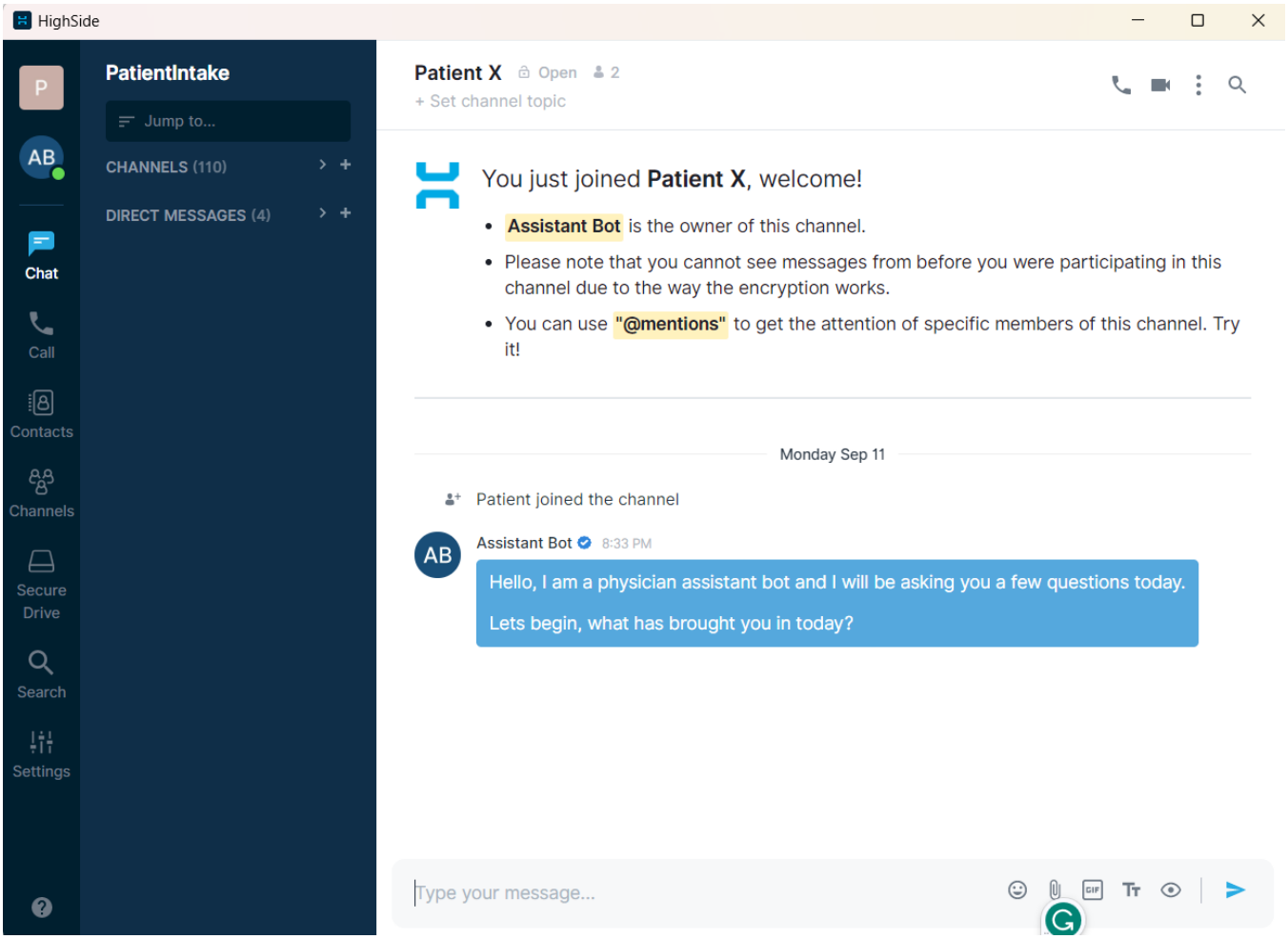
[81] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 3 (2020), 1–37.

[82] J.D. Zamfirescu-Pereira, Heather Wei, Amy Xiao, Kitty Gu, Grace Jung, Matthew G Lee, Bjoern Hartmann, and Qian Yang. 2023. Herding AI Cats: Lessons from Designing a Chatbot by Prompting GPT-3. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (, Pittsburgh, PA, USA,) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2206–2220. https://doi.org/10.1145/3563657.3596138

[83] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. 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 437, 21 pages. https://doi.org/10.1145/3544548.3581388

[84] Claudia Zanini, Paolo Maino, Jens Carsten Möller, Claudio Gobbi, Monika Raimondi, and Sara Rubinelli. 2016. Enhancing clinical decisions about care through a pre-consultation sheet that captures patients' views on their health conditions and treatments: A qualitative study in the field of chronic pain. *Patient education and counseling* 99, 5 (2016), 747–753.

A SUPPLEMENTARY TABLES AND FIGURES



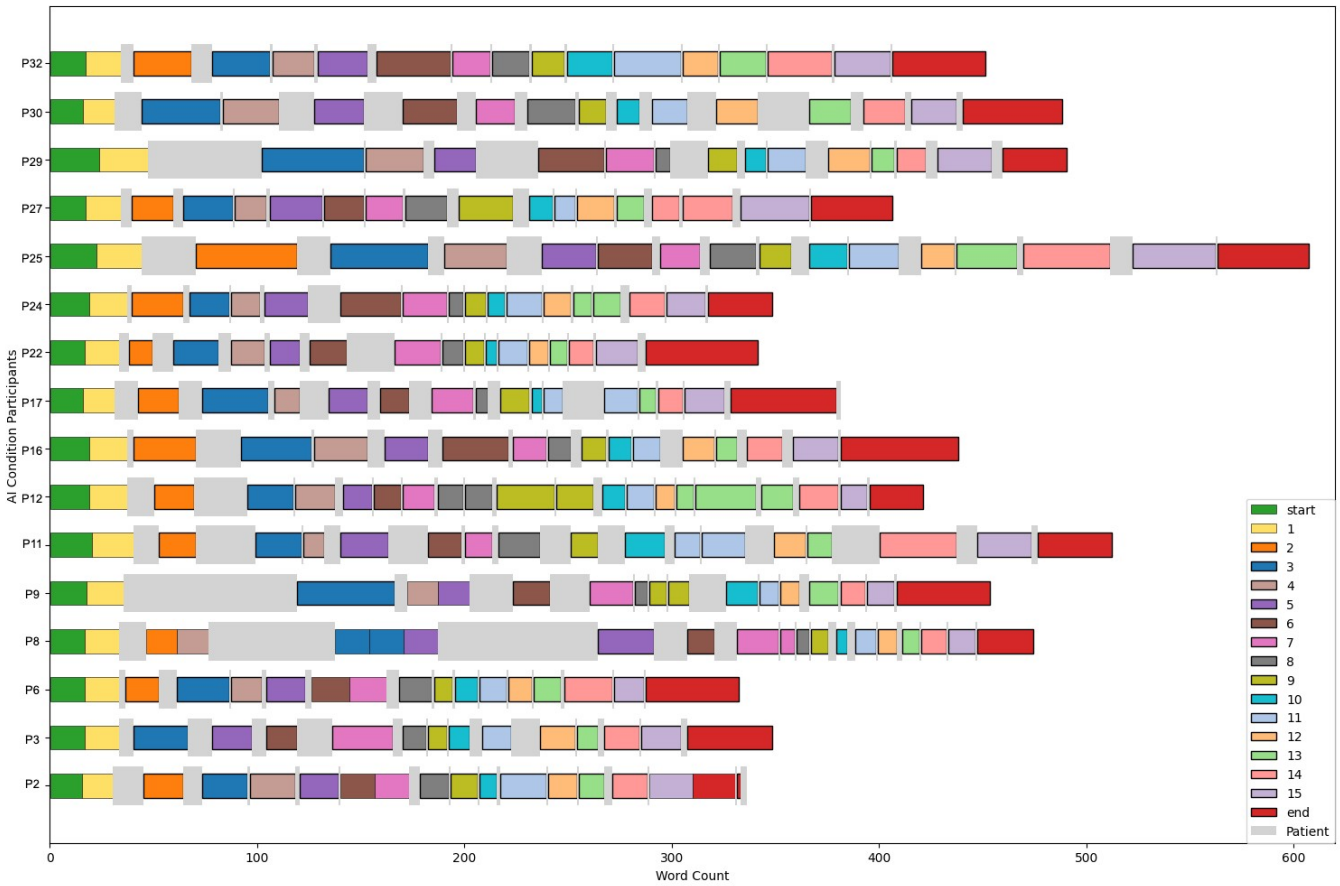
Supplementary Figure 1: An example of the Highside interface shown to participants.

Supplementary Table 1: Conversation messages sent and word count breakdown among the conditions and senders.

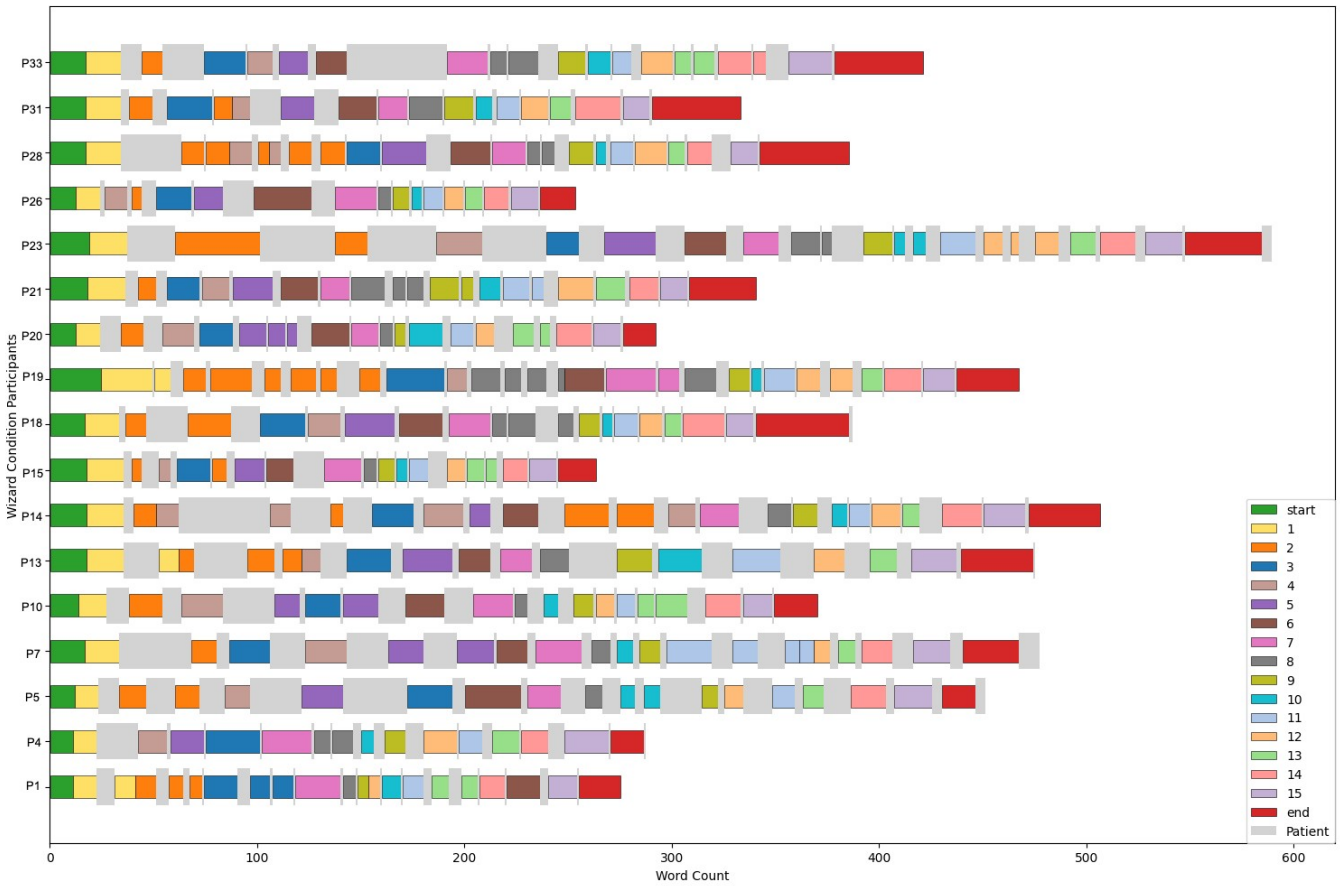
		Agents		Patients	
		AI Condition	Wizard Condition	AI Condition	Wizard Condition
Messages Sent	Mean ± Std (Min, Max)	16.06 ± 1.43 (14, 20)	18.94 ± 2.7 (15, 27)	15.87 ± 1.58 (13, 19)	19.76 ± 4.53 (15, 30)
Word Count	Mean ± Std (Min, Max)	335.19 ± 60 (270, 491)	284.59 ± 54.52 (206, 391)	92.25 ± 51.93 (33, 196)	102.06 ± 58.53 (46, 226)

Supplementary Table 2: Conversation codes and their frequency counts in both AI and Wizard agents' interactions with participants.

Conversation Code	AI Agent (count)	Wizard Agent (count)
start+1	16	17
1.1		1
1.1+2		1
2	11	14
2.1		7
2.2		5
2.3		3
2.4		1
2.5		1
2.1+4		3
2.2+4		1
2+4	1	1
3	15	17
3+3.1+5	1	
3.1		1
3.2		1
4	13	14
4.1		1
4+5	1	
5	15	16
5.1		2
5.2		1
6	14	16
6+7	2	
7	15	17
7.1		1
8	16	17
8.1	1	7
8.2		3
8.3		1
8.4		1
9	16	16
9.1	2	1
9+12		1
10	16	17
10.1		2
11	16	17
11.1	2	2
11.2		1
11.3		1
12	15	16
12.1		2
12.2		1
13	16	18
13.1	2	4
13.2	1	
14	16	16
14.1	1	1
15	15	17
Grand Total	255	321



Supplementary Figure 2: The order of questions asked by AI agent for each participant assigned to that condition.



Supplementary Figure 3: The order of questions asked by Wizard agents for each participant assigned to that condition.

Supplementary Table 3: Semi-structured interview themes, codes, and quotes

Theme	Codes	Quotes
Overall experience	Natural and engaging	I could have easily thought I was talking to a human
Overall experience	Conversation was relevant	All the questions were relevant...
Overall experience	Reflect on symptoms	Made me think about how I was feeling more, so I have like better answers for the doctor.
Overall experience	Reflect on medical background	If I had some recent surgeries or things like that I was not even considering. So when the chatbot asked me about that it actually helped me to remember those things.
Overall experience	Not feeling rushed	I can literally wait five minutes to think about how I feel, to think about things that have been happening to me, and can write them down.
Conversation content	More follow-up questions	Maybe the follow up questions that the chatbot can ask are not quite as like specific as a nurse would...
Conversation content	More follow-up questions	Not following up questions that I expected to be followed up.
Privacy concerns + Understanding pre-consultation process	Not concerned about privacy as much	Right at the beginning I was concerned about what kind of information I was gonna have to give out but when I saw it, it didn't seem like anything was compromising–so yeah, it didn't matter.
Privacy concerns	Important but part of existing system	I know that doctors keep files also, I don't think it [data privacy concerns] would be that different you know?
Privacy concerns	Data storage	I was wondering where is this information being stored? Is it going to my file? Is it going somewhere? It's being saved? Is it being destroyed?
Understanding pre-consultation process	Setting expectations about pre-consultation	I didn't know that, while I was chatting with the chatbot how that information would affect my appointment. Well i knew, but I didn't understand until later.
Understanding pre-consultation process	Comparing with surveys or other forms of preconsultation	There are sometimes medical services [that send surveys] and it is confusing what they exactly want [from you–in the surveys]. But here, there were many questions but they were really to the point.
Presumptions on chatbot capabilities	AI conversing more robotically	I mean that's a very unreal expectation of me... like an AI to talk to me as a human. That's something, if it's there, it's amazing, if it's not there, it's not like something I'm missing out on.
Presumptions on chatbot capabilities	AI more reliable than humans for documtentation	When you're doing it on the chatbot, it is better documented than human interactions, there's really no error on the transcription
Presumptions on chatbot capabilities	Doctors better at giving answers	They [doctors] would probably be able to provide answers better than the chatbot [that's why patient didn't bother asking the chatbot any questions].
Presumptions on chatbot capabilities	AI less sophisticated so need simplify response	I was trying to keep it short and not too detailed. Because I know it might be a little tougher for it... So yeah, maybe I was holding back.
Empathy	Too much acknowledgement	[the chatbot] was saying like, "Okay, thank you for providing this information. It seems like you've had this, this and that", but I just wanted it to move on...
Empathy	Too much compassion	When you know it's a chatbot and it's an automatic answer that you're getting, "I'm sorry to hear that", it just seems very very fake.
Empathy	Too much appreciation	The "thank you for providing the information", that was repetitive.
Guiding patients back to the clinic	Summary feedback	I'd like to have the option to receive the conversation and I would like to see the summary because maybe I misunderstood the question and provided the wrong information. .
Guiding patients back to the clinic	Clinical next steps	I didn't know the doctor was going to see the summary, but when he reviewed it with me, it just all makes sense now.