A Comparative Analysis of Information Gathering by Chatbots, Questionnaires, and Humans in Clinical Pre-Consultation

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

Saba Tauseef
Independent Researcher
Brampton, Ontario, Canada
sabaandtauseef@hotmail.com

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

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

Abstract

Information gathering is an important capability that allows chatbots to understand and respond to users' needs, yet the effectiveness of LLM-powered chatbots at this task remains underexplored. Our work investigates this question in the context of clinical pre-consultation, wherein patients provide information to an intermediary before meeting with a physician to facilitate communication and reduce consultation inefficiencies. We conducted a study at a walk-in clinic with 45 patients who interacted with one of three conversational agents: a chatbot, a questionnaire, and a Wizard-of-Oz. We analyzed patients' messages using metrics adapted from Grice's maxims to assess the quality of information gathered at each conversation turn. We found that the Wizard and LLM were more successful than the questionnaire because they modified questions and asked follow-ups when participants provided unsatisfactory answers. However, the LLM did not ask nearly as many follow-up questions as the Wizard, particularly when participants provided unclear answers.

CCS Concepts

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

Keywords

Pre-consultation, chatbot, LLM, primary care, walk-in clinic

ACM Reference Format:
Brenna Li, Saba Tauseef, Khai N. Truong, and Alex Mariakakis. 2025. A Comparative Analysis of Information Gathering by Chatbots, Questionnaires, and Humans in Clinical Pre-Consultation. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3706598.3713613

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713613

1 Introduction

Large language models (LLMs) have accelerated the capabilities of chatbots to carry out intelligible and human-like conversations. From customer service [13, 31, 70] to public health [30, 41, 107], chatbots are being used to collect information from users and provide personalized assistance tailored to individuals' needs in an automated manner. These two separate tasks — information gathering and task execution — are equally vital to the success of chatbots in their target applications. For instance, errors in diagnostic chatbots often arise from flawed decision-making driven by incomplete symptom information [28, 59, 111]. While most contemporary evaluations of LLM-powered chatbots are primarily concerned with task execution, our work focuses strictly on studying information gathering as a necessary and critical prerequisite.

In order to study information gathering without confounders from other objectives, we selected an application domain where a chatbot was strictly designed for this purpose: clinical pre-consultation. During pre-consultation, patients share preliminary information about their health concerns and medical history to an intermediary, which then relays key details to physicians to help them better prepare for their appointment [94]. The intermediary usually takes one of two forms. The first entails a web-based questionnaire that is easy to distribute [4, 44, 94] but suffers from the same survey fatigue associated with other static forms [22, 74, 80]. The second entails a human intake nurse asking questions. While nurses can convey empathy and adapt their line of questioning to patients' responses [8, 48, 68], they are a valuable resource in clinics that could be used for other services. Balancing the trade-offs between these two approaches, researchers have proposed chatbots that can be distributed at scale with adaptive and empathetic conversational abilities [69, 99]. Prior work has investigated the acceptability of such a chatbot in clinical settings [57]; however, to the best of our knowledge, the quality of the information gathered by a pre-consultation chatbot relative to existing solutions remains underexplored. We operationalize this goal with two research questions:

(RQ1) How do dynamic and static pre-consultation processes compare in terms of the quality of information gathered from participants' initial responses?

(RQ2) How do LLM-powered chatbots and humans compare in the ways that they follow up on unsatisfactory question responses?

To answer these questions, we conducted a study at a real walk-in clinic with 45 patients who underwent pre-consultation with one of three conversational agents: a static questionnaire, an LLM-powered chatbot, and a Wizard-of-Oz. The agents were all designed to gather answers to the same 15 questions commonly used in clinical pre-consultation, but only the latter two were able to adapt to participants' responses. We analyzed the dialogues using four metrics (clarity, depth, informativeness, and relevance) derived from Grice's maxims [33, 109], which outline principles that guide effective and cooperative communication between speakers. After assessing the different patterns associated with unsatisfactory responses by participants according to our conversation metrics, we examined how often each agent was able to solicit a satisfactory response with and without follow-up, as well as the scenarios where each of the dynamic agents failed to remedy an unsatisfactory response. We found that the dynamic questions generated by the LLM and Wizard agents were more effective than those from the static questionnaire in gathering information that was clear, detailed, informative, and relevant. However, the LLM agent still had room for improvement in its follow-up questioning, particularly when participants provided unclear answers. To summarize, our main contributions are as follows:

- An approach inspired by Grice's maxims [33, 109] to evaluate an LLM-powered chatbot's information-gathering capabilities,
- Evidence from the analysis of 45 pre-consultation interactions showing that dynamic agents are more effective at gathering information, particularly by using follow-up questions to improve the clarity, depth, informativeness, and relevance of patient responses, and
- Design considerations for future chatbot developers evaluators within and beyond health contexts.

2 Related Work

We begin by providing an overview of how chatbots utilize information-gathering techniques to understand their users. We then shift our focus towards clinical pre-consultation as a prime example where chatbots are being used strictly for information gathering without any other objectives. We conclude by enumerating various techniques researchers have used to evaluate chatbots across these domains.

2.1 Chatbots for Gathering Information

Gathering information is a crucial aspect of many chatbot interactions. Chatbots built using LLMs can be easily designed to engage in semantically coherent conversations, exhibiting human-like traits like acknowledgments of previous statements and adaptive questioning to promote self-disclosure from users [47, 49, 60, 99, 109]. These affordances have led to deployments of information-gathering chatbots across a variety of domains [13], including but not limited to education [38, 77, 85], customer service [13, 31, 70], and healthcare [30, 41, 107].

However, deploying LLM-powered chatbots for this purpose does not come without its challenges, particularly when it comes to debugging errors. When a chatbot fails to perform as expected, it can be difficult to determine whether the error stems from faulty information gathering or an issue with decision making and task execution [45, 106]. The black-box nature of LLMs further complicates this issue, especially in healthcare, where the risk of uncertainty greatly impacts adoption [28]. For instance, when a diagnosis chatbot provides an incorrect assessment, it is often unclear whether the chatbot failed to determine the correct diagnosis or the chatbot was given incomplete or ambiguous information [10, 59] This lack of transparency makes it challenging to address the root cause of such errors.

2.2 Clinical Pre-consultation and Health Chatbots

Clinical pre-consultation involves gathering preliminary information about a patient prior to their visit to facilitate the physician consultation [94, 95, 102]. This process is often structured yet open-ended [34, 44], starting with questions about patients' reasons for the visit and relevant medical history, and then ending with the opportunity for patients to voice any concerns they would like to have addressed during the consultation [25, 27, 78]. There are various benefits of pre-consultation for both physicians and patients, largely stemming from improved communication between the two parties. For physicians, pre-consultation improves efficiency and reduces documentation burden [94, 104]. Patients have also shown greater engagement, reduced anxiety, and appointment satisfaction when they go through a pre-consultation [86, 112].

Researchers have proposed that chatbots can automate many clinical processes to alleviate human resources. These studies are often motivated by the strong performance of LLMs in both answering medical questions and conducting engaging conversations with human users [1, 9, 101]. However, most existing work has focused on diagnostic chatbots that provide medical recommendations after the user has explained their symptoms or concerns [52, 93, 100]. Despite the potential benefits, diagnostic chatbots pose significant detriments in the event of misdiagnosis or improper guidance [6, 28]. The accuracy of these chatbots also depends on both the quality of information provided by the user and the reasoning of the chatbot itself [28, 59, 111].

In contrast, pre-consultation chatbots are strictly designed to collect comprehensive patient information to better prepare physicians for consultations [4, 63, 71]. This nuance makes pre-consultation chatbots an exemplary testbed for exploring the information-gathering capabilities of LLM-powered chatbots without confounders from other objectives. While Ni et al. [69] and Te Pas et al. [99] proposed potential designs for this purpose, we recently deployed and evaluated such an LLM-powered chatbot [57]. Our work was primarily concerned with how patients felt about interacting with a pre-consultation and how those opinions could inform the design of future systems. However, we did not examine the quality of the information exchanged and the potential chatbot behaviors that led to successes or failures in that process. This paper provides a rigorous analysis of pre-consultation chatbot transcripts to fill this gap.

2.3 Evaluating Chatbots and Conversations

To date, there are no widely accepted methods for robustly evaluating LLMs across domains, due in large part to the lack of consensus on appropriate evaluation criteria and metrics [2]. For the purposes of our work, we categorize existing approaches into three categories: objective evaluations, subjective evaluations by users, and subjective evaluations by researchers. We elaborate on these three categories below.

2.3.1 Objective Evaluations. Objective chatbot evaluations often involve developing an algorithm to assess a chatbot's performance, making them ideal for running repeatable and reproducible experiments at scale [7, 91]. Some of these approaches prioritize the semantics of the conversation, attributing success to formal definitions of reply coherence [91]. For example, BLEU [76] and ROUGE [61] scores have historically been used across natural language processing to rate the quality of machine-generated text against human-generated references. However, these scores are quickly becoming less applicable as LLM-powered chatbots increasingly excel at producing fluent and coherent sentences. A second set of objective approaches evaluates chatbots from the perspective of task completion rate [14], but this limits the evaluation to a specific domain, making it difficult to generalize the findings to other applications [65].

2.3.2 Subjective Evaluations by Users. Subjective chatbot evaluations often involve recruiting human participants to provide feedback on the chatbot interactions. Many researchers have argued that user-based evaluations are accurate predictors of a chatbot's performance and adoption [2, 90], yet the criteria that users are asked to assess can depend on the chatbot's intended purpose [16]. Metrics can include but are not limited to people's opinions of the chatbot's conversation efficiency [103], personality and tone [40, 98], interface [40], and engagement [21, 88]. Despite being regarded as more comprehensive than objective evaluations, subjective evaluations are often inconsistent and difficult to reproduce [16, 65]. Some qualitative evaluations are conducted at scale with crowd workers [89, 108], while others rely on small-sample usability studies [40]. Even with the same users, results may vary across evaluations [2]. Xiao et al. [108] attempted to standardize these metrics with a meta-metric evaluation, but their participants and data samples still exhibited significant variation.

2.3.3 Subjective Evaluations by Researchers. To evaluate chatbots in a more rigorous and methodical manner while still incorporating elements of human interpretation, researchers have proposed and applied various techniques to evaluate human-chatbot interactions themselves. Many of these approaches are based on Grice's maxims [33], which describe the rules followed by people in conversation. The four maxims — quantity, quality, relevance, and manner — have been widely used in chatbot evaluation studies to assess the clarity and appropriateness of information exchanged between users and conversational agents [42, 92, 108].

In healthcare specifically, existing evaluation metrics by researchers have largely focused on patient-physician communication, such as how well physicians allow patients to share their concerns using accessible language [39], or how physicians structure their questioning to maximize patient disclosure [20]. Furthermore, studies like the one by Ben-Shabat et al. [10] have shown that many existing healthcare chatbots struggle with satisfactory data collection, although this finding was supported by vignettes rather than formal metrics.

To the best of our knowledge, none of the aforementioned techniques have been used specifically to evaluate a chatbot's capacity to gather information, especially in the context of clinical pre-consultation. Our work adapts Grice's maxims for this exact purpose, comparing a chatbot against both a static questionnaire and a Wizard-of-Oz medical professional to identify opportunities for improvements.

3 Methods

This section describes the study protocol and analysis plan that was used to investigate our research questions. The study was approved by the Research Ethics Board at the University of Toronto (Protocol #41033) and the supervising manager at the clinic where we held our study.

3.1 Participants

Forty-five patients were recruited from a walk-in medical clinic in the Greater Toronto Area. The exclusion criteria included being under 18 years old and having difficulty reading and writing in English. Additionally, patients could only participate in the study if they were visiting the walk-in clinic for a new presenting symptom, as that would be the most relevant clinical scenario for pre-consultation. A full breakdown of the participants' demographics can be found in Table 1.

3.2 Study Conditions

Since the walk-in clinic where we conducted our study did not have a pre-consultation process in place but had an interest in exploring its use, we used the list of pre-consultation questions from our own prior work [57], which itself was compiled from existing pre-consultation questionnaires [68, 86, 104, 112]. To answer our research questions regarding the ability of an LLM-powered chatbot to gather information relative to humans and static questionnaires, we implemented three agents:

- **Questionnaire:** This agent acted as a turn-based questionnaire without any flexibility in changing the questions' order or content.
- **LLM:** Rather than creating our own LLM-powered chatbot, this agent utilized a chatbot powered by OpenAI's GPT-4 and initialized with the prompt we had previously tested [57]. The prompt instructed the chatbot to take a medically professional tone and not make any medical recommendations. The prompt also gave the chatbot the liberty to add, skip, reword, or reorder questions; however,

Table 1: The demographics of our patient participants (N = 45). The percentages reported are with respect to the overall study cohort.

Categories		Questionnaire, N (%)	LLM, N (%)	Wizard, N (%)	Total, N (%)
Gender	Male	7 (15.6%)	6 (13.3%)	9 (20.0%)	23 (48.9%)
	Female	6 (13.3%)	10 (22.2%)	7 (15.6%)	22 (51.1%)
Age	18–24	0 (0%)	0 (0%)	2 (4.4%)	2 (4.4%)
	25–34	10 (22.2%)	11 (24.4%)	7 (15.6%)	28 (62.2%)
	35–44	2 (4.4%)	2 (4.4%)	4 (8.9%)	8 (17.8%)
	45–54	1 (2.2%)	0 (0%)	2 (4.4%)	3 (6.7%)
	55–64	0 (0%)	3 (6.7%)	1 (2.2%)	4 (8.9%)
Education Level	High school	0 (0%)	2 (4.4%)	4 (8.9%)	6 (13.3%)
	College or technical certificate	1 (2.2%)	4 (8.9%)	1 (2.2%)	6 (13.3%)
	University Bachelor's degree	7 (15.6%)	8 (17.8%)	7 (15.6%)	22 (48.9%)
	Graduate or professional degree	4 (8.9%)	1 (2.2%)	4 (8.9%)	9 (20.0%)
	Prefer not to say	1 (2.2%)	1 (2.2%)	0 (0%)	2 (4.4%)
Technology Proficiency	Average	3 (6.7%)	3 (6.7%)	3 (6.7%)	9 (20.0%)
	Somewhat above average	8 (17.8%)	11 (24.4%)	7 (15.6%)	26 (57.8%)
	Far above average	2 (4.4%)	2 (4.4%)	6 (13.3%)	10 (22.2%)

Table 2: The list of pre-consultation questions for which all study agents were expected to collect answers from participants. These questions were taken from our previous work [57].

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?

it included an explicit warning not to double-barrel questions.

- **Wizard:** This agent was operated by one of two medically-licensed healthcare professionals who acted as a Wizard-of-Oz chatbot. Their instructions matched those that were given to the chatbot for the LLM agent.

The Questionnaire agent served as a control condition to isolate the effects of the adaptive interactions exhibited by the LLM and Wizard agents, addressing **RQ1**. The Wizard served the opposite role, providing a benchmark for the highest level of adaptability possible and allowing us to investigate **RQ2**.

All interactions with participants took place over Highside[1] — a HIPAA-compliant and secure online messaging platform with a basic text-based interface — to standardize the agents' presentation. Highside also served as a proxy for redacting personally identifiable information before messages were transmitted to OpenAI for the LLM agent. To do this, a researcher copy-and-pasted responses between the Highside and a separate OpenAI interface while removing any identifying information (e.g., name, address).

3.3 Study Design

Administrative staff provided a brief introduction to the study to all eligible patients who entered the walk-in clinic. If a patient expressed interest, a research team member approached them in the waiting room to explain the protocol in greater detail. To minimize disruption to patients' scheduled appointments, the research team only approached those with waiting times exceeding 30 minutes.

[1] https://highside.io/

After obtaining consent, participants were taken to a study room within the clinic, where they first completed a pre-study questionnaire that included questions about their demographics and experiences with chatbots. Participants were then randomly assigned to one of the three study conditions. Regardless of their assigned condition, participants were informed that they would interact with a chatbot that would ask questions about their medical concerns and relevant medical history. They were also made aware that a summary of their conversation with the chatbot would be provided to their doctor for review before their consultation; a medical resident actually generated this summary in real time. After participants were introduced to the chatbot interface, they completed the pre-consultation process which typically took between 10 and 15 minutes. Participants were then given a post-interaction survey to capture their experiences and feedback.

At the end of the study, participants were fully debriefed about the conditions and capabilities of the chatbot with which they conversed. They were then escorted back to the waiting room, where the clinic staff would call them into a consultation room for their normal consultation with a doctor.

3.4 Analysis Codebook

Our work draws inspiration from Xiao et al. [109] by focusing our analysis on the ability of chatbots to gather information. Using Grice's maxims [33], we created a brief codebook that would allow us to methodically characterize the information communicated at each conversation turn. Our new metrics — clarity, depth, informativeness, and relevance — are defined in Table 3. The "manner" and "relation" maxims were directly mapped to message codes that we called "clarity" and "relevance" respectively. Clarity was defined as the level of ambiguity in the patient's response, while relevance was defined as the degree to which the patient's response addressed the question rather than another topic. Finally, the "quantity" maxim was split into two separate message metrics: depth, which related to the level of detail in the patient's response, and informativeness, which was the degree to which the patient's response added new knowledge. Although each of these metrics lies on a spectrum, we split them into binary codes (i.e., clarity ∈ [clear, unclear]) for consistency in our analyses in light of diversity in patients' verbosity and case complexity.

It is important to note that, similar to Xiao et al. [109], we omitted Grice's maxim of "quality" from our analysis because participants generally have little incentive to provide false information to healthcare providers. This decision was also motivated by the fact that evaluating the robustness of the information would have required input from the consulting physicians themselves, which would have dramatically impacted their workflow in the busy walk-in clinic. While physicians received a summary of patients' conversations with the agents, the medical professionals who generated those summaries often removed information to promote efficiency and used their own language to enhance clarity, introducing additional variables warranting significant consideration.

Because of the lack of a ground truth for the pre-consultation process, we could not code the conversations according to whether participants provided a complete response to a question. Instead, we used the combination of our four conversation metrics to determine whether participants provided a satisfactory response to each question. For each question, we examined the first relevant conversation turn to see if they had four favorable codes indicating a response that was clear, deep, informative, and relevant. If it did, we considered the question to be answered in a satisfactory manner. If it did not, we examined any and all follow-up messages to see if deficiencies were remedied. Those with addressed deficiencies were considered satisfactory, while those that did not were considered unsatisfactory.

3.5 Analysis Procedure

Two researchers manually coded all 45 participant dialogues: one who is an expert in human-computer interaction and another with extensive medical training. The researchers coded 730 conversation turns according to the metrics listed in Table 3. The researchers also coded each turn according to the questions listed in Table 2 so that we could readily identify which questions were sufficiently answered. Note that this mapping was many-to-many for the Wizard and LLM agents. The same question may have been answered over multiple conversation turns; conversely, the same patient response may have addressed multiple questions.

Both researchers read and coded all of the conversations using an iterative process to maximize validity and rigor. The two researchers first independently coded ~10% of the conversations. After they convened to discuss the differences and established more consistent coding criteria, they reviewed the rest of the conversation data using the updated coding criteria. Each conversation turn was assigned four codes corresponding to our conversation metrics, resulting in a total of 2,920 codes being assigned by each researcher. According to Cohen's Kappa score, the inter-rater reliability between the researchers was 0.85, indicating substantial agreement. Using the codes, we established whether each response was satisfactory or not using the scheme described previously. An added benefit of this process was that it enabled us to identify seamless and problematic exchanges according to the number of unsatisfactory responses given by participants.

4 Results

The final dataset comprised transcripts from 13 participants who interacted with the Questionnaire agent, 16 participants who interacted with the LLM agent, and 16 participants who interacted with the Wizard agent. Since prior work has already investigated patients' perspectives on pre-consultation chatbots [57], we briefly discuss this topic from the lens of conversational efficiency and information gathering. We then delve into our two research questions, examining the contributing factors that led to deficiencies in participants' initial responses to agent questions (**RQ1**) and how the dynamic agents were able to address these initial shortcomings with follow-up questions (**RQ2**).

Recognizing that the answers to these questions may be influenced by the complexity of the medical concerns raised by participants, we forgo stratifying our analyses in this way;

Table 3: The metrics used to assess the quality of patient's responses to agents' questions. Messages were assigned a binary value for each metric (i.e., clarity ∈ [clear, unclear]) to simplify our analyses.

Conversation Metric	Relevant Gricean Maxim	Definition	Example of Failure
Clarity	Manner	The level of unambiguity in the patient's response	**Chatbot:** `Are you currently taking any medications?` **Participant:** `Yes, the pill.`
Depth	Quantity	The level of detail to which the patient's response answered the question	**Chatbot:** `Do you have any chronic medical conditions?` **Participant:** `Yes`
Informativeness	Quantity	The degree to which the patient's response added new insights about their medical concern or history	**Chatbot:** `What is the reason for your visit today?` **Participant:** `Pain on my right earlobe` **Chatbot:** `What symptoms are you having?` **Participant:** `Pain on my right ear`
Relevance	Relation	The degree to which the patient's response addressed the question rather than another topic	**Chatbot:** `How would you rate the discomfort these symptoms are causing you on a scale of 1-10?` **Participant:** `My primary concern is the lump on my back.`

however, an anonymized list of patients and the reasons for their visit can be found in Supplementary Table S1.

4.1 Overview of Agent Behaviors

Table 4 shows high-level trends associated with each agent: their success at eliciting satisfactory responses from participants, their propensity for modifying the initial set of questions included in their prompt, and their propensity for adding follow-up questions. We found that the Questionnaire agent performed the worst in eliciting satisfactory responses, only achieving a success rate of 79.7%. In contrast, the LLM and Wizard agents demonstrated higher success rates at 85.3% and 85.6%, respectively. While both agents reordered and skipped questions at a similar rate, the Wizard agent asked over ×5 as many follow-up questions compared to the LLM agent.

4.1.1 General Perceptions of Pre-consultation Agents. In the post-interaction survey, participants generally appreciated the value of a pre-consultation process. However, many could easily discern when they were interacting with the static agent. This was evident in comments like the one from P32, who remarked:

> "*It felt weird to write down my medical history in this questionnaire as opposed to talking to a human.*"
> (P32, Questionnaire agent)

The static nature of the questionnaire also frustrated some participants. For instance, P29 noted that they could not ask questions back to the agent, while P4 expressed:

> "*I wished it would learn from previous answers, so I don't have to repeat myself.*"
> (P4, Questionnaire agent)

These remarks reflect the common challenges associated with static approaches to information gathering.

In contrast, participants who interacted with the LLM agent described the experience as more thorough and dynamic compared to static intake methods like paper forms. Confirming the potential efficiency and engagement LLM-powered chatbots can afford, P5 commented:

> "*It was more thorough than a piece of paper with a box asking why I'm there, which would help save physicians time because they'd already have an idea of what's going on prior to seeing you.*"
> (P5, LLM agent)

However, some participants found that the LLM agent's questions lacked depth. For example, P44 shared:

> "*I like the fact that all the questions are simple and straightforward. But I was expecting more profound questions about my reason for being here.*"
> (P44, LLM agent)

This observation hints at the difference in the follow-up questions that were asked, which we examine later in our analyses.

Participants who conversed with the Wizard agent often reported being impressed by how human-like the interaction felt. P14 captured this sentiment by saying:

> "*Great interaction, doesn't feel like talking to a robot.*"
> (P14, Wizard agent)

This feedback confirms the importance of creating a more natural experience. However, the Wizard agent was not without its flaws. Some participants felt the interaction lacked the level of empathy they desired. As P2 noted:

Table 4: An overview of how each agent went through the pre-consultation process with participants.

Agent	Participants, N	Conversation Turns, N	Satisfactory Responses, N (%)	Reordered Questions, N (%)	Skipped Questions, N (%)	Follow-up Questions, N (%)
Questionnaire	13	198	158 (79.7%)	0 (0%)	0 (0%)	0 (0%)
LLM	16	248	211 (85.3%)	13 (5.2%)	5 (2.0%)	9 (3.6%)
Wizard	16	284	243 (85.6%)	14 (4.9%)	5 (1.8%)	48 (16.9%)

"*It was direct and polite, but it also didn't feel as personable as I would have hoped.*"
(P2, Wizard agent)

These responses suggest that while the Wizard agent excelled at creating a human-like interaction, there is still room to enhance its emotional resonance.

4.2 Responses to Initial Questions by the Agents

To answer **RQ1**, we examine participants' initial responses to the agents' questions and the factors that may have led to various deficiencies. We then investigate how the dynamic agents overcame some of these deficiencies during the information-gathering process.

4.2.1 Initial Response Deficiencies. Table 5 shows how often participants' initial responses were satisfactory across the different agents according to the questions included in the initial prompt. The table also reports the prevalence of the codes that led to these determinations, providing insight into the communication breakdowns that may have occurred during the interactions. We found that the most common problem associated with unsatisfactory responses was lacking depth, followed by lacking clarity and informativeness. In fact, we observed that deficiencies in depth and informativeness were often concurrent, indicating that responses were both non-descriptive and shallow. Most of the participants' responses were relevant, although there were exceptions across all three conditions.

Similar to the result reported in Section 4.1, the dynamic agents solicited higher rates of satisfactory responses after their first attempt at asking each question in the initial prompt compared to the Questionnaire agent. However, a chi-squared test deemed that the difference across the agents was not statistically significant ($\chi^2 = 3.34, n.s.$); the same held true for chi-squared tests on the four conversation metrics. There were more drastic differences across the agents with respect to individual questions, but we did not evaluate the significance of these differences due to a combination of our sample size, the multiple comparisons problem, and the heterogeneity of medical concerns in our study. Instead, we applied a qualitative approach to examining the differences across agents and questions.

Unsurprisingly, straightforward and closed-ended questions yielded higher satisfaction levels among responses. For example, Q3 required participants to rate their discomfort on a scale from 1 to 10, and Q10 required participants to list their allergies. We observed rates of satisfactory responses above 85% across all three agents, including the Questionnaire agent. Conversely, responses to open-ended questions showed a marked decrease in success. Q1 required participants to explain the reason for their visit, while Q8 required participants to provide their medication history. Both of these questions led to significantly worse rates of satisfactory responses, with the latter even approaching 50% by some of the agents.

A notable exception to this trend was Q13, a closed-ended question that resulted in low rates of satisfactory responses across all three conditions. Q13 required participants to report their smoking, drinking, and recreational drug use. These questions are often combined in clinical practice to understand patients' social history. However, asking about all these topics at once may have led to ambiguous or unclear responses. As we observed in our past work [57], participants may have also been reluctant to give lengthy responses to these questions due to their sensitivity or perceived irrelevance to their chief medical concern. Interestingly, the rate of satisfactory responses to this question was the lowest for the LLM agent (50.0%), followed by the Questionnaire agent (61.5%) and the Wizard agent (75.0%). The Questionnaire agent may have seen higher success compared to the LLM agent because by the time the Questionnaire agent reached Q13, participants realized that they had to be verbose since the agent was not adaptive in any way. Meanwhile, a possible explanation for the difference between the LLM and Wizard agents may be that the Wizard was more proactive in changing the question's wording to be more clear.

4.2.2 Patterns of Question Modifications Made by the Dynamic Agents. Upon further examination of the transcripts, we found that the increased success exhibited by the dynamic agents could be attributed to multiple changes they made to the wording of the original questions in their prompt. One way that the dynamic agents reworded the questions was by recognizing and acknowledging what participants had said in previous responses. For example, many participants mentioned their symptoms when explaining the reason for their visit in response to Q1. When the Questionnaire agent asked Q2 ("What symptoms are you experiencing?") without any modification, participants had to repeat information they had already provided, thereby detracting from the informativeness of their response. The dynamic agents, on the other hand, were able to acknowledge what participants had already said, allowing them to get richer information with the next conversation turn:

Q1 – Questionnaire agent: `What is the reason for your visit today?"`
P4: `Ear pain`
Q2 – Questionnaire agent: `What symptoms are you experiencing?`

Table 5: The percentage of questions with satisfactory responses across all three agents.

Question ID	Questionnaire, Satisfactory %					LLM, Satisfactory %					Wizard, Satisfactory %				
	Overall	C	D	I	R	Overall	C	D	I	R	Overall	C	D	I	R
Q1	69.2%	92.3%	84.6%	84.6%	100%	68.8%	87.5%	75.0%	87.5%	100%	81.3%	93.8%	93.8%	87.5%	100%
Q2	61.5%	84.6%	100%	76.9%	84.6%	83.3%	83.3%	100%	100%	91.7%	66.7%	80.0%	100%	86.7%	93.3%
Q3	92.3%	100%	100%	100%	100%	93.8%	93.8%	93.8%	100%	100%	100%	100%	100%	100%	100%
Q4	76.9%	100%	100%	84.6%	92.3%	86.7%	86.7%	100%	100%	100%	86.7%	93.3%	100%	93.3%	100%
Q5	76.9%	88.5%	84.6%	92.3%	92.3%	93.8%	100%	93.8%	93.8%	100%	93.3%	100%	100%	100%	93.3%
Q6	69.2%	69.2%	100%	100%	100%	81.3%	84.4%	100%	100%	93.8%	100%	100%	100%	100%	100%
Q7	92.3%	100%	100%	100%	100%	87.5%	87.5%	93.8%	100%	100%	93.8%	100%	100%	100%	93.8%
Q8	53.8%	96.2%	65.4%	65.4%	100%	87.5%	93.8%	87.5%	87.5%	100%	56.3%	81.3%	75.0%	75.0%	100%
Q9	100%	100%	100%	100%	100%	87.5%	100%	87.5%	87.5%	100%	87.5%	93.8%	93.8%	93.8%	93.8%
Q10	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	87.5%	93.8%	87.5%	93.8%	100%
Q11	80.8%	84.6%	84.6%	100%	100%	87.5%	87.5%	87.5%	100%	100%	62.5%	87.5%	75.0%	81.3%	100%
Q12	84.6%	92.3%	92.3%	92.3%	100%	93.8%	93.8%	93.8%	100%	100%	100%	100%	100%	100%	100%
Q13	61.5%	92.3%	61.5%	61.5%	100%	50.0%	87.5%	50.0%	68.8%	100%	75.0%	87.5%	75.0%	87.5%	100%
Q14	92.3%	96.2%	100%	96.2%	100%	81.3%	93.8%	87.5%	87.5%	100%	83.3%	93.3%	83.3%	93.3%	96.7%
Q15	84.6%	92.3%	100%	92.3%	92.3%	96.9%	96.9%	96.9%	100%	96.9%	100%	100%	100%	100%	100%
Overall	79.7%	92.6%	91.5%	89.7%	96.9%	85.3%	91.8%	89.8%	94.2%	98.8%	84.9%	93.6%	92.2%	92.8%	98.1%

C: clarity, D: depth, I: informativeness, R: relevance

> **P4:** `I am having ear pain`

> **Q1 – LLM agent:** `What is the reason for your visit today?"`
> **P50:** `insect bite`
> **Q2 – LLM agent:** `To help the physician prepare better for your visit, can you share what symptoms you're experiencing due to this insect bite?`
> **P50:** `rash, itchiness, redness, skin is swollen near the bite`

Another way that the dynamic agents reworded questions was by prefacing them with empathetic remarks expressing condolences or reassurance that they would receive sufficient care. These explicit utterances of empathy may have encouraged greater disclosure from participants, as illustrated by this example with P24:

> **Q1 – LLM agent:** `What is the reason for your visit today?`
> **P24:** `anxiety and depression`
> **Q2 – LLM agent:** `I'm sorry to hear that you've been experiencing anxiety and depression. What specific symptoms are you experiencing?`
> **P24:** `difficulty sleeping, don't want to wake up in the morning, constant fear, irritability, loss of interest in things I used to like`

4.3 Responses to Follow-up Questions by the Dynamic Agents

To answer **RQ2**, we examine the utility and characteristics of the follow-up questions that were asked by the dynamic agents.

4.3.1 Impact of Follow-up Questions. Table 6 shows how follow-up questions by the dynamic agents affected their success in collecting satisfactory responses. Across all questions, the LLM agent improved its success rate by 2.9%, while the Wizard agent improved its success rate by 7.2%. These improvements led to a significant difference in the rate of satisfactory responses across the three agents ($\chi^2 = 17.28, p < .01$). Pairwise tests with Bonferroni corrections revealed that the only statistically significant difference was between the Questionnaire and Wizard agents ($\chi^2 = 15.77, p < .01$), although the difference between the Questionnaire and LLM agents was nearly significant ($\chi^2 = 4.79, p = .085$). Most of the improvements demonstrated by the dynamic agents came from the open-ended questions since there was more room for improving participants' initially unsatisfactory responses. Significant improvement was also observed in the sensitive Q13, with the Wizard agent achieving a 25% improvement in response quality.

4.3.2 Likelihood of Follow-up Questions. Table 7 summarizes the frequency of follow-up questions based on whether participants' initial responses were satisfactory. When we examined the likelihood of follow-up questions in reaction to initially unsatisfactory responses, we found that the LLM agent added a question 8 out of 37 (21.6%) possible times, while the Wizard agent added a question 22 out of 41 (53.7%) possible times; this difference was statistically significant ($\chi^2 = 7.13, p < .01$). Although the Wizard agent was far more proactive in remedying unsatisfactory responses, there were still several instances where both dynamic agents failed to ask follow-up questions. At the same time, there were several instances when the Wizard agent asked follow-up questions even when the initial response was satisfactory; the LLM only did this once across 16 participants.

Table 8 illustrates how different deficiencies in participants' responses influenced the likelihood of a follow-up question. Given the aforementioned considerations of heterogeneity and sample

Table 6: The impact follow-up questions by the dynamic agents had on their ability to improve response quality from participants. Empty entries indicate instances when the agent never asked a follow-up question.

Question ID	Questionnaire, Final Satisfactory %	LLM, Satisfactory %		Wizard, Satisfactory %	
		Final	Difference with Follow-up	Final	Difference with Follow-up
Q1	69.2%	68.8%	–	87.5%	+6.3%
Q2	61.5%	83.3%	–	80.0%	+13.3%
Q3	92.3%	93.8%	–	100%	–
Q4	76.9%	86.7%	–	86.7%	–
Q5	76.9%	93.8%	–	100%	+6.7%
Q6	69.2%	81.3%	–	100%	–
Q7	92.3%	87.5%	–	93.8%	–
Q8	53.8%	93.8%	+6.3%	81.3%	+25.0%
Q9	100%	100%	+12.5%	93.8%	+6.3%
Q10	100%	100%	–	100%	+12.5%
Q11	80.8%	93.8%	+6.3%	68.8%	+6.3%
Q12	84.6%	93.8%	–	100%	–
Q13	61.5%	62.5%	+12.5%	100%	+25%
Q14	92.3%	87.5%	+6.3%	90.0%	+6.7%
Q15	84.6%	96.9%	–	100%	–
Overall	79.7%	88.2%	+2.9%	92.1%	+7.2%

Table 7: The rate at which each agent asked a follow-up question depending on the quality of participants' initial responses.

Initial Response Satisfactory?	Agent Follow-up?	Questionnaire, N (%)	LLM, N (%)	Wizard, N (%)
No	No	40 (20.3%)	29 (11.7%)	19 (6.7%)
	Yes	–	8 (3.2%)	22 (7.7%)
Yes	No	158 (79.7%)	210 (84.7%)	217 (76.4%)
	Yes	–	1 (0.4%)	26 (9.2 %)

Table 8: The number of times the agents followed up on participants' messages when they were lacking in one of the conversation metrics, and the degree to which those follow-up questions led to satisfactory responses. Note that messages missing multiple codes were double-counted in this analysis.

Response Deficiency	LLM, N (%)			Wizard, N (%)		
	Never Followed Up	Followed Up with Satisfactory Result	Followed Up with Unsatisfactory Result	Never Followed Up	Followed Up with Satisfactory Result	Followed Up with Unsatisfactory Result
Lack of clarity	19 (95.0%)	1 (5.0%)	–	6 (40.0%)	8 (53.3%)	1 (6.7%)
Lack of depth	18 (69.2%)	7 (26.9%)	1 (3.9%)	4 (21.0%)	14 (73.7%)	1 (5.3%)
Lack of informativeness	8 (57.1%)	6 (42.9%)	–	8 (47.0%)	7 (41.2%)	2 (11.8%)
Lack of relevance	4 (100%)	–	–	3 (60.0%)	2 (40.0%)	–

size, we again resorted to qualitative methods to comment on how the response deficiencies may have led to certain types of follow-up questions.

4.3.3 Patterns of Follow-up Questions Asked by the Dynamic Agents. Reviewing the transcripts associated with the dynamic agents revealed four categories of follow-up questions. The first category, probes for more detail when participants gave shallow responses, was equally prevalent in the LLM and Wizard agents' transcripts. The remaining three categories — conversational flow management, clarification, and prompts for specificity — were more commonly exhibited by the Wizard agent. We elaborate on these categories below.

LLM & Wizard Agents – Probing for More Detail: While the LLM agent added fewer questions than the Wizard agent, it was equally adept at following up on responses that lacked informativeness and depth. This was particularly true when participants gave generic and brief responses such as "sometimes" and "recently". The LLM agent often overcame these deficiencies by encouraging participants to provide more information, emphasizing the utility of the information for the participant's impending consultation.

> **Q14 – LLM agent:** `Understood. Do you have a personal or family history of mental health conditions?`
> **P41:** `Yes`
> **Q14 – LLM agent:** `Could you please provide a bit more detail? Your privacy is respected and this information helps us in providing a complete care for you.`
> **P41:** `Depression, anxiety, and BPD`

Wizard Agent – Managing Conversation Flow: The Wizard agent was more adept at recognizing situations when participants mentioned multiple concerns. In these situations, the agent was able to guide the conversation in a way that allowed participants to address each concern separately. This capability enhanced the clarity of participants' responses, allowing them to focus on each of their concerns more precisely. The following example illustrates the Wizard agent using follow-up questions to this effect when probing P33's symptoms:

> **Q1 - Wizard agent:** `What is the reason for your visit today?`
> **P33:** `Hello, the reason for my visit today is skin issues like ingrown hair and excess of hair. Also, sensitivity to noise and headache`
> **Q2 - Wizard agent:** `I am sorry to hear about it. Let's go through them one by one. First, tell me about your hair issues. Where on the body are the ingrown hairs? And are you experiencing excess hair growth all over the body?`
> **P33:** `The ingrown hair is mostly on the legs. The excess hair is on the breasts, near the hips and maybe legs?`
> **Q2 Follow up – Wizard agent:** `Thank you for this information. Now tell me more about the light and noise sensitivity please?`
> **P33:** `It is just noise issues. It is like the noise [that is] louder that disturbed [me] a lot. It is like the noise is too near my ears, and gives me a headache`

Wizard Agent – Clarifying Participant Response: The Wizard agent also used their medical knowledge to determine when information about participants' medication history needed clarification. As illustrated in the transcript below, the Wizard agent guessed the type of medication P30 was referencing and asked a follow-up question to confirm this assumption, which removed the ambiguity from the initial response.

> **Q8 – Wizard agent:** `Are you currently taking any medications? This also includes over-the-counter medications and herbal medications.`
> **P30:** `yes - the pill`
> **Q8 Follow up – Wizard agent:** `Does this mean oral contraceptive pill.`
> **P30:** `yes`
> **Q8 Follow up – Wizard agent:** `Since how long have you been taking it?`
> **P30:** `about two years`

Wizard Agent – Prompting for More Specific Information: We observed that 26 out of 48 (54.2%) of the follow-up questions from the Wizard agent happened even when the initial response was satisfactory. As an example, P25 provided an adequate response to a question about their symptoms, but the Wizard still prompted further to inquire about when the symptoms were most prevalent.

> **Q2 — Wizard agent:** `What is the reason for your visit today?`
> **P25:** `Lower back pain`
> **Q2 — Wizard agent:** `Sorry to hear about that. What symptoms are you experiencing?`
> **P25:** `severe pain when I stand for a while and I feel hard to do the regular day-to-day activities`
> **Q2 Follow up — Wizard agent:** `That seems really bothersome. Does it always come when you are standing? or does it bother you at rest as well?`
> **P25:** `bothers when only when I am standing and as soon as I wake up`

4.3.4 Patterns of Lacked Follow-up Questions by the Dynamic Agents.

While both of the dynamic agents were generally effective in asking follow-up questions that elicited more satisfactory participant responses, there were occasions when they failed to ask follow-up questions despite receiving an unsatisfactory initial response. We report the trends in these scenarios below:

LLM Agent — Instances of No Follow-up: Unlike the Wizard agent, the LLM agent often struggled with cases when participants presented with multiple symptoms. The LLM agent typically assumed that all symptoms were related to the same medical concern and, therefore, rarely asked clarifying follow-up questions to investigate the issues separately. This approach led to ambiguous responses regarding discomfort levels and symptom timelines.

> **Q1 – LLM agent:** `What is the reason for your visit today?`
> **P3:** `Brown patches on the face and after car mishap I lose balance. Therefore meeting doctor`
> **Q2 – LLM agent:** `I'm sorry to hear that. Let's gather some more information about your symptoms. What other symptoms are you`

experiencing?"
P3: Due to fibromyalgia, always body pain, depression and anxiety.
Q3 – LLM agent: Thank you for providing that information. How would you rate the discomfort these symptoms are causing you on a scale of 1-10?
P3: 6
Q4 – LLM agent: How long have you been experiencing these symptoms, specifically the brown patches on your face and balance issues?
P3: 2 years

We also observed inconsistencies in whether the LLM agent asked follow-up questions for similar sets of response deficiencies. The following example bears many similarities to the one above, as P31's initial response was lacking in both depth and informativeness. Instead of probing the participant for more in this situation, however, the LLM agent moved on to the next question.

Q13 – LLM agent: Do you have a personal or family history of mental health conditions?
P31: Yes.
Q13 – LLM agent: Thank you for sharing that information. Do you have anything else you want to discuss about your medical history?
P31: no

Wizard Agent — Instances of No Follow-up: Despite the Wizard agent's experience as a trained medical professional, there were still several instances when they did not follow up on unsatisfactory responses according to our conversational metrics. We observed that the Wizard agent's decision to ask follow-up questions appeared to depend on the perceived relevance of the information to the visit. In the example that follows, the Wizard agent did not attempt to clarify the ambiguity surrounding P2's family medical history. Perhaps also sensing that the participant did not want to go into detail, the Wizard decided to move on to the next question.

Q1 – Wizard agent: How can I help you today?
P2: I currently have unknown swelling within the jaw region.
...
Q11 – Wizard agent: Is there any medical conditions that run in your family?
P2: Too many to list
Q12 – Wizard agent: Alright. Have you had any hospitalizations?
P12: no

5 Discussion

In our discussion, we first relate our findings back to our initial research questions. As we revisit these questions, we describe potential design recommendations for future information-gathering chatbots. We then discuss how our findings might apply in other domains beyond clinic pre-consultation. Finally, we discuss the limitations of our work and opportunities for future investigation.

5.1 RQ1: Importance of Dynamic Question Generation

We found that a dynamic and adaptive question-asking agent, whether human or LLM-driven, outperforms a static one in eliciting satisfactory responses from users. The following subsections reflect on potential explanations for these observations.

5.1.1 Initial Question Structure. We observed that some questions given to the agents received relatively high rates of satisfactory responses, even with the Questionnaire agent. These questions were typically closed-ended, requiring simple yes/no answers or numerical ratings along a scale. Closed-ended questions have the advantage of being objective and easier to answer than their open-ended counterparts, but they can also prematurely narrow the scope of the conversation [114]. This is often problematic in the medical domain, as patients may struggle to express their concerns through a numeric rating or predefined options [50]. Robinson and Heritage [84] also found that patients do not appreciate conversations that only involve closed-ended questions since they feel it limits their sense of agency over their health concerns. This is why physicians often use a combination of open-ended and closed-ended questions [84, 87].

5.1.2 Changing the Line of Questioning. Even without follow-up questions, we found that the LLM and Wizard agents were able to gather more satisfactory responses from participants. This result may be partly attributed to the flexibility the agents were given to modify their question messages based on participants' responses. We observed several instances where these dynamic agents skipped and reordered questions, limiting repeated questions already answered in earlier conversation turns and grouping related questions based on participants' responses. These findings align with previous research suggesting that streamlined conversations with chatbots can lead to increased user engagement and self-disclosure [18, 109]. In the medical context, heightened engagement is particularly important because patients are often required to provide a wide range of information [12]. Preparing for a clinical consultation can be stressful and anxiety-inducing, as patients may worry about remembering critical details or being misunderstood [11]. Therefore, gathering information in a way that adapts to the natural flow of conversation can alleviate some of the cognitive load on patients, allowing them to focus on providing accurate and complete information.

5.1.3 Rewording with Acknowledgment. Another factor that may have led to better response quality without follow-up questions was wording changes made by the dynamic agents. We previously observed that our LLM agent frequently integrates phrases that repeat or acknowledge what patients had previously told it [57]. This form of verbal mimicry, sometimes referred to as an "echo effect", has been shown to scaffold rapport and comfort in human conversations [51], including clinical consultation [5, 36, 58]. The same may have happened in our study, ensuring that the conversation stayed on topic and that new information was being gathered with each turn.

5.2 RQ2: Overcoming Response Deficiencies with Dynamic Follow-Ups

While we observed that both dynamic agents were capable of adapting to participants' responses, the LLM agent was not nearly as proactive in asking follow-up questions as the Wizard agent, particularly when it came to lacking clarity. In the following subsections, we propose several causes for these differences and potential design implications for future conversational agents.

5.2.1 Adding Follow-up Questions. Even though the dynamic agents exhibited different tendencies in asking follow-up questions, these questions were almost always successful in overcoming unsatisfactory responses when they were asked. This confirmed our hypothesis that giving a conversation agent the ability to follow up on messages enhances the quality of information gathered from the user. Similar to echoing participants' previous responses, asking relevant follow-up questions is only possible when the agent adapts to what the other person is saying. This affordance leads to a more personalized and responsive interaction, fostering a sense of being understood.

5.2.2 Handling Ambiguous Language. Human conversations frequently involve subtle cues such as tone, phrasing, or context to make conversations more efficient [53]. Missing these cues can result in misunderstandings [29], so humans develop the ability to identify ambiguity in language and ask clarifying questions when needed. LLMs, on the other hand, are often trained on curated and vetted data [19], so they are not explicitly trained to deal with ambiguous content. This problem is exacerbated by the fact that ambiguity can be specific to domains including medicine [32, 35, 96].

For vague terms that do not require domain knowledge (e.g., "sometimes", "recently"), we observed that follow-up questions by the dynamic agents help provide more specificity. At the same time, we observed that the LLM agent was not always consistent in when it asked for clarification. This inconsistency may be due to some combination of the LLM's temperature setting, which regulates the creativity and randomness in its outputs, and the context accumulated from earlier conversation turns. One way to predispose a chatbot to ask more follow-up questions is by using a fine-tuning technique called alignment with perceived ambiguity (APA) [46]. This approach encourages the LLM to make its own assessments of perceived ambiguity and the amount of information gain that can be had from disambiguation. Even though the prompt given to our LLM agent had an explicit statement to clarify vague responses (*"You should follow up on questions whenever the response given by the user is vague"*), encoding this behavior at the level of model training and alignment may prove more fruitful.

While the aforementioned techniques can help an LLM detect and disambiguate general language, domain-specific training would inevitably be beneficial to account for terminology that requires expert knowledge (e.g., "the pill"). The LLM in our study used a generic GPT-4 model rather than a model like Med-PaLM [93] trained specifically for medical applications. We made this choice to ensure that the LLM agent had sufficient capabilities to lead an engaging conversation with participants. Patients typically do not explain their medical concerns or history using domain-specific language, so there were not any concerns about having a model that could not understand expert vocabulary. Nevertheless, future designs for pre-consultation chatbots could consider incorporating medically trained LLMs to help identify when topics require disambiguation.

5.2.3 Handling Multiple Concerns Simultaneously. The LLM agent's occasional inability to track multiple medical topics at the same time may also be partly attributed to the fact that it did not include the domain-specific knowledge that would have helped it decide when topics should be separated. However, there are domain-agnostic techniques that could be used to remedy this solution. Specific prompting can be added to help LLMs account for situations when patients present with multiple medical issues, encouraging them to separate out a new line of questioning for each one. Chain-of-thought prompting helps LLMs break down complex conversations [66, 113], while retrieval-augmented generation helps LLMs retrieve and incorporate earlier information [56]. However, both approaches have limitations: chain-of-thought can miss subtle details, and retrieval-augmented generation may introduce inconsistencies if the retrieved data does not align with the ongoing dialogue. To address these issues, future designs could explore hybrid techniques like retrieval-augmented thought (RAT) [105] to dynamically track and integrate user inputs across extended conversations.

5.2.4 Scoping the Goal of the Conversation. While Grice's maxims provided us with metrics for evaluating dialogue at every conversation turn, they overlook some subtleties in the context of the conversation at a higher level. In our analysis of the follow-up questions, we observed that the Wizard agent's decision to initiate further inquiry was often guided by the broader context of participants' expectations and needs. This awareness enabled the Wizard agent to prompt for additional information when it perceived participants had more to share, as evidenced by the follow-up questions about symptoms even when the initial responses were already satisfactory. This judgment may have also influenced the Wizard agent's decision to forgo asking follow-up questions to unsatisfactory responses that seemed less pertinent, instead prioritizing keeping the conversation on track.

Clinical pre-consultation questionnaires are designed with the sole goal of collecting answers to a pre-defined list of questions. We argue that intake nurses who conduct pre-consultation operate on a much broader and patient-centered goal: gathering all the information patients wish to express and may not know to express within a reasonable amount of time [97]. This skillful balance between proactive probing and efficient conversation is a core aspect of patient-centered care [26, 72]. While the LLM agent in our study was proficient at keeping the conversation on track, the balance was not always in favor of encouraging patients to speak more about their concerns and priorities. As our work strives to make pre-consultation chatbots more along the lines of intake nurses, we believe that chatbots in this space should do more than react to patients' responses; they should actively listen and infer relevant topics to give patients more opportunities to share what is on their mind [37, 43, 55, 83]. One way to do this is by encouraging chatbots to consider pertinent negatives, which are symptoms that are expected but absent [75]. Although differential diagnosis is explicitly not a goal of pre-consultation, provoking questions related to pertinent negatives may provide one avenue of operationalizing probing beyond the one-size-fits-all questions of pre-consultation. This affordance would help patients not only experience more efficient consultations but also become better advocates of their own care.

5.3 Generalizing Beyond Clinical Pre-consultation

While some of our design considerations are specific to clinical pre-consultation, our analysis method and many of our findings can be generalized. The conversation metrics we developed — clarity, depth, informativeness, and relevance — can be used to assess the information-gathering quality of many existing health chatbots. Examples include triaging chatbots that prioritize patient cases based on information gathered about the severity of patient concerns [54, 62, 111], post-operative chatbots that monitor recovery and gather feedback following surgery [15, 24, 110], and mental health chatbots that provide companionship and deliver CBT treatments over extended periods [3, 23, 79]. By applying these metrics, developers and researchers can systematically improve the information-gathering capabilities of their chatbots before focusing their attention on task execution (e.g., patient ranking and personalized feedback).

Furthermore, our findings on how dynamic agents adapt and follow up on unsatisfactory user responses can be extended to areas beyond healthcare. In customer service, for instance, chatbots are often criticized for failing to understand customer needs [31]. Designing chatbots that ask follow-up questions to clarify and confirm their understanding of customers' concerns can improve service interactions and increase customer satisfaction. Similarly, in technical support, chatbots frequently struggle to identify and address multiple interconnected problems accurately [82]. Equipping chatbots with the ability to ask follow-up questions to untangle complex queries can help manage conversations and address issues one at a time. Finance [73] and legal consultation [81] are other areas with a growing chatbot presence. These use cases often require gathering detailed and precise information to provide accurate and tailored advice. Designing chatbots that can effectively prompt and probe for more specific and relevant details ensures they address user needs comprehensively and minimize miscommunication errors. These examples show that, regardless of the domain, adaptive information gathering and effective conversation management are crucial for improving chatbot utility. By applying these principles, designers in various industries can build more efficient, user-friendly systems that offer meaningful, productive interactions.

5.4 Privacy and Ethical Considerations

Unlike health chatbots designed to triage or diagnose patients, pre-consultation chatbots delegate decision-making to trained medical professionals, but this does not absolve pre-consultation chatbots from privacy and ethics considerations. In many parts of the world, doctor-patient confidentiality protects patients' privacy when they discuss sensitive topics with physicians [17]. This principle applies to intake nurses as well, but since patients know they will eventually consult a physician, they may avoid extraneous disclosure if they deem the information too sensitive or less relevant to their medical concerns. Intake questionnaires and pre-consultation chatbots must also adhere to strict privacy regulations, enforcing that information is only accessible to healthcare providers who need it. Still, the knowledge of having information logged electronically may discourage some patients from sensitive disclosure, especially if the data goes through a third-party software developer [64, 67]. While all of these factors may deter disclosure, we previously observed that some patients might actually feel more comfortable disclosing sensitive information with a chatbot because they do not have to worry about being directly observed by a human who may judge their responses [57].

One way of addressing these considerations around privacy and confidential disclosure is by carefully considering how information gathered by the chatbot is shared and stored with other entities. In our study, a medical resident summarized each encounter with the chatbot before handing off the information to the patient's assigned physician. These summaries provided a couple of benefits over complete transcripts. It was less time-consuming for physicians to read the summaries between their consultations, and the summaries were concise enough that physicians could use them to document cases in their electronic medical records (EMR). Our study design also allowed physicians to decide what information to document and what to omit, which may not be the case for deployments that are fully integrated with EMR systems. Therefore, future deployments could consider designing summaries for different purposes. For instance, a chatbot could provide a complete summary to physicians but abstract or redact sensitive information in the one stored in the EMR; this arrangement would still give physicians the power to document sensitive information provided they deem it important and receive patient consent. Even if only a single summary is generated, this step in the workflow yields a prime opportunity for patients to review, modify, redact, and consent to the information that is being shared and documented.

5.5 Limitations and Future Work

While both the LLM and Wizard agents were more successful in eliciting satisfactory responses from participants compared to the Questionnaire agent once follow-up questions were included in the analysis, the other tests and comparisons did not lead to statistically significant results. This may be due to our modest study cohort of 45 patients, so a larger sample size may strengthen the claims that can be made about our observed trends. The lack of a strong result may also be attributed to the diversity of medical concerns participants presented in our study. Some participants had simple concerns like physical injuries, in which case the Questionnaire agent was likely just as effective as the dynamic agents in performing intake. Other participants had more complicated or multi-faceted concerns, providing more opportunities for the dynamic agents to excel. We found it difficult to consolidate these cases into cohesive categories, but future work could consider repeating our study in a setting other than a walk-in clinic where patients present with complex but homogeneous concerns.

Another limitation of our analysis lies in how we categorized the information gathered. We annotated transcripts with binary codes corresponding to our conversation metrics, and we examined these codes at the level of conversation turns. This procedure provided consistency in how we compared transcripts across patients with varying levels of verbosity or case complexity. However, we may have lost some of the nuances in the broader context of the consultation process, namely physicians' satisfaction with the resulting information. We envision that future pre-consultation chatbots will include a mechanism for automatically summarizing the conversation transcripts into an easily digestible format for physicians. With enough instrumentation, it may be easy to track whether elements from these summaries are incorporated into patients' electronic medical records. This workflow from pre-consultation to medical documentation provides ample space for exploration.

6 Conclusion

In this work, we examined the extent to which a pre-consultation chatbot is able to elicit information from patients relative to a medical professional and a static questionnaire. We observed that both dynamic agents — the Wizard and LLM — adapted their language and used follow-up questions to gather more satisfactory responses according to the metrics we adapted from Grice's maxims: clarity, depth, informativeness, and relevance. Nevertheless, we found that the wording of the initial question and the deficiencies in participants' initial responses influenced when the agents decided to follow up further. Although these observations are grounded in a clinical context, the insights were drawn using an evaluation approach that dissected the information-gathering process. Our work highlights many design implications for chatbots across domains ranging from customer service to public health, including the importance of having explicit features that encourage follow-up questions and account for concurrent topics.

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] Bhashithe Abeysinghe and Ruhan Circi. 2024. The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches. *arXiv preprint arXiv:2406.03339* (2024).

[3] Farah Ahmad, Wendy Lou, Yogendra Shakya, Liane Ginsburg, Peggy T Ng, Meb Rashid, Serban Dinca-Panaitescu, Cliff Ledwos, and Kwame McKenzie. 2017. Preconsult interactive computer-assisted client assessment survey for common mental disorders in a community health centre: a randomized controlled trial. *Canadian Medical Association Open Access Journal* 5, 1 (2017), E190–E197.

[4] 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.

[5] Daniel Angus, Bernadette Watson, Andrew Smith, Cindy Gallois, and Janet Wiles. 2012. Visualising conversation structure across time: Insights into effective doctor-patient consultations. *PloS one* 7, 6 (2012), e38014.

[6] 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.

[7] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (Eds.). Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909

[8] 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.

[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. doi:10.1145/3313831.3376718

[10] Niv Ben-Shabat, Gal Sharvit, Ben Meimis, Daniel Ben Joya, Ariel Sloma, David Kiderman, Aviv Shabat, Avishai M Tsur, Abdulla Watad, and Howard Amital. 2022. Assessing data gathering of chatbot based symptom checkers-a clinical vignettes study. *International Journal of Medical Informatics* 168 (2022), 104897.

[11] Jozien M Bensing, William Verheul, and AM Van Dulmen. 2008. Patient anxiety in the medical encounter: a study of verbal and nonverbal communication in general practice. *Health education* 108, 5 (2008), 373–383.

[12] Parvathy Bowes, Fiona Stevenson, Sanjiv Ahluwalia, and Elizabeth Murray. 2012. 'I need her to be a doctor': patients' experiences of presenting health information from the internet in GP consultations. *British Journal of General Practice* 62, 604 (2012), e732–e738.

[13] 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.

[14] Wanling Cai, Yucheng Jin, and Li Chen. 2022. Task-oriented user evaluation on critiquing-based recommendation chatbots. *IEEE Transactions on Human-Machine Systems* 52, 3 (2022), 354–366.

[15] Kevin Campbell, Philip Louie, Brett Levine, and Jeremy Gililland. 2020. Using patient engagement platforms in the postoperative management of patients. *Current reviews in musculoskeletal medicine* 13 (2020), 479–484.

[16] Jacky Casas, Marc-Olivier Tricot, Omar Abou Khaled, Elena Mugellini, and Philippe Cudré-Mauroux. 2021. Trends & Methods in Chatbot Evaluation. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 280–286. doi:10.1145/3395035.3425319

[17] William A Challener Jr. 1949. The Doctor-Patient Relationship and the Right to Privacy. *U. Pitt. L. Rev.* 11 (1949), 624.

[18] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758. doi:10.1080/10447318.2020.1841438

[19] Xu Chu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang. 2016. Data Cleaning: Overview and Emerging Challenges. In *Proceedings of the 2016 International Conference on Management of Data* (San Francisco, California, USA) *(SIGMOD '16)*. Association for Computing Machinery, New York, NY, USA, 2201–2206. doi:10.1145/2882903.2912574

[20] Clint Cuffy, Nao Hagiwara, Scott Vrana, and Bridget T McInnes. 2020. Measuring the quality of patient–physician communication. *Journal of biomedical informatics* 112 (2020), 103589.

[21] Kate Daley, Ines Hungerbuehler, Kate Cavanagh, Heloísa Garcia Claro, Paul Alan Swinton, and Michael Kapps. 2020. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. *Frontiers in digital health* 2 (2020), 576361.

[22] 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.

[23] Kerstin Denecke, Alaa Abd-Alrazaq, and Mowafa Househ. 2021. Artificial intelligence for chatbots in mental health: opportunities and challenges. *Multiple perspectives on artificial intelligence in healthcare: Opportunities and challenges* (2021), 115–128.

[24] Tim Dwyer, Graeme Hoit, David Burns, James Higgins, Justin Chang, Daniel Whelan, Irene Kiroplis, and Jaskarndip Chahal. 2023. Use of an artificial intelligence conversational agent (chatbot) for hip arthroscopy patients following surgery. *Arthroscopy, Sports Medicine, and Rehabilitation* 5, 2 (2023), e495–e505.

[25] 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.

[26] Ronald M Epstein and Richard L Street. 2011. The values and value of patient-centered care. 100–103 pages.

[27] 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.

[28] 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.

[29] Caitlin M Fausey and Lera Boroditsky. 2010. Subtle linguistic cues influence perceived blame and financial liability. *Psychonomic bulletin & review* 17, 5 (2010), 644–650.

[30] Giuseppe Fenza, Francesco Orciuoli, Angela Peduto, and Alberto Postiglione. 2023. Healthcare Conversational Agents: Chatbot for Improving Patient-Reported Outcomes. In *Advanced Information Networking and Applications*, Leonard Barolli (Ed.). Springer International Publishing, Cham, 137–148.

[31] Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 1, 9 pages. doi:10.1145/3342775.3342784

[32] Shiphra Ginsburg, Glenn Regehr, Lorelei Lingard, and Kevin W Eva. 2015. Reading between the lines: faculty interpretations of narrative evaluation comments. *Medical Education* 49, 3 (2015), 296–306.

[33] Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics* 3 (1975), 43–58.

[34] 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.

[35] Benjamin M Gyori, Charles Tapley Hoyt, and Albert Steppi. 2022. Gilda: biomedical entity text normalization with machine-learned disambiguation as a service. *Bioinformatics Advances* 2, 1 (2022), vbac034.

[36] Steven R Hahn, David S Friedman, Harry A Quigley, Sameer Kotak, Elizabeth Kim, Meaghan Onofrey, Corey Eagan, and Jack Mardekian. 2010. Effect of patient-centered communication training on discussion and detection of nonadherence in glaucoma. *Ophthalmology* 117, 7 (2010), 1339–1347.

[37] M Jawad Hashim. 2017. Patient-centered communication: basic skills. *American family physician* 95, 1 (2017), 29–34.

[38] Michael A. Hedderich, Natalie N. Bazarova, Wenting Zou, Ryun Shim, Xinda Ma, and Qian Yang. 2024. A Piece of Theatre: Investigating How Teachers Design LLM Chatbots to Assist Adolescent Cyberbullying Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 668, 17 pages. doi:10.1145/3613904.3642379

[39] Baptiste Jacquet, Alexandre Hullin, Jean Baratgin, and Frank Jamet. 2019. The impact of the gricean maxims of quality, quantity and manner in chatbots. In *2019 international conference on information and digital technologies (idt)*. IEEE, 180–189.

[40] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 designing interactive systems conference*. 895–906.

[41] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[42] Prathyusha Jwalapuram. 2017. Evaluating dialogs based on Grice's maxims. In *Proceedings of the Student Research Workshop associated with RANLP*. 17–24.

[43] Lianna Karp. 2015. Can empathy be taught? Reflections from a medical student active-listening workshop. *Rhode Island medical journal* 98, 6 (2015).

[44] 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.

[45] Anjali Khurana, Hariharan Subramonyam, and Parmit K Chilana. 2024. Why and when llm-based assistants can go wrong: Investigating the effectiveness of prompt-based interactions for software help-seeking. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 288–303.

[46] Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024. Aligning Language Models to Explicitly Handle Ambiguity. *arXiv preprint arXiv:2404.11972* (2024).

[47] 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. doi:10.1145/3290605.3300316

[48] 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.

[49] 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/

[50] Jon A Krosnick and Duane F Alwin. 1987. An evaluation of a cognitive theory of response-order effects in survey measurement. *Public opinion quarterly* 51, 2 (1987), 201–219.

[51] Wojciech Kulesza, Dariusz Dolinski, Avia Huisman, and Robert Majewski. 2014. The echo effect: The power of verbal mimicry to influence prosocial behavior. *Journal of Language and Social Psychology* 33, 2 (2014), 183–201.

[52] 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.

[53] Marta Kutas and Steven A Hillyard. 1980. Reading between the lines: Event-related brain potentials during natural sentence processing. *Brain and language* 11, 2 (1980), 354–373.

[54] Lucinda Lai, Kelley A Wittbold, Farah Z Dadabhoy, Rintaro Sato, Adam B Landman, Lee H Schwamm, Shuhan He, Rajesh Patel, Nancy Wei, Gianna Zuccotti, et al. 2020. Digital triage: novel strategies for population health management in response to the COVID-19 pandemic. In *Healthcare*, Vol. 8. Elsevier, 100493.

[55] Forrest Lang, Michael R Floyd, and Kathleen L Beine. 2000. Clues to patients' explanations and concerns about their illnesses: A call for active listening. *Archives of family medicine* 9, 3 (2000), 222.

[56] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[57] Brenna Li, Ofek Gross, Noah Crampton, Mamta Kapoor, Saba Tauseef, Mohit Jain, Khai N. Truong, and Alex Mariakakis. 2024. Beyond the Waiting Room: Patient's Perspectives on the Conversational Nuances of Pre-Consultation Chatbots. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 438, 24 pages. doi:10.1145/3613904.3641913

[58] 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. doi:10.1145/3544548.3581014

[59] Brenna Li, Amy Wang, Patricia Strachan, Julie Anne Séguin, Sami Lachgar, Karyn C Schroeder, Mathias S Fleck, Renee Wong, Alan Karthikesalingam, Vivek Natarajan, Yossi Matias, Greg S Corrado, Dale Webster, Yun Liu, Naama Hammel, Rory Sayres, Christopher Semturs, and Mike Schaekermann. 2024. Conversational AI in health: Design considerations from a Wizard-of-Oz dermatology case study with users, clinicians and a medical LLM. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 88, 10 pages. doi:10.1145/3613905.3651891

[60] 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. doi:10.1145/3571884.3604305

[61] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[62] Riley J Lyons, Sruthi R Arepalli, Ollya Fromal, Jinho D Choi, and Nieraj Jain. 2024. Artificial intelligence chatbot performance in triage of ophthalmic conditions. *Canadian Journal of Ophthalmology* 59, 4 (2024), e301–e308.

[63] 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.

[64] Mason Marks and Claudia E Haupt. 2023. AI chatbots, health privacy, and challenges to HIPAA compliance. *Jama* (2023).

[65] Wari Maroengsit, Thanarath Piyakulpinyo, Korawat Phonyiam, Suporn Pongnumkul, Pimwadee Chaovalit, and Thanaruk Theeramunkong. 2019. A survey on evaluation methods for chatbots. In *Proceedings of the 2019 7th International conference on information and education technology*. 111–119.

[66] Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Pajaree Krisanapan, Yeshwanter Radhakrishnan, and Wisit Cheungpasitporn. 2024. Chain of thought utilization in large language models and application in nephrology. *Medicina* 60, 1 (2024), 148.

[67] Tala Mirzaei, Leila Amini, and Pouyan Esmaeilzadeh. 2024. Clinician voices on ethics of LLM integration in healthcare: a thematic analysis of ethical concerns and implications. *BMC Medical Informatics and Decision Making* 24, 1 (2024), 250.

[68] 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).

[69] 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.

[70] Krishna Kumar Nirala, Nikhil Kumar Singh, and Vinay Shivshanker Purani. 2022. A survey on providing customer and public administration based services using AI: chatbot. *Multimedia Tools and Applications* 81, 16 (2022), 22215–22246.

[71] 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).

[72] Julian Oates, W Wayne Weston, John Jordan, et al. 2000. The impact of patient-centered care on outcomes. *Fam Pract* 49, 9 (2000), 796–804.

[73] Takuma Okuda and Sanae Shoda. 2018. AI-based chatbot service for financial industry. *Fujitsu Scientific and Technical Journal* 54, 2 (2018), 4–8.

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

[75] Clifford D Packer and Clifford D Packer. 2019. Pertinent Positives and Negatives. *Presenting Your Case: A Concise Guide for Medical Students* (2019), 57–71.

[76] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. doi:10.3115/1073083.1073135

[77] Hyanghee Park and Daehwan Ahn. 2024. The Promise and Peril of ChatGPT in Higher Education: Opportunities, Challenges, and Design Implications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.

[78] 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.

[79] Kay T Pham, Amir Nabizadeh, and Salih Selek. 2022. Artificial intelligence and chatbots in psychiatry. *Psychiatric Quarterly* 93, 1 (2022), 249–253.

[80] 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.

[81] Marc Queudot, Éric Charton, and Marie-Jean Meurs. 2020. Improving access to justice with legal chatbots. *Stats* 3, 3 (2020), 356–375.

[82] Miguel Angel Quiroz Martinez, Segundo Estuardo Mayorga Plua, Monica Daniela Gomez Rios, Maikel Yelandi Leyva Vázquez, and Daniel Humberto Plua Moran. 2020. Chatbot for Technical Support, Analysis of Critical Success Factors Using Fuzzy Cognitive Maps. In *International Conference on Applied Technologies*. Springer, 363–375.

[83] Kathryn Robertson. 2005. Active listening: more than just paying attention. *Australian family physician* 34, 12 (2005).

[84] Jeffrey D Robinson and John Heritage. 2006. Physicians' opening questions and patients' satisfaction. *Patient education and counseling* 60, 3 (2006), 279–285.

[85] 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).

[86] 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.

[87] Debra Roter and Susan Larson. 2002. The Roter interaction analysis system (RIAS): utility and flexibility for analysis of medical interactions. *Patient education and counseling* 46, 4 (2002), 243–251.

[88] Ryan M Schuetzler, G Mark Grimes, and Justin Scott Giboney. 2020. The impact of chatbot conversational skill on engagement and perceived humanness. *Journal of Management Information Systems* 37, 3 (2020), 875–900.

[89] João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. ChatEval: A Tool for Chatbot Evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 60–65. doi:10.18653/v1/N19-4011

[90] João Sedoc and Lyle Ungar. 2020. Item response theory for efficient human evaluation of chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. 21–33.

[91] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696* (2020).

[92] Vidya Setlur and Melanie Tory. 2022. How do you converse with an analytical chatbot? revisiting gricean maxims for designing analytical conversational behavior. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–17.

[93] 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.

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

[95] 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.

[96] Mark Stevenson and Yikun Guo. 2010. Disambiguation in the biomedical domain: the role of ambiguity type. *Journal of biomedical informatics* 43, 6 (2010), 972–981.

[97] Moira A Stewart. 1995. Effective physician-patient communication and health outcomes: a review. *CMAJ: Canadian medical association journal* 152, 9 (1995), 1423.

[98] Tianjun Sun, Alan R Dennis, Michelle Zhou, et al. 2024. Perception is reality? Understanding user perceptions of chatbot-inferred versus self-reported personality traits. *Computers in Human Behavior: Artificial Humans* 2, 1 (2024), 100057.

[99] 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.

[100] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.

[101] 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.

[102] 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.

[103] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. 2018. On evaluating and comparing open domain dialog systems. *arXiv preprint arXiv:1801.03625* (2018).

[104] 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.

[105] Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. 2024. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313* (2024).

[106] 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]

[107] Lee Wilson and Mariana Marasoiu. 2022. The development and use of chatbots in public health: scoping review. *JMIR human factors* 9, 4 (2022), e35882.

[108] Ziang Xiao, Susu Zhang, Vivian Lai, and Q Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. *arXiv preprint arXiv:2305.14889* (2023).

[109] 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 Trans. Comput.-Hum. Interact.* 27, 3, Article 15 (jun 2020), 37 pages. doi:10.1145/3381804

[110] Yi Xie, Ishith Seth, Warren M Rozen, and David J Hunter-Smith. 2023. Evaluation of the artificial intelligence chatbot on breast reconstruction and its efficacy in surgical research: a case study. *Aesthetic Plastic Surgery* 47, 6 (2023), 2360–2369.

[111] Yue You, Chun-Hua Tsai, Yao Li, Fenglong Ma, Christopher Heron, and Xinning Gui. 2023. Beyond Self-diagnosis: How a Chatbot-based Symptom Checker

Should Respond. *ACM Trans. Comput.-Hum. Interact.* 30, 4, Article 64 (Sept. 2023), 44 pages. doi:10.1145/3589959

[112] 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.

[113] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024. LLM as a Mastermind: A Survey of Strategic Reasoning with Large Language Models. *arXiv preprint arXiv:2404.01230* (2024).

[114] Ronggang Zhou, Xiaorui Wang, Leyuan Zhang, and Haiyan Guo. 2017. Who tends to answer open-ended questions in an e-service survey? The contribution of closed-ended answers. *Behaviour & Information Technology* 36, 12 (2017), 1274–1284.