Scaffolded Turns and Logical Conversations: Designing Humanized LLM-Powered Conversational Agents for Hospital Admission Interviews

Dingdong Liu
The Hong Kong University of Science and Technology
Hong Kong, China
dliuak@connect.ust.hk

Yujing Zhang
KTH Royal Institute of Technology
Stockholm, Sweden
yujingzh@kth.se

Bolin Zhao
The Hong Kong University of Science and Technology
Hong Kong SAR, China
bzhaoan@connect.ust.hk

Shuai Ma
The Hong Kong University of Science and Technology
Hong Kong, China
shuai.ma@connect.ust.hk

Chuhan Shi
Southeast University
Nanjing, China
chuhanshi@seu.edu.cn

Xiaojuan Ma[*]
The Hong Kong University of Science and Technology
Hong Kong, China
mxj@cse.ust.hk

Abstract

Hospital admission interviews are critical for patient care but strain nurses' capacity due to time constraints and staffing shortages. While LLM-powered conversational agents (CAs) offer automation potential, their rigid sequencing and lack of humanized communication skills risk misunderstandings and incomplete data capture. Through participatory design with clinicians and volunteers, we identified essential communication strategies and developed a novel CA that implements these strategies through: (1) dynamic topic management using graph-based conversation flows, and (2) context-aware scaffolding with few-shot prompt tuning. Technical evaluation on an admission interview dataset showed our system achieving performance comparable to or surpassing human-written ground truth, while outperforming prompt-engineered baselines. A between-subject study (N=44) demonstrated significantly improved user experience and data collection accuracy compared to existing solutions. We contribute a framework for humanizing medical CAs by translating clinician expertise into algorithmic strategies, alongside empirical insights for balancing efficiency and empathy in healthcare interactions, and considerations for generalizability.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**; **Empirical studies in interaction design**; **User studies**; **Participatory design**; • **Applied computing → Health care information systems**; • **Computing methodologies → Discourse, dialogue and pragmatics**.

Keywords

Conversational Agents, Clinical Communication, Hospital Admission Interview, Healthcare Automation, Scaffolded Dialogue, Participatory Design, Large Language Models

ACM Reference Format:
Dingdong Liu, Yujing Zhang, Bolin Zhao, Shuai Ma, Chuhan Shi, and Xiaojuan Ma. 2025. Scaffolded Turns and Logical Conversations: Designing Humanized LLM-Powered Conversational Agents for Hospital Admission Interviews. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3706598.3714196

1 Introduction

Conducting admission interviews is a crucial step in the hospital admission process. Typically, these interviews involve structured interactions between patients and care providers, guided by a standardized assessment form [28, 41, 64, 75]. The care provider, often a nurse, listens to the patient's narrative and captures critical medical information, including symptoms, medications, personal histories, and preferences, as answers and auxiliary notes to the questions [36, 49, 51]. Admission interviews are vital in reducing medication errors, building pre-understandings between patients and care providers, and creating a caring environment [23].

However, in hospital practices, the execution of admission interviews can be obstructed by nurses' limited availability due to manpower shortage. Among nurses' various responsibilities, admission interviews are particularly demanding, requiring substantial time commitment while offering little scheduling flexibility [19, 23, 43]. Nurses may need to rush interviews or handle frequent interruptions during interviews, leading to incomplete documentation and compromised patient satisfaction [49, 77]. While volunteer recruitment helps address immediate staffing needs, they are not always available. Volunteers' limited medical expertise can also compromise interview quality [42]. Similarly, self-reported questionnaires often yield imprecise information due to patients' misinterpretation of medical terminology or questions [63, 72]. Given the growing adoption of conversational technologies in healthcare, we propose

[*] Corresponding Author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714196

exploring CAs to automate the admission interview process. CA-assisted systems could potentially enhance healthcare efficiency by consistently capturing comprehensive information without fatigue, keeping the same level of expertise across all interviews. This approach has shown promise in various contexts [45, 47, 89, 91], including medical scenarios [3, 54, 86].

However, the admission interview scenario presents unique challenges revealed by our formative study with nurses and hospital volunteers. First, the interview is guided by an assessment form with medical jargon. Nurses must translate medical jargon into plain language while ensuring patient comprehension, such as explaining that a "thickener for dysphagia" refers to swallowing assistance supplements. Second, nurses must extract information from patients' narratives and translate it into the assessment form's predefined format. This task is challenging because patients may not follow the form's sequence when sharing information, often providing details that either address future questions or contradict previous responses, requiring careful tracking and cross-referencing throughout the interview. Previous research also highlights that patients require specific conversational support in medical interviews, including scaffolding for question comprehension [48, 59, 66] and topic management for logical coherence [25, 91]. While existing CA-assisted survey systems have attempted to address these needs, they remain limited by rigid question sequencing and basic scaffolding approaches [29, 37], often appearing inflexible and inauthentic to users.

Recent Large language model (LLM)-driven CAs have been proposed as interactive, highly expressive tools with the potential to convey information in a manner resembling that of medical professionals [38, 55, 56]. They could mitigate the flexibility and authenticity problem but faces challenges including limited long-term memory [30], insufficient active information seeking [55, 56], empathy showing [55], and potential hallucination issues [30]. These limitations particularly hindered their direct application in admission interviews. Given the contextual challenges identified via our formative study and the limitations of existing technology, we therefore raised the following research questions:

- **RQ1:** What are the key challenges and techniques in hospital admission interviews that need to be addressed by CAs?
- **RQ2:** How can these interview techniques be effectively implemented in an LLM-driven CA?
- **RQ3:** How effective is the developed system in conducting admission interviews compared to existing approaches?

To answer **RQ1**, we designed a formative study with clinicians and hospital volunteers to dissect the challenges in the aforementioned admission interview scenario. The study revealed three key techniques used by clinicians: scaffolding, topic management, and empathy with re-orientation. We also collected a corpus of how clinicians do scaffolding, present empathy and re-orient patients. To address **RQ2**, a CA-assisted survey system incorporating these techniques through graph-based conversation modeling and few-shot prompt tuning. To answer **RQ3**, we conducted a technical assessment using Gricean Maxims and a user study with 44 recently-recovered patients, evaluating communication quality, the CA's information quality, and users' subjective experience. The study workflow is shown in Figure 1.

Results show that our system achieved comparable or superior performance to human benchmarks at the turn level, consistently outperforming baseline approaches. However, limitations in conveying empathy and human touch remain, suggesting directions for future improvements. In summary, our main contribution are as follows:

- We systematically analyzed and documented the essential communication skills employed by healthcare professionals during hospital admission interviews, and collaboratively designed these competencies for implementation in conversational agents,
- We developed and validated a novel framework that equips conversational agents with aforementioned interview capabilities, conducting both turn-level and conversation-level evaluations through human participant studies, and
- We provided empirical insights into the impact of conversational agents on hospital admission workflows, along with design guidelines for deploying CAs across diverse patient interview contexts.

2 Related Works

2.1 Hospital Admission Interviews

Hospital admission interviews typically involve a narrative-based information collection process focusing on patients' medical history, current health status, and other relevant information. This process entails a conversation between a patient and a healthcare professional, usually a nurse, guided by a patient assessment form (PAF) and supplemented by physical examinations [36, 49, 51]. The collected information is then recorded in the Electronic Health Record (EHR) system [70]. This narrative-based approach offers several benefits. It embraces idiosyncratic experiences of illness [22, 46], provides a caring environment [64, 80], and empowers individuals with medical conditions in the face of dominant biomedical knowledge and language [41, 64, 75].

However, hospital admission interviews are time-consuming and labor-intensive [23], exacerbating the strain on an already overburdened healthcare profession [27]. Simply converting this narrative-based process into a structured questionnaire (and back) presents significant challenges. PAF forms contain numerous medical jargon and nuanced descriptions that are difficult to capture in a structured format [63]. Moreover, narrative elicitation is not merely a transition from traditional medical history-taking or a type of structured interview; it requires communication skills and strategies tailored to case-specific and personalized needs [66, 92]. Current research explores CAs to replicate the communication skills and strategies required for other medical tasks [17, 29, 92]. However, some systems rely on selecting responses from a pre-existing candidate pool [29], which constrains variability, while others use generic, unified responses [92], which have been criticized for lacking authenticity [37]. Several critical questions remain unanswered for developing effective, authentic, and adaptable automated systems for hospital admission interviews. First, how are these essential communication skills and strategies shown in this specific setting? Second, how to design CAs not only replicate these capabilities but dynamically generate contextualized responses to meet patients' needs?In this work, we proposed a novel approach to address these challenges

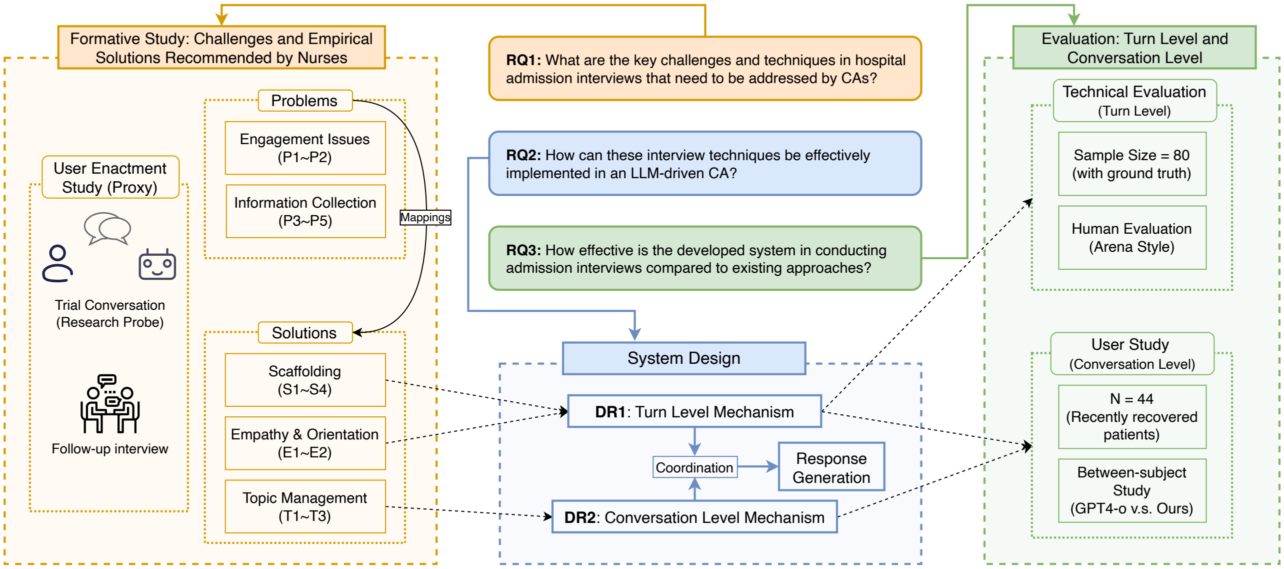

Formative Study: Challenges and Empirical Solutions Recommended by Nurses
User Enactment Study (Proxy)
Trial Conversation (Research Probe)
Follow-up interview
Problems
Engagement Issues (P1~P2)
Information Collection (P3~P5)
Mappings
Solutions
Scaffolding (S1~S4)
Empathy & Orientation (E1~E2)
Topic Management (T1~T3)
RQ1: What are the key challenges and techniques in hospital admission interviews that need to be addressed by CAs?
RQ2: How can these interview techniques be effectively implemented in an LLM-driven CA?
RQ3: How effective is the developed system in conducting admission interviews compared to existing approaches?
System Design
DR1: Turn Level Mechanism
Coordination
Response Generation
DR2: Conversation Level Mechanism
Evaluation: Turn Level and Conversation Level
Technical Evaluation (Turn Level)
Sample Size = 80 (with ground truth)
Human Evaluation (Arena Style)
User Study (Conversation Level)
N = 44 (Recently recovered patients)
Between-subject Study (GPT4-o v.s. Ours)

Figure 1: The workflow of our study. We first conducted a formative study to identify the challenges in the hospital admission interview scenario and the techniques clinicians use to address these challenges. We then designed a CA using the identified techniques and conducted a technical evaluation to assess the system's performance. Finally, we conducted a user study to trace the user's perception of the system.

by abstracting the essential communication skills through a set of challenge-strategies mappings and integrating them into an LLM to generate context-aware responses.

2.2 Conversational Agents for Interview Tasks

Respondent engagement in healthcare questionnaires is fundamental to ensuring adequate response rates for service and care quality evaluation. The heavy workload and time constraints of hospital staff call for automation of questionnaire and survey processes [23]. Conventional survey designs, such as digital questionnaires, are often perceived as dull and unengaging, resulting in negative respondent behaviors like "survey-taking fatigue" [91]. CAs, including chatbots, are proposed as a solution to provide more attractive and motivating survey completion experiences [86].

Compared to traditional web-based surveys, CA-powered surveys have demonstrated advantages in response rate, user engagement, and response quality due to natural conversation and interactive features [45, 86]. This trend is observed across various domains, including internet usage surveys [45], mental health assessments [54], social needs screenings [47], user satisfaction surveys [91], patient-reported outcome measures [86], and course evaluations [89]. The advantages of CA-powered surveys stem from several factors. They can reduce perceived completion time [86], provide context-sensitive questions and personalized framings [74, 90], proactively manage survey processes [91], and create a non-judgmental environment that encourages self-disclosure [54].

However, designing CAs for interview tasks presents challenges. In mental health assessments, form-based interactions lead to higher assessment credibility for closed-ended questions, which improves response quality in open-ended questions [37]. This suggests that CA design for patient interviews should balance proficiency with plain language and consider question types. Moreover, conversational information collection may require more follow-up questions, resulting in scattered information [91] and potential deviation from the questionnaire [59]. Thus, CA design for interview tasks must carefully balance natural conversation with structured questionnaires while considering the nature of the questions being asked.

2.3 Language Support Needed in Interview Tasks

Interview tasks require sophisticated language capabilities to ensure effective communication and data collection. In general contexts, active listening skills and the ability to ask clarification questions are essential [83, 90]. However, healthcare settings present unique challenges that demand even more advanced language support. These specialized contexts require additional capabilities to effectively scaffold conversations, addressing the complex and often sensitive nature of health-related discussions.

In NCD screening, for instance, reactive back-channeling encourages participants to continue sharing their experiences [17], while proactive scaffolding questions guide them through the conversation [29, 59]. Medical surveys often require disambiguation questions to clarify meanings, such as time references or degrees [63], especially for patients with low health literacy [47]. However, existing solutions are often script-based, lacking the flexibility to adapt to user responses [29], and have been criticized for lacking authenticity [37].

Beyond the immediate challenges of eliciting and clarifying information, the structure and sequence of questions in interview tasks play a crucial role in shaping the quality and accuracy of responses. The spreading activation theory suggests that activating one concept can spread to related concepts, aiding information recall [4]. Consequently, the order of presenting questions significantly influences survey responses [78]. For example, self-rated health appeared worse when asked before chronic conditions rather than after [53]. Similarly, overall evaluation questions asked before specific service questions tended to elicit more neutral or negative responses [6]. These findings underscore the need for logical question organization and validation mechanisms to ensure data quality.

While managing conversation flow is important, maintaining participant engagement throughout the interaction is equally crucial [59]. Empathy skills are essential for sustaining engagement and ensuring data quality [58]. Both cognitive and emotional empathy positively influence recipients' feelings and emotions [79]. Interestingly, most participants prefer simple detection and acknowledgment of their emotions with scripted messages [14, 57]. However, these skills have primarily been explored in short conversations, leaving the challenge of maintaining engagement throughout long surveys and gracefully exiting conversations when necessary.

2.4 LLMs for Medical Scenarios

LLMs are increasingly used in healthcare dialog systems [34]. These models show particular promise in medical triage, screening, and detection tasks [34]. Recent advancements in LLMs have led to breakthroughs in open-domain dialog systems, enabling free-form conversations on diverse topics [31]. Such capabilities offer potential benefits for public health interventions, particularly in providing empathetic interactions for individuals facing challenging health experiences and reaching underserved populations [31, 39]. Integrating LLM-powered CAs into healthcare systems has been explored to elicit patient health information and support public and personal health needs, such as medical self-diagnosis [61] and social isolation intervention [38].

However, LLM-driven CAs still have demonstrated limitations in recent medical-related communication tasks like hospital admission interviews, particularly in managing conversations and collected information. Conversational behavioral issues such as providing inappropriate follow-ups, repetitiveness, and less context continuity have significantly impacted user experience [61]. More specifically, issues include premature conclusion-jumping and insufficient information-seeking behaviors [55]. Minor deviations from human-like behavior, such as inadequate follow-up questions and overlooking important details [55], repetitive or irrelevant responses [61], have also been observed. Other important language capabilities, such as keeping users informed about the conversation's progress and providing clear explanations, are not garrenteed in LLMs by default [52]. Besides, from a technical perspective, these systems often struggle with long-term memory integration, which affects their ability to reference personal health history information from past interactions, reducing user engagement and decreasing the effectiveness of systems [38]. Recent works explored the possibility of integrating long-term memory into LLMs to address this issue, but included further discussion over privacy [39]. Additionally, the phenomenon of hallucination poses a significant challenge, where CAs generate responses that are either ungrounded in the conversation or factually incorrect [30, 52].

While technical challenges may be addressed through improvements in LLM architecture, addressing behavioral issues requires more sophisticated design approaches to ensure high-quality conversations and reliable data collection. The question of how to extend the existing capabilities of LLM-driven CAs to overcome these challenges remains open and warrants further research.

3 Formative Study

To address **RQ1**, we conducted a formative study to examine current practices in hospital admission interviews based on the Hospital Authority PAF, a 7-page document with 13 assessment blocks currently used by local public hospitals. Our collaborating hospital identified five blocks, namely *respiratory status*, *pain*, *ambulation*, *communication*, and *nutrition*, that are filled during patient interviews. The remaining blocks are completed by nurses or doctors based on observations, electronic medical records, or physical examination results. Two nurses provided video demonstrations showing how they orally present PAF questions and medical terminology to patients. We then created a mock-up system (detailed in Section 3.1) accordingly to demonstrate the use of an LLM-powered CA for hospital admission interviews and invited six nurses and four hospital volunteers to act as simulated patients and experience this system. We collected feedback on participants' perceptions of the idea and the system's performance and identified challenges patients may face during the interviews. We conducted follow-up semi-structured interviews that explored the participants' strategies for overcoming these challenges in real-world practices, summarized in Table 2 and Table 3.

3.1 Prototype System

We developed a preliminary prototype voice CA that can conduct hospital admission interviews using relevant PAF blocks as a guide with off-the-shelf technologies [60]. It incorporates features aligned with previous research on patient interviews [2, 32, 82], including welcoming messages, scripted questions, responses to user answers, and expressions of gratitude at the end. User interactions are managed by a dialogue manager implemented with GPT-4 [69]. Upon receiving a response, the dialogue manager extracts information to perform questionnaire slot-filling and determines whether to proceed to the next question (if sufficient information is provided) or repeat the question (if the response is unclear). The system also summarizes the information collected into a completed 5-block PAF at the end of the interview. The system aims to probe the feasibility of using CA for hospital admission interviews and to identify potential challenges and strategies for addressing them.

3.2 Participants and Procedure

Due to hospital privacy regulations and ethical concerns regarding observing vulnerable patients, we conducted experiments with nurses and hospital volunteers as patient proxies [7], an approach consistent with medical training practices [10]. While healthcare providers may not fully represent subjective patient experiences, their professional expertise and extensive experience with diverse patient populations make them particularly suitable for evaluating standardized admission interviews. Their ability to simulate concrete interactions and provide insights from multiple patient cases offers valuable perspectives that would be difficult to obtain from individual patients [84]. With IRB approval, we recruited six nurses and four hospital volunteers, all with prior experiences in administrating hospital admission interviews, from our collaborating hospital in an East Asian city via internal communication channels. Demographic information is presented in Table 1. Note that all nurses had professional training on communicating with patients, while the volunteers are part-time staff who assist during peak times. None of the participants had used or seen CAs for patient interviews. It was thus important for them to interact with such a system to align their understanding before eliciting feedback.

Figure 2 illustrates our formative study workflow. The study was carried out in a quiet room simulating the waiting areas where the patient interviews typically occur in our collaborating hospital. Each participant engaged in four rounds of role-play, where in each round, a participant role-played one of the two general categories of patients they typically encounter in their daily work – those who need more additional assistance or those do not – and responded to the CA's PAF questions as real patients would (step 1 and 2 in Figure 2). The order of the patient categories participants role-played was counterbalanced in 4 rounds to mitigate order effects. As individual differences among patients can be vast, we asked the participants to act twice in each round with different responses and behaviors (step 4 in Figure 2). The goal of this activity was for the nurses and volunteers to gain first-hand knowledge of the performance and interactive capabilities of current state-of-the-art LLM-based CA technologies in hospital admission tasks. After each role-playing session, we carried out a semi-structured interview (step 3 in Figure 2) to gather participants' feedback on (1) the characteristics of the patient they just simulated, (2) challenges faced by the patient during the interview, (3) nurses' and volunteers' empirical strategies to address these challenges, and (4) other feedback on the CA. Upon completion of the entire role-playing activity, we conducted an exit interview with participants to confirm their suggestions and discuss CA's potential application in the hospital. The entire study lasted about 90 minutes for each participant. With the participants' consent, we video-recorded the entire study, logged the conversations in the role-playing sessions, and transcribed all the speech data from the interviews for later analysis.

3.3 Key Findings from Formative Study

Nurses and volunteers expressed positive attitudes toward using CAs in hospital admission interviews, particularly appreciating the LLM's ability to understand responses and summarize information in the PAF. While they noted occasional ASR errors and stilted question delivery, they emphasized that CAs should complement rather than replace human involvement, as real-world scenarios present greater complexity than simulated interactions.

3.3.1 *The Admission Interview Task and Scope of CA.*

Participants identified admission interviews as particularly suitable for CA assistance in the whole admission process. Unlike their other critical duties (vital signs monitoring, medical examination preparation, patient care, etc), interviews follow a more structured format and easier to be handled by automated system. They viewed the CA as valuable tools for information collection and documentation, especially beneficial when nurses face interruptions or urgent calls during interviews. A dedicated CA could maintain interview continuity while allowing nurses to focus on essential care responsibilities.

Nurses and volunteers raised concerns that not all patients are appropriate for CA interactions, emphasizing that cases involving physical or cognitive limitations (hearing impairments, disorientation), or patients feeling unfit to respond demand human expertise and inference. However, they noted that such challenging cases can be identified through preliminary EHR review and brief initial patient interactions. This early assessment enables informed decisions about patient suitability, suggesting that CAs could be selectively deployed for straightforward interactions with communicative patients while healthcare professionals manage complex cases. They emphasized that the system should promptly alert nursing staff when communication difficulties arise during CA-patient interactions.

Even for more communicative patients deemed suitable for CA interaction, nurses highlighted the need for CAs to better handle diverse and nuanced patient responses. We explored these challenges and nurses' empirical strategies for addressing them in our semi-structured interviews, with detailed findings presented in Table 2 and Table 3.

3.3.2 *Challenges and Strategies in Conducting Admission Interviews.*

To identify the challenges nurses face in hospital admission interviews and what strategy they use to resolve respectively, we first conducted an inductive analysis of the interview transcripts, extracting the related information from the participant's responses and setting the responding code for each [26]. Then, we refined and organized our themes and codes through a comparative process, incorporating insights from related work in similar contexts [29, 59, 66]. While some themes overlapped conceptually with prior work, we preserved the unique nuances of our research context. Due to space constraints, we present a summary of the challenges and strategies identified in the formative study in Table 2 and 3. We also attached a full table with detailed examples in the supplementary materials and discussed its generalizability in Section 7.4.

According to our inductive codes, the main challenges nurses face fall into two types: *information collection* and *engagement issues. Information collection* issues include cases where patients provide incomplete, inaccurate, or additional, conflicting information, risking misdiagnosis or unsuitable treatment (P3.1 - P3.3). To handle additional and conflicting information issue, nurses suggested maintaining detailed records, logically guiding questions to prevent distractions (T1), following up questions to ensure accurate information (T2), and validating related information when reaching the later corresponding blocks in PAF (T4). For example, if a patient mentions "eating less" and "poor sleep" due to coughing during respiratory questions, these hints of "weight loss" should be cross-validated when reaching "weight loss" related questions (N1). To handle ambiguous or incomplete answers (P4.1, P4.2), nurses suggested providing scaffolding questions and follow-ups to gather details (S2, S4). For example, when asked if any part of the body feels uncomfortable, patients only describe how it feels but miss which part (N10), requiring follow-up questions for clarification.

Table 1: Participants' demographic information. IDs starting with 'V' are hospital volunteers, and those starting with 'N' are nurses. The 'Experience' column indicates years of experience, while 'Patients/Day' represents the average number of patients they interact with daily.

ID	Gender	Age	Experience	Patients/Day	ID	Gender	Age	Experience	Patients/Day
V1	Female	65+	15+ yrs	5-10	V6	Female	55-64	1-5 yrs	20+
V2	Female	55-64	1-5 yrs	20+	N7	Male	25-34	6-10 yrs	20+
N3	Female	45-54	15+ yrs	20+	N8	Female	45-54	15+ yrs	20+
N4	Female	35-44	15+ yrs	10-20	N9	Female	25-34	1-5 yrs	20+
V5	Female	65+	<1 yr	20+	N10	Female	35-44	11-15 yrs	20+

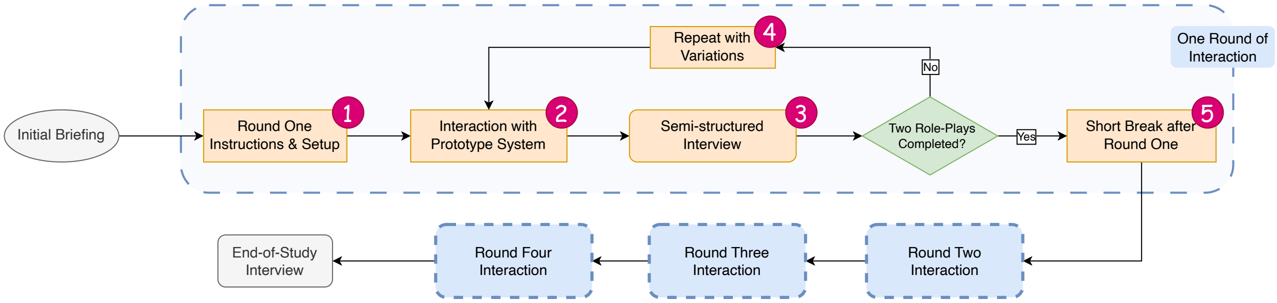


Figure 2: Formative study workflow. Participants first receive an introduction to the CA, then engage in 4 rounds of role-play. In each session, participants receive instructions on the patient category they are role-playing (step 1), interact with the CA (step 2), and participate in a semi-structured interview to provide feedback (step 3). They then role-play the same category of patient with different behavior (step 4, followed by steps 2 and 3). A short break is provided at the end of each round (step 5). Each participant completes 4 rounds of interactions, resulting in 8 role-play sessions. After the 4 rounds, an exit interview collects nurses' overall thoughts and expectations regarding the CA.

When patients struggle with uncertainty, measurement lacking, or miscomprehension (P5.1-P5.3), strategies like explaining questions (S3), prompting memory (S1), and contextual follow-ups (S4) are commonly used. For example, asking if clothes feel looser can help patients recall weight loss (N4, N9). Besides, patients often meet with *engagement issues* during admission interviews, like impatience, sadness, anger, or silence can affect their participation (P1.1 - P1.4). For example, patients asked to repeat answers may grow impatient or even wish to directly end the conversation (N3, N4). While most remain cooperative, providing emotional support (E1) can help them relax and engage.Physical or cognitive issues, like hearing impairments (P2.1), disorientation (P2.2), or feeling unfit to answer (P2.3), also impact interviews. Suggested strategies include repeating questions with emphasis (S3.1), orienting patients (S3.2), or gracefully exiting the conversation when necessary (T3). For further details, please see the code-level mappings in the last column of Table 3 and more examples in Section 4 of the supplementary materials.

Building on previous research insights [29, 59, 66], we categorized all strategies (S1-S4, E1-E2, T1-T4) into two core system design requirements. **Turn-Level Language Adaptation (DR1)**: Focuses on *immediately* adjustments in current conversational turns to scaffold questions (S1-S4), show empathy (E1), and provide orientation support (E2) based on patient responses. **Conversation-Level Topic Management (DR2)**: Addresses *long-term* management block by block by following a logical order within a block, altering the block order smoothly according to syndromes mentioned (T1), recording and following up on previously mentioned related syndromes in other questions (T2), exiting gracefully if necessary (T3), and cross-validating information across blocks for accuracy (T4).

In the following sections, we detail our designed CA based on these requirements and evaluate its performance in hospital admission interviews with a technical evaluation and a 44-subject between-group user study.

4 System Design

In our collaborative hospital, PAF questions are typically completed in quiet waiting rooms. Given the inconvenience of manual writing or input for patients, nurses prefer verbal interviews. Following this real-world process, we incorporated a voice CA in our formative study, validating its effectiveness and adaptability, and chose it as the primary interface for our system. Moreover, since our study focuses on improving language capability, the CA served as an embodiment to deploy and assess this capability. However, our system is not limited to CAs, which could also be applied to other embodiment systems like tablets, smart speakers, or robots. Besides, the CAs in this study were designed to focus only on completing the PAF in an EHR-compatible format and providing descriptive notes to nurses. Pre-recorded patient information from the EHR system or known to the admitting nurses was not considered in our design, because the collaborating hospital and the Hospital Authority restrict access to EHR data. Letting nurses input such

Table 2: Challenges found in hospital admission interviews. The "P" in the ID stands for Problem.

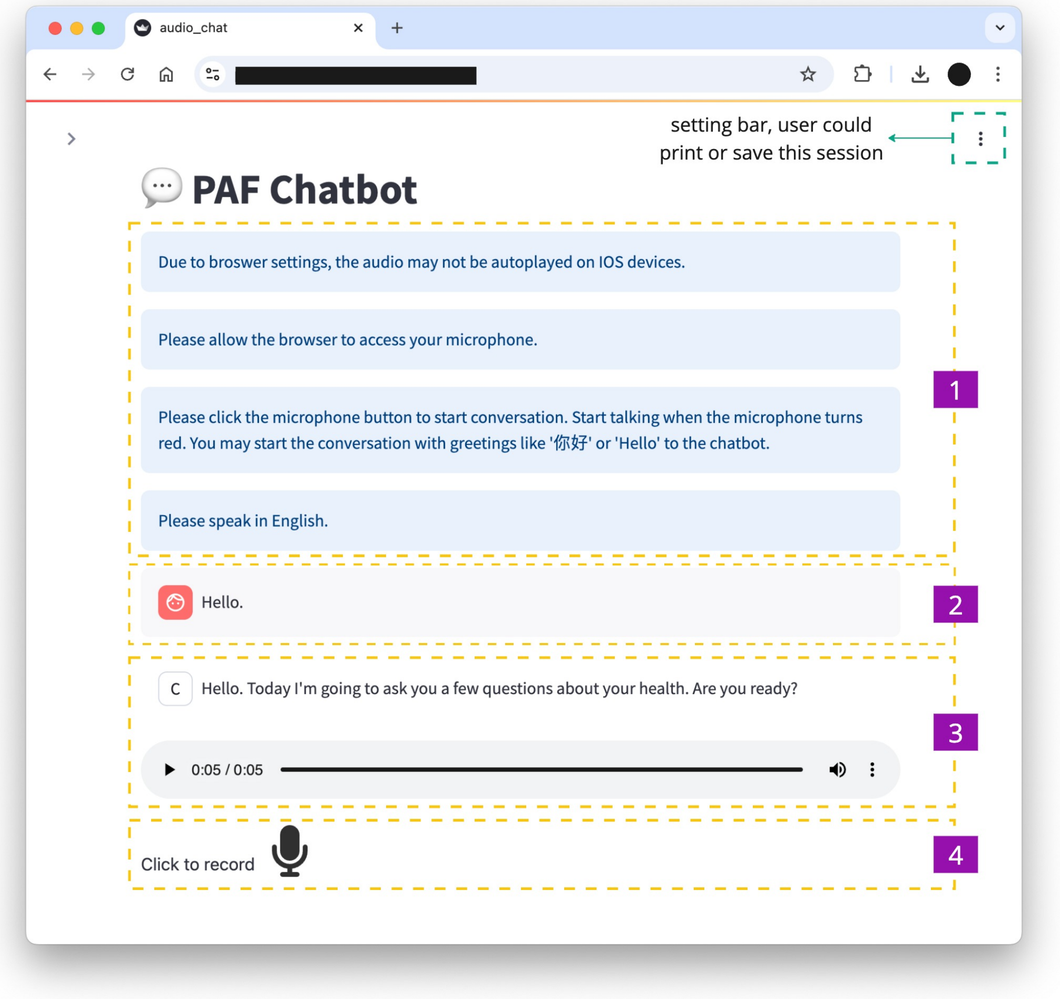
Theme (Challenges)	Sub-theme	ID	Code
Engagement Issue	Emotional Problems	P1.1	Impatient
		P1.2	Agitated or Angry
		P1.3	Silent / Avoidance
		P1.4	Sad
	Physical or Mental Issues	P2.1	Hearing Problem
		P2.2	Disoriented
		P2.3	Unfit / Unwell
Information Collection	Information Retrieval	P3.1	Overloaded / Pre-emptive Answer
		P3.2	Jumping Around
		P3.3	Irrelevant Information
	Inaccurate Answer	P4.1	Ambiguous Answer
		P4.2	Answer Lacks Details
	Cannot Answer	P5.1	No Relevant Information Cannot Remember / Unsure
		P5.2	Lacking a Measure
		P5.3	Do Not Understand the Question

Table 3: Solutions for challenges found in hospital admission interviews and the mappings to challenges. The "S, E, T" in the ID stands for Solution, Empathy, and Topic Management, respectively.

Theme (Solutions)	Sub-theme	ID	Code	Solvable Challenges
Scaffolding	Provoke memory	S1.1	Make comparison	P5.1, P5.2
		S1.2	Frame around the concept	P5.1, P5.2
	Simplify question	S2.1	Give a level or choice	P4.1, P5.2
		S2.2	Ask a simpler question (but keep the original intention)	P5.1, P5.2
	Further explain the question	S3.1	Repeat with emphasis	P2.1, P3.2
		S3.2	Explain keywords (word level)	P5.3
		S3.3	Rephrase with easier sentences (sentence level)	P5.2, P5.3
	Follow-up question	S4.1	Further check relevant information	P4.1, P4.2
		S4.2	Further check in details (focus one current question)	P4.2
Empathy & Orientation	Orientation support	E1.1	Explain current situation	P1.1–1.4
		E1.2	Address what the agent will do	P1.1, P1.2
		E1.3	Address what patient should do	P1.1, P1.2
	Provide emotional support	E2.1	Comfort sentences	P1.3–1.4, 2.3, 3.3
		E2.2	Be polite	P1.1–1.4, P3.2
		E2.3	Encourage / Motivate patient	P1.3–1.4, P2.3
Topic Management	Adjust question order	T1.1	Group similar questions by block	P3.1
		T1.2	General order of question groups	P3.1, P3.2
	Stay on topic	T2.1	Pick-up descriptions and set remarks	P3.1, P3.2
		T2.2	Follow up important syndromes (deviate and set remarks)	P3.2
	Gracefully exit	T3.1	Check Orientedness	P2.2
		T3.2	Guide back	P2.2
	Cross validation	T4.1	Check Consistency	P3.2

information manually to our system is also not feasible due to the time constraints of the hospitalization process.

Based on these considerations and the design requirements, we implemented two LLM-based CA prototypes: the baseline prototype and our prototype. Concerning user data privacy, we initially developed our system using a local LLM [20], but its frequent hallucinations, such as off-topic responses and fabricated information, had significantly decreased performance. Given our study focuses on exploring CAs in their early development stages, we finally built



audio_chat

[redacted]

PAF Chatbot

Due to broswer settings, the audio may not be autoplayed on IOS devices.

Please allow the browser to access your microphone.

Please click the microphone button to start conversation. Start talking when the microphone turns red. You may start the conversation with greetings like '你好' or 'Hello' to the chatbot.

Please speak in English.

Hello.

C Hello. Today I'm going to ask you a few questions about your health. Are you ready?

0:05 / 0:05

Click to record

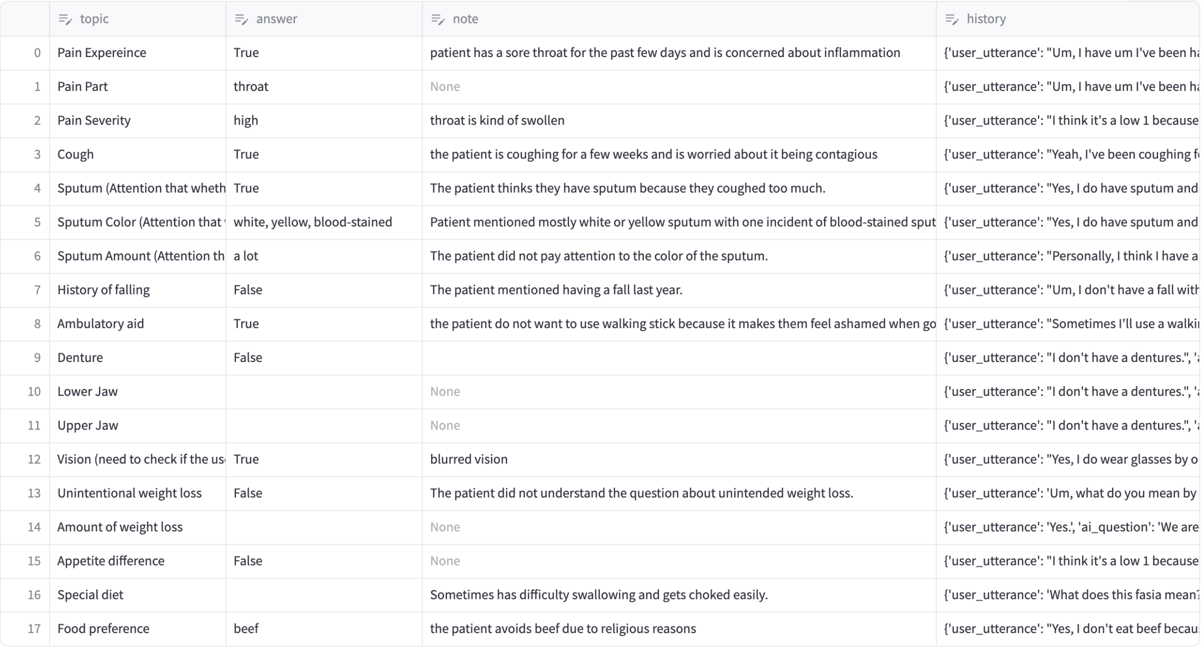
Figure 3: CA Interface

the system using GPT-4o[1]. Both prototypes share a Flask-based web interface [71] for user interactions (see Figure 3). The interface mimics turn-based interaction, inspired by CAs like Siri [33]. The interface provides a simple user guide (see box $\boxed{1}$ in Figure 3). During each turn, the user could click the microphone button (see box $\boxed{4}$ in Figure 3) to input the response via voice. The CA employed ASR [11] to recognize the user-spoken content(see box $\boxed{2}$ in Figure 3) and generated a spoken response via Text-to-Speech [12], which is automatically played (see box $\boxed{3}$ in Figure 3). At the end of the conversation, the controller summarizes the recorded answer into a completed 5-block PAF (see Figure 4). To ensure a fair comparison, both baseline and our system shared the same voice-to-text and text-to-voice interface. Besides, both systems obeys the same task requirements of admission interview.

4.1 Baseline CA Prototype

Previous study [44] has validated the significant priority of the GPT-4o [68] in the performance on a range of language tasks, especially multilingual ones where the generative answers are required to be coherent and varied; and the open-ended nature of creative writing tasks. Therefore, GPT-4o is considered to have huge potential to generate language content that meets writers' needs without any extra engineering. Following this, we implemented the baseline CA system adapted from Li et al.'s work [55] with GPT-4o, which is fully controlled by one module with prompt chaining and requirement concatenation. Detailedly, the prompt defines the role of the bot as a nurse, outlining the core elements of hospital interviews and the requirements for conducting medical interviews professionally. It also reminds the collection of accurate patient information by using the LLM to detect potential mismatches in responses and autonomously generate corrective conversations (As shown in Supplementary Material 3.1).

[1] Version gpt-4o-2024-05-13



	topic	answer	note	history
0	Pain Expereince	True	patient has a sore throat for the past few days and is concerned about inflammation	{'user_utterance': "Um, I have um I've been ha
1	Pain Part	throat	None	{'user_utterance': "Um, I have um I've been ha
2	Pain Severity	high	throat is kind of swollen	{'user_utterance': "I think it's a low 1 because
3	Cough	True	the patient is coughing for a few weeks and is worried about it being contagious	{'user_utterance': "Yeah, I've been coughing f
4	Sputum (Attention that wheth	True	The patient thinks they have sputum because they coughed too much.	{'user_utterance': "Yes, I do have sputum and
5	Sputum Color (Attention that	white, yellow, blood-stained	Patient mentioned mostly white or yellow sputum with one incident of blood-stained sput	{'user_utterance': "Yes, I do have sputum and
6	Sputum Amount (Attention th	a lot	The patient did not pay attention to the color of the sputum.	{'user_utterance': "Personally, I think I have a
7	History of falling	False	The patient mentioned having a fall last year.	{'user_utterance': "Um, I don't have a fall with
8	Ambulatory aid	True	the patient do not want to use walking stick because it makes them feel ashamed when go	{'user_utterance': "Sometimes I'll use a walki
9	Denture	False		{'user_utterance': "I don't have a dentures.", '
10	Lower Jaw		None	{'user_utterance': "I don't have a dentures.", '
11	Upper Jaw		None	{'user_utterance': "I don't have a dentures.", '
12	Vision (need to check if the us	True	blurred vision	{'user_utterance': "Yes, I do wear glasses by o
13	Unintentional weight loss	False	The patient did not understand the question about unintended weight loss.	{'user_utterance': 'Um, what do you mean by
14	Amount of weight loss		None	{'user_utterance': 'Yes.', 'ai_question': 'We are
15	Appetite difference	False	None	{'user_utterance': "I think it's a low 1 because
16	Special diet		Sometimes has difficulty swallowing and gets choked easily.	{'user_utterance': 'What does this fasia mean?
17	Food preference	beef	the patient avoids beef due to religious reasons	{'user_utterance': "Yes, I don't eat beef becau

Figure 4: Filled PAF Result Form. The displayed form comes from P43 who interacted with our model in the user evaluation.

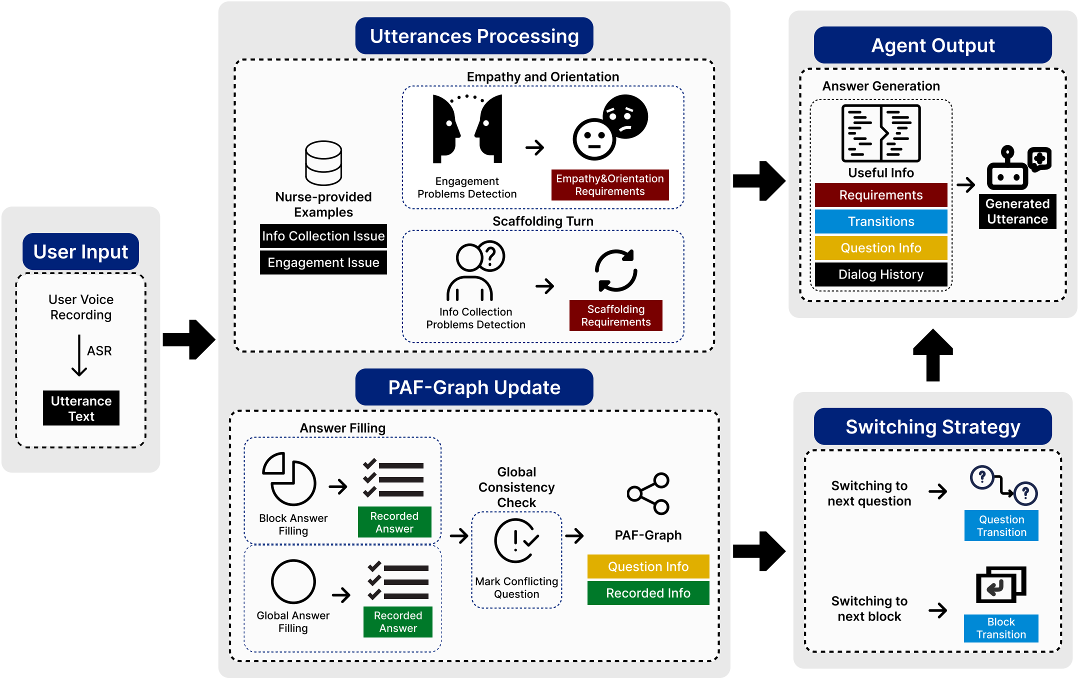


Figure 5: Overall system design of our CA prototype

4.2 Our CA Prototype

Our system is managed by a dialogue management controller (see Figure 5), which is also based on GPT-4o. The dialogue interaction follows a turn-based structure, with each turn comprising two utterances: one from the user and one from the CA. To achieve *turn-level* language adaptation (DR1), we implemented the Empathy & Orientation Module and the Scaffolding Turn Module, which detect and address the scaffolding challenges (S1 to S4), as well as the empathy & orientation requirements (E1, E2). To achieve *conversational-level* topic management (DR2), we designed the Answer-Filling Module, the Consistency Check Module, and the graph-structured PAF. These modules are developed to manage the block order with the user-mentioned information, support the logical and smooth transition between questions and blocks (T1, T2), and control cross-validation for conflicting answers (T4). Within each turn of dialogue, the utterances of both the agent and the user are parallel analyzed across the Empathy and Orientation Module, the Scaffolding Module, and the Answer-Filling Module. Besides, we designed the Response Generation Module to phrase the agent's utterance in the next turn. The prompts used in our system are specifically tailored to achieve distinct functions for each block (As shown in Supplementary Material 3.2).

4.2.1 Empathy & Orientation Module and Scaffolding Turn Module. These modules are implemented based on few-shot learning prompting, which leveraged the real-life utterances (see Table 2) we obtained from the formative study. Specifically, the prompts in these blocks include instructions to identify issues such as accuracy or sufficiency in the answers provided by the patient, or any emotional concerns that could affect the interview (As shown in Supplementary Material 3.2.1, 3.2.2). With identified issues, the conversation controller assigns corresponding predefined strategies (see Table 3) in the later response generation. The controller is designed to terminate the conversation in scenarios where the user remains disoriented or uncooperative (T3).

4.2.2 Answer-Filling Module and Global Consistency Check Module. We parsed PAF questionnaire into a graph structure (see Figure 6). The Answer-Filling Module processes each question block individually following a pre-defined answer-filling chain and extracting relevant information for pre-defined specific information points in each block (As shown in Supplementary Material 3.2.3). *Block-level* filling checks if the current conversation contains answers to the current block, while *global-level* filling reviews all blocks to find any related information mentioned. The extracted results are then updated to the PAF graph, and relative information for future questions is marked to prevent repeated questions. Specifically, global-level answer-filling may trigger a block reordering, adjusting the sequence to prioritize blocks with mentioned information. If inconsistencies arise between newly extracted and previously recorded answers, the system initiates a consistency check, flagging conflicting responses for later clarification after the current question block.

4.2.3 Switching Strategy and Response Generation Module. The controller assesses whether to proceed to the next question based on the sufficiency and the clarity of the recorded information for the current question. When the communication switches to the next question, the system generates transitional sentences to inform the user of the question switching and the possible block topic changing. The Response Generation Module generates the agent's utterance for the next turn, incorporating the selected information provided by the previous modules and the latest chat history. The prompt here (As shown in Supplementary Material 3.2.5) includes the relevant information from previous interactions, such as the extracted answers, identified issues, and example questions. It also includes an explanation of how to combine all the information to generate contextually appropriate responses in the current turn. When all the questions in the questionnaire are detected to be finished by status checking, the controller concludes the conversation.

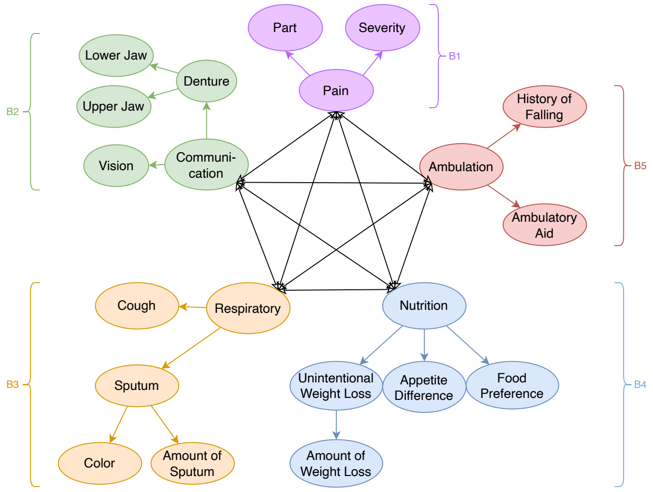


Figure 6: The PAF Graph. It involves five blocks assess through admission interview, which are pain, communication, respiratory, nutrition and ambulation (serve as root nodes). Each question block is represented as a branch with information points as nodes. The root nodes for each branch are interconnected, which achieves the transition between the blocks. Related questions are organized hierarchically, with parent and child nodes. Each node stores information such as the current block topic, the medical question information, and the recorded results. Related questions are organized hierarchically, with parent and child nodes. To manage the process of the questionnaire in the graph, we implemented a graph algorithm to control the node answer filling and updating, question skipping, node marking, and status checking for the questionnaire.

5 Experiment 1: Technical Evaluation

To assess the performance and user experience our proposed prototype, we conducted two evaluation studies: **a technical evaluation** and **a user study**. We followed Maroengsit et al.'s survey on CA evaluation methods [62] to assess our prototype at both *turn-level* and *conversation-level*.

The **technical evaluation** aimed to assess the prototype's performance in *turn-level* language capabilities (DR1) and ensure that the CA can generate language based on the context of the conversation (RQ2). We thus framed CA's *turn-level* capabilities as a language generation task and compared our prototype with ground truth data collected from nurses as well as a baseline LLM model adapted from Li et al.'s work [55]. We deployed a turn-level evaluation with reference to [62]. The following sections describe the task formulation, participants, experiment protocol, and results of the technical evaluation study.

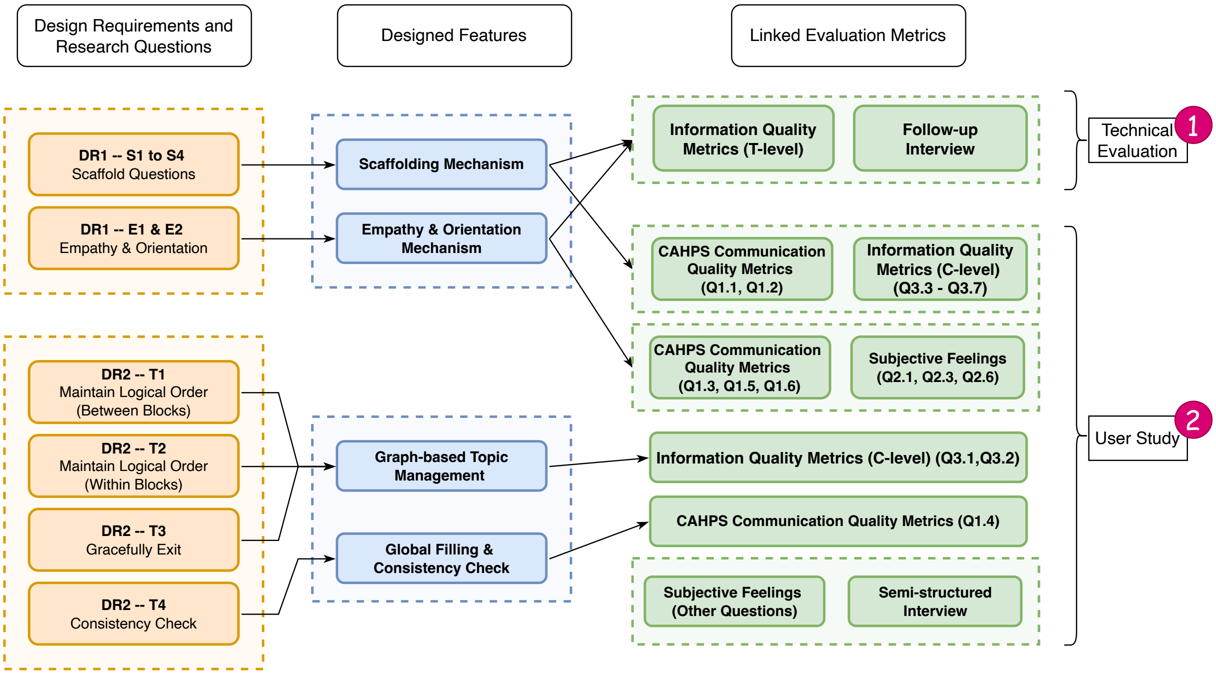

Design Requirements and Research Questions
Designed Features
Linked Evaluation Metrics
DR1 -- S1 to S4 Scaffold Questions
DR1 -- E1 & E2 Empathy & Orientation
DR2 -- T1 Maintain Logical Order (Between Blocks)
DR2 -- T2 Maintain Logical Order (Within Blocks)
DR2 -- T3 Gracefully Exit
DR2 -- T4 Consistency Check
Scaffolding Mechanism
Empathy & Orientation Mechanism
Graph-based Topic Management
Global Filling & Consistency Check
Information Quality Metrics (T-level)
Follow-up Interview
CAHPS Communication Quality Metrics (Q1.1, Q1.2)
Information Quality Metrics (C-level) (Q3.3 - Q3.7)
CAHPS Communication Quality Metrics (Q1.3, Q1.5, Q1.6)
Subjective Feelings (Q2.1, Q2.3, Q2.6)
Information Quality Metrics (C-level) (Q3.1,Q3.2)
CAHPS Communication Quality Metrics (Q1.4)
Subjective Feelings (Other Questions)
Semi-structured Interview
Technical Evaluation
1
User Study
2

Figure 7: Evaluation logistics. Two evaluation studies are planned: a technical evaluation (marked with purple 1) and a user study (marked with purple 2). The correspondence between design requirements (DR), corresponding evaluation metrics, evaluation studies. For information quality metrics, "T-level" refers to turn-level metrics, and "C-level" refers to conversation-level metrics.

5.1 Task Formulation, Evaluation Metrics, and Data Preparation

We formulated a language generation task based on the immediate preceding turn of questioning (by the CA) and answering (by the patient). We thus framed the technical evaluation task as: given a previous turn as context, assessing the quality of the subsequent scaffolding question generated by our model – concerning how well it follows the Grice's maxims of conversation [24] – in comparison to the human ground truth and that produced by a baseline LLM. We derived most of our metrics for human ratings from the Gricean Maxims as described in [91], with slight modification to the narrative of some items to fit the context of hospital admission interview. We added four additional metrics: *Factual Consistency*, *Coherence*, *Conciseness* and *Empathy*. *Factual Consistency* intended to evaluate the severity of hallucinatory errors in the generated questions, as LLMs might generate factually incorrect information [30]. *Coherence* corresponded to the "orderly" statement in the Manner maxim [24]. We also added *Empathy* to evaluate the emotional supportiveness of the generated questions, as empathy is crucial in patient interviews [58]. The full list of metrics are presented in Table 4.

For the technical evaluation study, we invited nurses and hospital volunteers from our formative study to provide conversation contexts and suggest subsequent scaffolding questions for each problem-solving skill they mentioned. After removing samples already used in our system's training data, we selected 40 conversation contexts, comprising two samples for each of the 20 turn-level question-asking skills, as the data used for technical evaluation. Each context was associated with ground truth data and two on-the-fly generated questions: one from the baseline model and one from our prototype.

5.2 Experiment Interface, Participants, and Evaluation Protocol

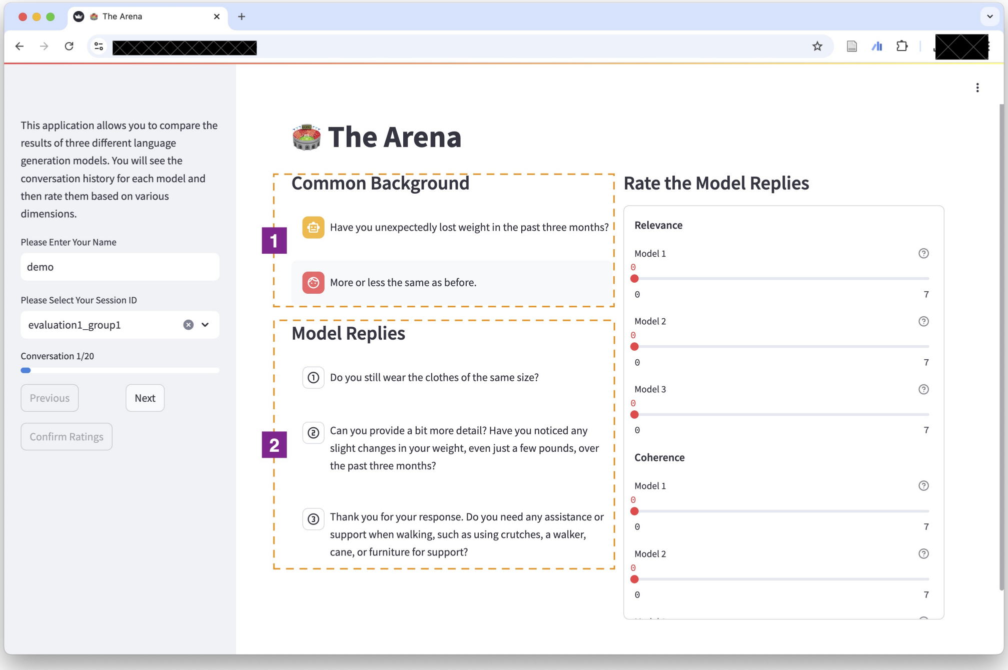
We designed an Arena-style interface for participants to evaluate the performance of three models side by side: our prototype (denoted as *ours*), the human model (i.e., the ground truth data generated by nurses and volunteers *gt*), and the baseline model (*baseline* using GPT-4o[2]). In each round of evaluation, the participants were presented with a conversation context (box $\boxed{1}$ in Figure 8) and three generated questions (box $\boxed{2}$ in Figure 8) from the three models simultaneously. Participants were asked to rate each generated question on a 7-point Likert scale (1 for very poor and 7 for excellent) for every item listed in Table 4.

We performed a prior power analysis to determine the sample size required for the study. We calculated the sample size based on an effect size of 0.5, a significance level of 0.05, a power of 0.95, and the need of three pairwise comparisons to distinguish the three models. The result indicated that we would need at least 47 data points for each model. Since our dataset included 40 conversation context samples (two for each of the 20 turn-level question-asking skills), we recruited four human raters (3 male and 1 female, aged 21 to 34) from a local university, each evaluating 20 context samples –

[2] The specific version we chose is gpt-4o-2024-05-13. Prompts provided in supplementary materials

Table 4: Gricean Maxims used to guide the development of quality metrics for evaluating the turn-level language generation task. We developed information quality metrics (V1) from Gricean Maxims to evaluate the quality of the generated questions.

Gricean Maxims	Original Definition	Our Quality Metrics	Definition
Quantity	One should be as informative as possible	Informativeness	To what degree does the question or response have the potential to elicit informative answers?
		Specificity	How detailed and contextually tailored is the question?
Quality	One should communicate truthfully	Factual Correctness	To what degree is the question free from factual errors or misleading implications?
Relevance	One should provide relevant information	Relevance	To what extent is the question or response relevant to the given context or topic?
Manner	One should communicate in a clear and orderly manner	Clarity	How clear and unambiguous is the question or response in conveying its intent?
		Coherence	How logically structured and understandable is the question or response?
		Empathy	How emotionally supportive is the question or response?



[redacted]

[redacted]

The Arena

This application allows you to compare the results of three different language generation models. You will see the conversation history for each model and then rate them based on various dimensions.

Please Enter Your Name: demo

Please Select Your Session ID: evaluation1_group1

Conversation 1/20

Previous Next

Confirm Ratings

Common Background

Have you unexpectedly lost weight in the past three months?

More or less the same as before.

Model Replies

① Do you still wear the clothes of the same size?

② Can you provide a bit more detail? Have you noticed any slight changes in your weight, even just a few pounds, over the past three months?

③ Thank you for your response. Do you need any assistance or support when walking, such as using crutches, a walker, cane, or furniture for support?

Rate the Model Replies

Relevance

Model 1: 0 (0–7)

Model 2: 0 (0–7)

Model 3: 0 (0–7)

Coherence

Model 1: 0 (0–7)

Model 2: 0 (0–7)

Figure 8: The Arena-style evaluation interface used in the technical evaluation study.

one for each skill. They were all fluent in English and knowledgeable about the challenges in doctor-patient communications. Two of them were HCI researchers in the healthcare domain (P1 and P3), one was an investment analyst in the healthcare industry (P2), and one was a patient recently discharged from the hospital (P4). We counterbalanced the order of selected samples to ensure that each conversation context was evaluated by two participants, which resulted in a total of 80 (>47) data points.

To mitigate potential bias, we provided a brief introduction to the participants about the research background and the evaluation metrics. We also conducted a practice session to familiarize them with the evaluation interface and the evaluation process. Additionally, we anonymized the generated questions from the three models as model 1, model 2, and model 3, and randomized their display order to avoid learning effects and order effects. To simulate the randomness of LLMs, we generated the questions from conversation contexts on the fly during the evaluation study, and the questions were not stored in the LLM cache. The order of the conversation context samples was also randomized to avoid sequence effects. After the evaluation, we conducted an exit interview to collect general feedback from the participants. The evaluation study was carried out in a quiet room, and the participants were given sufficient time to complete their ratings. The results of the technical evaluation are presented in the following Section 5.3.

5.3 Results from Technical Evaluation

We first applied the Friedman test to examine within-group differences, confirming significant differences between the three models across all evaluation metrics ($p < 0.05$). Subsequently, we conducted post-hoc Wilcoxon signed-rank tests to compare our prototype's performance with the ground truth data and the baseline model, using the Holm-Bonferroni correction to adjust for multiple comparisons. The results, presented in Figure 9 and Table 5, show that our model significantly outperformed the baseline in five metrics: Relevance, Informativeness, Factual Correctness, Specificity, and Conciseness. The effect size was large for "Relevance" and medium for the other metrics. Differences in "Coherence" and "Clarity" were marginal, possibly because both systems used the same underlying language model. Besides, in this technical evalaution, we only focused on the turn-level language generation task. The assessment of "Coherence" and "Clarity" might require seeing a longer conversational history over multiple turns [62].

Interestingly, our model received significantly higher human ratings than the ground truth in most metrics, except for "Conciseness." All four participants found the agent-generated sentences significantly and consistently more verbose than those used by the nurses, despite being blind to the conditions during evaluation. There are several possible reasons. First, the nurses are familiar with the conversation context and master question asking skills through practices; they are thus able to generate more prompt and direct questions. Second, the limited man power in the local public hospitals might require the nurses to get through the admission interviews efficiently. However, though our system use more words than the ground truth data, the participants still perceived our system's responses as significantly more informative, specific, relevant, clear, factually correct and empathetic than the human ones. This could be due to the fact that our participants are not medical professionals, and they would prefer receiving more detailed and contextually tailored scaffolding questions and follow ups in the conversation from the perspective of a patient.

In summary, the technical evaluation demonstrates that our model effectively balances between human-like conciseness and AI-enhanced informativeness and specificity. It outperforms both the baseline AI model and human-generated responses in most metrics, suggesting its potential to provide comprehensive and tailored patient interviews.

6 Experiment 2: User Study

The **user study** focused on evaluating the prototype's language capabilities at both *turn-level* (DR1) and *conversation-level* (DR2), as well as user experience (RQ3) in a full hospital admission interview. It was designed as a between-subjects study involving 44 participants. This study compared our system against the baseline, a state-of-the-art LLM model, in terms of both turn-level (DR1) and conversation-level (DR2) language capabilities, as well as user experience (RQ3). We evaluated DR1 in both studies because we believed that turn-level language capabilities might be influenced by the context of the conversation, allowing us to understand how the system performs in different scenarios. The evaluation logistics are shown in Figure 7.

6.1 Participants

To avoid carryover effects, we designed our experiment as a between-subjects study. We carried out a prior power analysis with an effect size of 0.8, a significance level of 0.05, and a power of 0.80. The result indicated that a minimum sample size of 21 in each group was necessary. With IRB approval, we recruited who recently recovered from a medical condition that require hospital treatment through email, social media, and local community groups. A total of 44 (>42) participants were recruited (24 described themselves as male, 19 as female and 1 prefer not to say), with age from 19 to 32 years old. All participants are self-reported fluent in English. The participants were randomly assigned to one of the two conditions and were asked to interact with the assigned model in a simulated hospital admission interview described in the previous subsection. Demographic details are in supplementary material section 5.

6.2 Experiment Protocol

The experiment was conducted in a quiet room with a desk and a chair, simulating a hospital admission interview environment (Figure 10b). This study took place in a major metropolitan area in East Asia, characterized by high-volume public hospitals with significant patient loads. The experiment workflow is shown in Figure 10a. After an initial briefing and obtaining consent, participants completed a pre-study questionnaire to collect demographic information and their prior experience with VAs. We then invited participants to a practice session to familiarize them with the VA interface.

Since around 40% of the hospital admission interview questions concern the patient's symptoms, we held a warm-up session to help participants recall their recent experiences with serious illnesses that required emergency treatment or hospitalization. Participants were asked to recollect their symptoms and personal feelings at the time, and researchers took notes during the conversation for information verification after the main task.

After that, participants were instructed to situate themselves in the previous scenario and interact with the assigned model (baseline or ours, described in Section 4.1 and Section 4.2) as if they were in a hospital admission interview during their visit to the inpatient department. The participants interacted with the VA through a laptop computer with a microphone and speaker. The VA was presented as a web-based interface, and the participants were asked to speak to the agent as if they were in a hospital admission interview. During

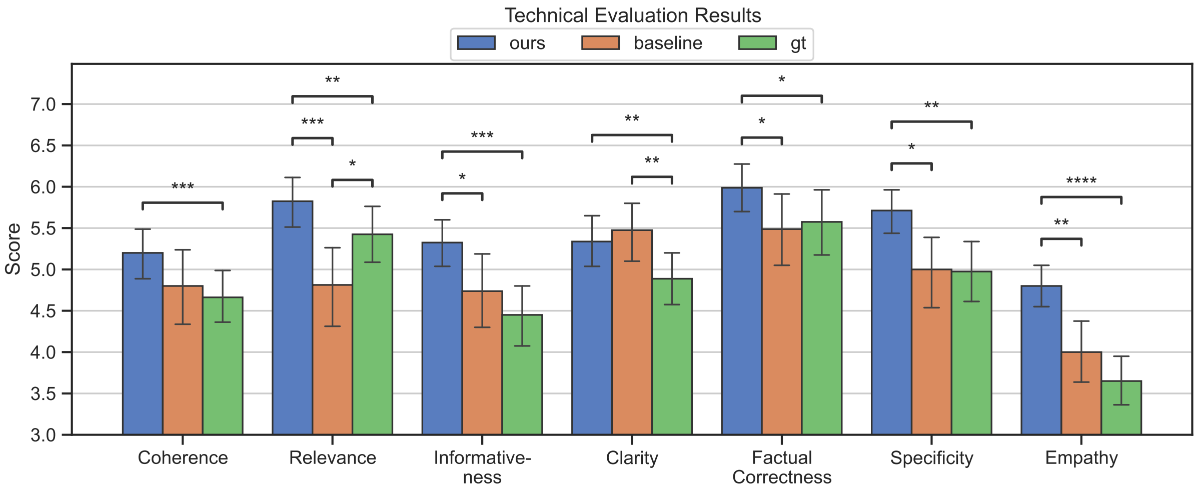


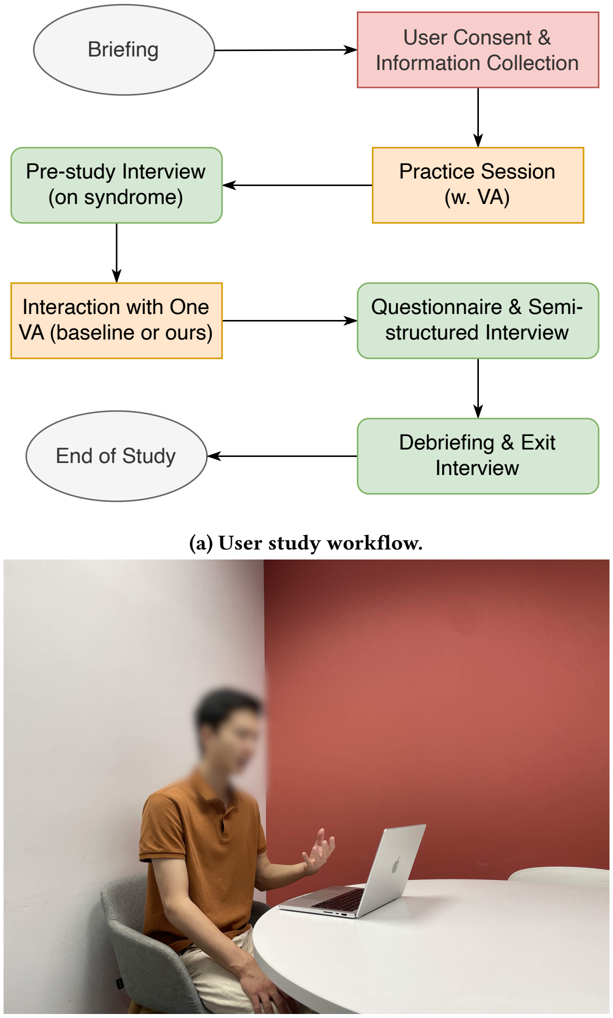
Figure 9: Results of the technical evaluation study. The bar plot shows the distribution of evaluation scores for three models: our prototype (*ours*), the ground truth data (*gt*), and the baseline model (*baseline*). Error bars represent the 95% confidence interval (CI). Significance levels are indicated as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5: Results of the technical evaluation study. The table shows the mean scores of the three models: our prototype (*ours*), the ground truth data (*gt*), and the baseline model (*baseline*), and the statistical significance of the differences between the models.

Factor	Ours Mean(S.D.)	Baseline Mean(S.D.)	Ground Truth Mean(S.D.)	Ground Truth vs Baseline			Ground Truth vs Ours			Baseline vs Ours		
				W	p-value	eff. Size	W	p-value	eff. Size	W	p-value	eff. Size
Coherence	5.20 (1.40)	4.80 (2.03)	4.66 (1.40)	773.50	0.142	-0.18	297.00	**0.001**	-0.55	551.00	0.142	-0.23
Relevance	5.83 (1.40)	4.81 (2.18)	5.42 (1.61)	819.50	**0.019**	0.34	115.50	**0.005**	-0.56	198.00	**0.000**	-0.63
Informativeness	5.33 (1.32)	4.74 (1.98)	4.45 (1.64)	1016.00	0.124	-0.16	447.50	**0.000**	-0.54	568.50	**0.038**	-0.31
Clarity	5.34 (1.40)	5.47 (1.54)	4.89 (1.45)	401.50	**0.002**	-0.48	523.50	**0.009**	-0.39	426.50	0.305	0.09
Factual Correctness	5.99 (1.33)	5.49 (1.99)	5.58 (1.76)	477.00	0.272	0.11	186.00	**0.017**	-0.47	179.50	**0.017**	-0.46
Specificity	5.71 (1.23)	5.00 (2.01)	4.97 (1.67)	868.50	0.451	-0.02	253.00	**0.002**	-0.53	362.00	**0.012**	-0.41
Empathy	4.80 (1.13)	4.00 (1.68)	3.65 (1.30)	903.00	0.097	-0.18	165.00	**0.000**	-0.82	575.00	**0.002**	-0.45

the interaction, the VA administered the admission interview, and participants responded to the questions and prompts from the VA. The VA also presented the filled PAF form to the participants at the end of the interview, and invited them to review the form and make any necessary corrections.

After the interaction, participants completed a post-study questionnaire to evaluate the model's performance and report their user experience. We followed up on key responses through a semi-structured interview. Additionally, we presented the original PAF questionnaire to the participants and asked them to imagine the scenario where they needed to fill out the form themselves. We invited them to compare the use of the VA to completing the PAF form manually. The entire study took approximately 50 minutes for each participant, and the interaction was recorded for further


Briefing
User Consent & Information Collection
Practice Session (w. VA)
Pre-study Interview (on syndrome)
Interaction with One VA (baseline or ours)
Questionnaire & Semi-structured Interview
Debriefing & Exit Interview
End of Study

(a) User study workflow.

(b) User study setup.

Figure 10: User study workflow and setup.

analysis with consent. The participants were compensated with a gift card for their time as a token of appreciation.

6.3 Evaluation Metrics and Hypothesis

To evaluate the performance of the two models, we adopted a set of subjective and objective evaluation metrics. Their mappings to the design requirements are shown in Figure 7. In particular, the subjective evaluation metrics (7-point Likert Scale Ratings with 1 being the lowest score) include the following (detail questionnaire items are attached in supplementary materials):

(1) **Communication Quality**. To evaluate communication quality between patients and healthcare providers, we adopted the corresponding section in the Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey [15, 73] (Q1.1 to Q1.6 in Figure 11a).

(2) **Information Quality**. We extended the aforementioned Gricean Maxims metrics (Table 4) used in the technical evaluation to evaluate the quality of conversation-level language generation with the following modifications (Q3.1 to Q3.7 in Figure 11a). We replaced "Empathy" with "Manners" metric to evaluate if the VA was asking answerable questions and do not push the user to provide information that they are not willing to share. The original empathy metric was removed because it was duplicated with the user's subjective feelings metric.

(3) **User's Subjective Feelings**. We invited the participants to rate their self-confidence in providing correct answers, engagement, and perceived performance of the agent according to metrics by Li et al. [55] and Abd-Alrazaq [1] (Q2.1 to Q2.8 in Figure 11a).

For objective measures, we verified the completeness and accuracy of the information collected by the VA through the admission interview by counting the number of corrections the participants made on the filled PAF form after interaction. We also computed the number of words per response and the number of conversation turns.

To conclude with, we derived the following hypothesis:

H1 Compared to the baseline model, our system's scaffolding module will improve the quality of information provided to the user. Our model can provide more informative (Q3.3, *H1a*), clear (Q3.4, *H1b*), easier to answer (Q3.5, *H1c*), factual correct (Q3.6, *H1d*), and specific (Q3.7, *H1e*) responses. User can also feel they can get an answer to their questions (Q1.1, *H1f*), and the response is clearer (Q1.2, *H1g*).

H2 Compared to the baseline model, our system's empathy and orientation mechanism can improve the communication experience with the user. Our model present better abilities in attentive listening (Q1.3, *H2a*), showing respect to the patient (Q1.5, *H2b*), and spend enough time with the patient (Q1.6, *H2c*). User are also more willing to engage in the conversation (Q2.2, *H2d*), can share their feelings more comfortably (Q2.3, *H2e*), and feel the conversation is more engaging (Q2.6, *H2f*).

H3 Compared to the baseline model, our system's graph-based topic management mechanism can organize the conversation more logically. Our model can provide more coherent (Q3.1, *H3a*) and relevant (Q3.2, *H3b*) responses.

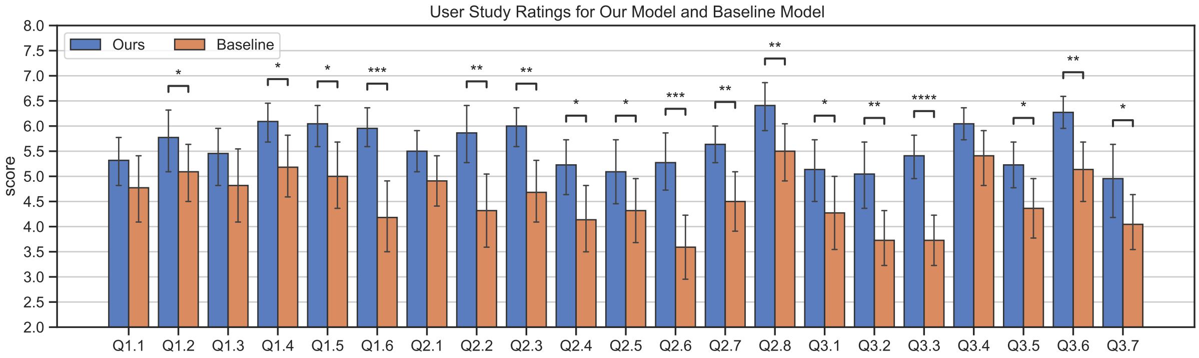
H4 Compared to baseline, our global filling & consistency check module can help the model remember the patient's medical history (Q1.4, *H4a*) and addressing conflicting answers. Users can perceive our model as more responsible (Q2.4, *H4b*), and has more medical knowledge (Q2.7, *H4c*).

H5 Compare with the baseline model, our system can capture patient's information more accurately. Our model can provide more complete (*H5a*) and accurate (*H5b*) information.

6.4 Results from User Study

In this section, we discuss the user study results based on the evaluation metrics and hypotheses proposed in Section 6.3.

We first verified within-module consistency for each hypothesis's measures using Cronbach's alpha. Results showed acceptable



(a) The users' ratings on the overall performance of the two models.

(b) Mann-Whitney U test results for the evaluation metrics. Within module consistency was calculated using Cronbach's alpha. Comparison of the two models was performed using Mann-Whitney U tests. The U-val, p-value, and effect size are reported for each evaluation metric. Effect size was calculated using Rank-Biseral Correlation. We noted the significant differences in bold. We marked the effect sides in Small (S), Medium (M), and Large (L) with the recommended cut-offs: 0.10 -< 0.3 (S), 0.30 -< 0.50 (M), >= 0.5 (L).

Module	Q_id	Metric	Experiment Mean(S.D.)	Baseline Mean(S.D.)	U	p-value	Effect Size	Hypothesis
Scaffolding	Q3.3	Informativeness	5.41 (1.01)	3.73 (1.32)	80.00	**0.000**	0.67 (L)	H1a Acc.
	Q3.4	Clarity	6.05 (0.79)	5.41 (1.33)	178.50	0.061	0.26 (S)	H1b Rej.
	Q3.5	Manner	5.23 (1.11)	4.36 (1.40)	152.00	**0.016**	0.37 (M)	H1c Acc.
	Q3.6	Factual Correctness	6.27 (0.70)	5.14 (1.42)	122.50	**0.002**	0.49 (M)	H1d Acc.
	Q3.7	Specificity	4.95 (1.79)	4.05 (1.40)	150.50	**0.015**	0.38 (M)	H1e Acc.
	Q1.1	Answer Quality	5.32 (1.17)	4.77 (1.57)	201.50	0.164	0.17 (S)	H1f Rej.
	Q1.2	Explanation Clarity	5.77 (1.48)	5.09 (1.44)	164.00	**0.030**	0.32 (M)	H1g Acc.
Empathy & Orientation	Q1.3	Attentive Listening	5.45 (1.37)	4.82 (1.79)	194.00	0.126	0.20 (S)	H2a Rej.
	Q1.5	Respect	6.05 (1.00)	5.00 (1.54)	148.00	**0.012**	0.39 (M)	H2b Acc.
	Q1.6	Time Investment	5.95 (0.90)	4.18 (1.59)	89.00	**0.000**	0.63 (L)	H2c Acc.
	Q2.2	Engagement Willingness	5.86 (1.36)	4.32 (1.76)	122.50	**0.002**	0.49 (M)	H2d Acc.
	Q2.3	Comfort Sharing	6.00 (0.98)	4.68 (1.52)	122.00	**0.002**	0.50 (M)	H2e Acc.
	Q2.6	Conversation Quality	5.27 (1.39)	3.59 (1.59)	108.00	**0.001**	0.55 (L)	H2f Acc.
Topic Management	Q3.1	Coherence	5.14 (1.52)	4.27 (1.78)	172.50	**0.050**	0.29 (S)	H3a Acc.
	Q3.2	Relevance	5.05 (1.53)	3.73 (1.32)	117.50	**0.001**	0.51 (L)	H3b Acc.
Consistency Check	Q1.4	History Recall	6.09 (0.97)	5.18 (1.47)	152.50	**0.015**	0.37 (M)	H4a Acc.
	Q2.4	Responsibility	5.23 (1.31)	4.14 (1.67)	146.00	**0.011**	0.40 (M)	H4b Acc.
	Q2.7	Medical Knowledge	5.64 (0.95)	4.50 (1.50)	130.50	**0.004**	0.46 (M)	H4c Acc.
Overall	Q2.1	Confidence	5.50 (0.96)	4.91 (1.15)	180.50	0.066	0.25 (S)	–
	Q2.5	Understanding	5.09 (1.51)	4.32 (1.55)	173.00	**0.050**	0.29 (S)	–
	Q2.8	Language Proficiency	6.41 (1.14)	5.50 (1.47)	143.50	**0.006**	0.41 (M)	–

Figure 11: User's ratings on the overall performance of the two models.

internal consistency for all modules except graph-based topic management:

- Scaffolding: 0.77 (95% CI: [0.642, 0.858])
- Empathy and orientation: 0.84 (95% CI: [0.751, 0.902])
- Graph-based topic management: 0.53 (95% CI: [0.139, 0.744])
- (Global filling and) Consistency check: 0.70 (95% CI: [0.511, 0.828])

The lower consistency in the graph-based topic management module may be due to the limited number of qualitative questions assessing this module.

We then conducted Mann-Whitney U tests to compare the performance of our system against the baseline model across three sets of evaluation metrics (see Section 6.3). Results are presented in Figure 11a and Figure 11b. Our system outperformed the baseline in all metrics, with significant differences in 17 out of 21 measures.

We report test statistics (U-value), p-value, and effect size[3] for each metric, following best practices [76, 85, 88].

6.4.1 Task Fulfillment and Overall Feedback. All participants successfully completed the simulated hospital admission interview, except for P37, who triggered the emergency response mechanism by mentioning severe cough with blood in their sputum. In this case, the agent appropriately halted the conversation and reported the emergency.

While both systems showed comparable performance in conversation metrics (turns, delay, response length), our system demonstrated significant improvements in information capture accuracy. Participants made fewer PAF corrections ($U = 65.0; p < 0.001; r = 0.73$, medium effect) and rated information accuracy higher ($U = 317.0; p = 0.024; r = 0.31$, small effect) with our system. Detailed data are presented in Table 6, supporting our hypothesis (**H5**) about improved information capture. Our system received significantly higher ratings for "understanding user responses" and "proficiency in English" (Figure 11b, Figure 11a). Participants also reported greater confidence in providing answers, though this difference wasn't statistically significant. These results support our hypothesis (**H5**).

Qualitative feedback revealed our system's superior performance could be attributed to several factors. First, our system better identified and handled nuanced responses. For instance, when P3 responded with *"medium but uh"* regarding headache pain, the system detected the hesitation and followed up: *"Okay, thank you. Could you describe if the headache is steady or does it come and go?"*. This attention to uncertainty, observed by P3 and six other participants, contrasted with the baseline system's tendency to accept unclear responses without further inquiry, ultimately reducing comprehension errors. This improvement can be attributed to our system's ability to detect concrete problems in patient responses beyond simple slot filling. Second, participants (P9 and five others) also noted that the baseline system unexpectedly skipped questions in the latter part of the PAF, requiring manual completion post-interview. Analysis of conversation logs revealed this likely stemmed from the baseline LLM's limited long-turn memory capacity, a problem not observed in our system's graph-based topic management approach. Participants also highlighted additional differences between the two systems, including the quality of scaffolding, communication manner that influenced their willingness to share information, and question ordering that reduced confusion. These differences are discussed in the following sections.

6.4.2 Scaffolded Turns: Enhancing Information Elicitation and User Engagement. User evaluation demonstrates that our model outperforms the baseline across all metrics, with statistically significant differences in all but the "Clarity" and "Get answer to my question" metrics (shown in Figure 11b, **H1** is generally supported). In experiments with our system, scaffolding was triggered 64 times in 20/22 conversations, where new information was successfully triggered 79.7% of the times. Among these cases, "S3 further explain questions" was the most frequently used (52.9%), followed by "S4 follow-up questions" (25.5%), "S1 Provoke memory" (11.8%), and "S2 Simplify question" (9.8%). Although both systems can respond to users' inquiries to keywords and provide clear explanations, accounting for the similarity in these two metrics, the baseline system tends to use more medical terms that often require users to seek further clarification or conduct online searches (P24, P40, P44). Moreover, P12 and 7 others reported that the lack of follow-up abilities in baseline system make them feel like the CA is just "rush you through the questions". After discovering this they tends to flash through the interview with simple "yes, no" answers.

In contrast, our system excels in providing more informative, specific, and factually correct responses while maintaining a more appropriate conversational demeanor. A key advantage of our system lies in its ability to ask pertinent follow-up questions, which users frequently praised. These questions help them to recall the more details (reported by P18 and 6 others), especially helpful for recalling time and degree related information. The scaffolded turns implemented in our system play a crucial role in this enhanced performance. By systematically guiding users through the interview process and probing for additional details when necessary, our system creates a more engaging and productive dialogue.

6.4.3 Empathy and Orientation: Fostering Trust and Emotional Support. Our study also investigated differences in users' emotional responses when interacting with the two agents. Results show that our system outperformed the baseline in the "Empathy and Orientation" module, with significant differences in all metrics except for "Attentive Listening" (see Figure 11b). The similarity in "Attentive Listening" scores likely stems from both systems sharing the same ASR and TTS modules. These findings support **H2**.

Patients in hospital settings often experience negative emotions such as anxiety, irritation, and pain. While some participants noted that the TTS's plain tone limited its ability to convey support, our model still provided more empathetic responses in terms of content. This contrasts sharply with the baseline agent, which, as P44 observed, *"did not provide any empathetic support in the whole process,"* leaving users feeling as if they were merely *"filling out a questionnaire."*

Our system's approach to emotional support and engagement created a more positive experience for users. It respected their responses and offered useful emotional support to alleviate negative feelings. P19 summarized this effect, stating, *"The agent gives me a unique experience and improves my mood in the hospital."* This empathetic interaction fostered a trustworthy environment between the patient and the virtual care provider, encouraging users to share more details about their conditions openly and comfortably.

6.4.4 Graph-based Topic Management: Enhancing Conversational Flow and Adaptability. Our design of the PAF Questionnaire Graph and Switching Strategy, as described in Sections 4.2.2 and 4.2.4, aims to construct a logical and fluent conversational flow. This feature was triggered 209 times across 22 experiments with our system - 56.3% for "T2 stay on the topic", 33.0% for "T1 Adjust question orders", 9.6% for "T4 Cross validation", and 1 time for "T3 Gracefully exit" (P37). The effectiveness of this approach is evident in our system's performance, which provided significantly more "logically coherent" and "relevant" responses compared to the baseline model (shown in Figure 11b). These results support **H3**.

[3] Using Rank-Biserial Correlation for effect size. Recommended cut-offs: 0.10 -< 0.3 (small), 0.30 -< 0.50 (moderate), >= 0.5 (large)

Table 6: Comparison of the two models in terms of the number of turns, response time, and response length, as well as the number of corrections and subjective accuracy. The mean (S.D.) values are reported for each metric. All the metrics, except for "Response Delay" were tested with Mann-Whitney U tests as they are not normally distributed. "Response Delay" was tested with a t-test. Metrics with significant differences are highlighted in bold.

Model	# Turns	Response Delay	Response Length	# Corrections	Subjective Accuracy
Ours	22.68 (5.92)	1.96 (0.37)	19.22 (1.26)	**0.41 (0.58)**	**4.36 (0.48)**
Baseline	23.00 (7.65)	1.86 (0.34)	10.73 (2.09)	2.55 (1.67)	3.82 (0.94)

User feedback indicates widespread appreciation for our agent's topic transitions, which were described as clear and easily understood. The system's ability to dynamically adjust question ordering and selection based on users' described symptoms was particularly well-received. For instance, when P20 explained a fall experience, our agent promptly switched to the fall question block to gather more detailed information. In contrast, users frequently highlighted limitations in the baseline agent's performance. Many participants (P9, P10, P22, among others) expressed a need for improved transitions in the baseline system. Others (P12, P25, P34) pointed out confusing logical flows, further emphasizing the advantages of our graph-based approach. The Graph-based Topic Management not only enhances the logical coherence of the conversation but also allows for a more patient-centered interview process. By adapting to user inputs and prioritizing relevant topics, our system creates a more engaging and efficient interview experience. This adaptive questioning flow represents a significant improvement over traditional linear questionnaires, potentially leading to more comprehensive and accurate patient information gathering during hospital admissions.

6.4.5 Global Consistency Check: Enhancing Information Accuracy and User Confidence. The answer filling and consistency check module, detailed in Section 4.2.3, demonstrates our system's ability to capture and validate user-provided information effectively. This feature significantly enhances the quality and reliability of the data collected during the interview process. Our system's capability to record and appropriately file additional information mentioned by patients is exemplified in P41's interaction. When P41 mentioned fall history and use of ambulatory aids during the pain question block, our agent recorded this key information and adjusted the subsequent questions accordingly. This adaptive approach ensures a more efficient and less repetitive interview experience. Furthermore, our agent's ability to identify and address contradictions in patients' responses adds another layer of data validation. P43 praised this feature, stating, *"I like the validation part, which asks for a clarification of my sputum color. I think it was good that the agent could understand this conflict and ask further, rather than ignoring and directly recording."* This approach not only improves data accuracy but also instills confidence in users about the thoroughness of the interview process. The effectiveness of these features is reflected in our user study results. Our system significantly outperformed the baseline in metrics such as "Remember the patient's medical history," "Responsibility," and "Medical knowledge" (see Figure 11b). These findings support **H4**. By implementing these consistency checks and adaptive information gathering techniques, our system demonstrates a level of professionalism, responsibility, and medical knowledge that users find reassuring. This enhancement contributes to a more trustworthy and efficient hospital admission interview process, potentially improving both the quality of collected data and the patient experience.

7 Discussion

In this section, we first summarize some of the benefits and limitations participants perceived when interacting with our system. We then reflect on how our findings connect to the idea of patient-centered care and the existing hospitalization process to suggest future design considerations. To conclude, we propose ideas on how to better situate hospital admission interview CAs into patients' clinical experiences and discuss insights that may generalize to other information-gathering CAs.

7.1 Balancing Task Efficiency and Patient-Centered Care

Our research focused on the hospital admission interview process, a critical communication point that helps establish caring relationships between patients and healthcare providers [23, 28]. Due to concerns about AI systems providing improper medical recommendations [5, 21, 35], deliberately limited our system's scope to information collection and descriptive documentation per the PAF form. Drawing from patient-centered care principles [67], we implemented scaffolding and empathy mechanisms, such as open-ended follow-up questions and examples, to encourage detailed health condition descriptions while focusing on listening rather than diagnosing.

Our study revealed divergent expectations between nurses and patients regarding the system's role. Nurses advocated for a task-oriented approach, expecting the system to *"collect 100% correct and complete information"* and *"stick to the questionnaire."* In response, we designed our CA to maintain focus on the structured questionnaire. Patients, however, showed greater engagement with their syndrome-related questions, often providing rich contextual details about symptom timing, sensations, and perceived causes beyond the PAF form's scope. While both CAs allowed brief narrative exploration before guiding users back to the questionnaire, some patients (notably P18 and P28) expressed preference for deeper exploration of primary symptoms rather than transitioning to seemingly unrelated topics like food preferences.

This tension between task efficiency and patient-centered care presents both challenges and opportunities. While nurses prioritize systematic data collection, patients value opportunities to share their health narratives. However, these perspectives can be complementary rather than contradictory. Detailed patient narratives can help nurses identify potential discrepancies, develop comprehensive understanding, and gather valuable diagnostic information. The CA's advantage lies in its ability to dedicate more time to patient listening while maintaining structured data collection. By setting clear expectations and providing space for narrative sharing, the system can enhance patient comfort and understanding [21, 57].

Future work could explore enriching PAF topics with additional symptom-related questions or implementing dynamic, database-driven follow-up questions based on user input, potentially achieving a better balance between structured data collection and patient-centered care.

7.2 Challenges in Humanizing CA Interactions

In designing our CA, we aimed to humanize it to potentially increase user trust and improve self-disclosure intentions [54]. This humanization effort focused on two aspects: enhancing the CA's proficiency and incorporating more human-like elements into the conversation. We improved proficiency by mimicking the language abilities of human nurses, as identified in our formative study. To add a human touch, we implemented empathy mechanisms, acknowledging user feelings and providing emotional support, following designs from previous works [14, 57, 79]. While participants generally received the CA's proficiency well, the human touch aspect fell short of expectations. We identified several challenges contributing to this shortfall. Technical limitations, such as inadequate ASR accuracy for detecting users' emotional needs (P8, P11) and unexpressive TTS failing to convey empathy effectively (P10, P13), hindered the CA's ability to engage empathetically. Our task-oriented design, focused on mimicking real-world nurses' behavior, may have compromised the human touch. Unlike human nurses constrained by time, the CA could potentially spend more time with users, allowing for deviations from the questionnaire to collect additional relevant information. Additionally, the CA's empathy mechanisms could be extended. For instance, providing actionable advice to alleviate discomfort, as suggested by P43 (a rehabilitation therapy expert), could make users perceive the CA as more helpful and caring. This approach differs from diagnosis by offering simple suggestions or feedback to manage uncomfortable symptoms. Future work could address these challenges by exploring techniques to boost context-related words in ASR, or fine-tuning TTS tone to enhance empathy conveyance. By addressing these challenges, future iterations of the CA could strike a better balance between task efficiency and human-like interaction, potentially improving user experience and data collection in hospital admission interviews.

7.3 Privacy Concern and System Adaptability

While our experimental results demonstrate the potential of CAs in admission interviews, several gaps need to be bridged for clinical implementation. Transitioning from experimental setup to healthcare environments requires addressing both privacy regulations and system adaptability. Healthcare privacy regulations necessitate robust security mechanisms beyond our current cloud-based implementation. Our modular system architecture, which abstracts patient support into problem-strategy mappings, offers a compelling solution through local LLM deployment. By hosting competent open-sourced models on hospital servers, we can maintain functionality while processing data locally, addressing privacy concerns without compromising performance. Our experimental results demonstrate that the system can achieve consistent or better performance when using the same underlying LLM architecture, suggesting that local deployment is viable. Furthermore, this modular design facilitates transitions between different LLM models without compromising the core interview structure, provided the new model can identify user challenges and generate strategy-aligned responses. While alternative approaches like homomorphic encryption (HE) with commercial LLMs exist [18], they face significant challenges including computational overhead and limited support for complex architectures [16, 93]. These findings suggest that local LLM deployment, combined with our modular architecture, might presents the most promising path forward for implementing privacy-preserving CAs in healthcare settings while maintaining system flexibility and performance.

7.4 Generalizability to Other Patient Interview Tasks

Hospital admission interviews represent just one of many patient interview tasks in the healthcare system. Other common tasks include but are not limited to oral based NCD screening [17, 29], pre-consultation interviews [55], and discharge interviews [13, 50]. The generalizability of our system to these tasks can be examined from two perspectives: question structure and language support.

Concerning question structure, at the question level, the five blocks selected from the PAF form include both closed-ended and open-ended questions and cover common question types [37]. It suggests that our system can handle commonly used question types in questionnaires. Additionally, our system allows for human suggested "sample questions" for each topic, providing flexibility for customization in specific domains. At the questionnaire level, our system employs structure-wise parsing, allowing the interview to include any number of topic blocks with predefined transition logic or order if needed.

Regarding language support, while different tasks may require varying language capabilities due to their distinct goals and requirements, they share commonalities with the hospital admission interview task. All these tasks require a certain degree of scaffolding, such as follow-up questions, to guide patients through the interview process [29, 55, 59]. We acknowledge that the hospital admission interview task might be more familiar to our participants compared to pre-consultation interviews [55] and does not require interaction with a complex system like NCD screening [29]. Consequently, the language capabilities required for different medical interviews may vary. However, the general principles of providing scaffolding turns, organizing logical conversations, and showing empathy to maintain a smooth conversation flow remain essential across tasks [29, 55, 59]. Unique scaffolding needs, such as disambiguating the role of the interview [55] or explaining system functions [29], could be incorporated as parallel language generation modules in our system. This modular approach we employed allows for task-specific customization while maintaining the core functionality of the interview system.

7.5 Limitations and Future Work

Our system has several limitations that need to be addressed in future work. First, our study participants were recently-recovered patients who had experienced hospitalization or required emergency care. While we asked them to recall their illness experiences and imagine themselves in the hospitalization process, their recollections may be imperfect, and their current emotional states likely differ from those during actual hospitalization [40]. Due to hospital regulations, we were unable to test our system with currently hospitalized patients due to privacy concerns. Although we adopted various privacy protection techniques, our current implementation still involves transmitting data to a third party.This violates our collaborating hospital's regulations on patient data privacy, especially when discussing sensitive health information. We attempted to build our system with a local LLM, but the performance was not satisfactory, as described in our Design section. Moreover, ethical concerns arise when testing a CA in early development with real patients [9]. To address these challenges, we chose to use co-design [8] and participatory design [65] approaches, involving different stakeholders at various stages of the design process. Future work may consider developing local LLMs with sufficient performance or adding more advanced privacy-protection mechanisms when using commercial LLMs. Additionally, future research at a later stage of design should focus on establishing deeper collaborations with hospitals to conduct studies involving actual patients, ensuring a healthcare professional's presence to supervise the interaction and maintain patient safety.

Second, the CA currently relies solely on user narratives for information acquisition. However, users may not always provide accurate information due to privacy concerns, lack of knowledge, or simply forgetfulness [70, 81, 87]. For instance, some participants in our study reported normal eyesight despite wearing prescription glasses, not considering it a vision problem. Such discrepancy is hard to detect by the CA, as it only analyses narratives and lacks the ability look at the user's face. Future work could explore integrating additional sensors into the CA to collect more objective, multimodal information for response triangulation, such as cameras to detect facial expressions or microphones to analyze voice tone.

Third, the potential for collaboration between the CA and nurses remains underexplored. In our study, the CA's role was limited to completing the PAF in an EHR-compatible format and providing descriptive notes to nurses. Besides, we did not considered existing information about patient in the EHR system as well as those information known by the nurses who help to admit the patient. Although getting such information could benefit the CA to provide more tailored care to users, currently we cannot access EHR information due to regulations of the collaborating hospital and Hospital Authority. Letting nurses input such information manually is not feasible due to the time constraints of the hospitalization process. However, as suggested by [39], having long term information could benefit the admission process by further enhancing user engagement and self-disclosure. Future work may consider keep a separate record of the CA's conversation with a patient, and include such information as prefilled nodes in the graph records in our proposed pipeline. However, future research are still require to investigate the best way to integrate the CA into the hospitalization process and enhance its ability to directly support front-line healthcare workers.

8 Conclusion

In this paper, we investigated the communication skills and strategies employed by healthcare professionals during hospital admission interviews. Based on these insights, we designed and implemented an LLM-powered CA to replicate these skills and strategies. Specifically, we proposed a novel approach to address these challenges by abstracting the essential communication skills through a set of challenge-strategies mappings and integrating them into an LLM to generate context-aware responses.We evaluated our system through both technical assessments and user studies, comparing it against a prompt-engineering based baseline and, in the technical evaluation, human-written ground truth data. Our results demonstrated that, with proper system design, the LLM-powered CA can achieve performance similar to or surpassing human performance at the turn level. Moreover, it consistently outperformed the prompt-engineering based baseline across all evaluation metrics. Our work contributes to the growing body of knowledge on designing CAs for medical interviews. It provides valuable insights for future research in this critical area of healthcare technology. By bridging the gap between human expertise and artificial intelligence in medical communications, we hope to improve the efficiency and quality of patient care while maintaining the essential human elements of empathy and understanding.

Acknowledgments

We would like to express our sincere gratitude to all the nurses, clinicians and hospital volunteers who generously contributed their time and expertise to this research project. We also thank the anonymous reviewers for their valuable feedback and constructive suggestions. The work described in this paper was partially supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Reference Number: AoE/E-601/24-N and T45-407/19N).

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 (June 2020), e18301. https://doi.org/10.2196/18301

[2] Ho Seok Ahn, Wesley Yep, Jongyoon Lim, Byeong Kyu Ahn, Deborah L. Johanson, Eui Jun Hwang, Min Ho Lee, Elizabeth Broadbent, and Bruce A. MacDonald. 2019. Hospital Receptionist Robot v2: Design for Enhancing Verbal Interaction with Social Skills. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1–6. https://doi.org/10.1109/RO-MAN46459.2019.8956300

[3] Zahraa Al-Hilli, Ryan Noss, Jennifer Dickard, Wei Wei, Anna Chichura, Vincent Wu, Kayla Renicker, Holly J. Pederson, and Charis Eng. 2023. A Randomized Trial Comparing the Effectiveness of Pre-test Genetic Counseling Using an Artificial Intelligence Automated Chatbot and Traditional In-person Genetic Counseling in Women Newly Diagnosed with Breast Cancer. *Ann Surg Oncol* 30, 10 (Oct. 2023), 5990–5996. https://doi.org/10.1245/s10434-023-13888-4

[4] John R. Anderson. 1983. A Spreading Activation Theory of Memory. *Journal of Verbal Learning and Verbal Behavior* 22, 3 (June 1983), 261–295. https://doi.org/10.1016/S0022-5371(83)90201-3

[5] 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 (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376718

[6] J. Edwin Benton and John L. Daly. 1993. Measuring Citizen Evaluations: The Question of Question Order Effects. *Public Administration Quarterly* 16, 4 (1993), 492–508. jstor:40861564

[7] Jordan L. Boyd-Graber, Sonya S. Nikolova, Karyn A. Moffatt, Kenrick C. Kin, Joshua Y. Lee, Lester W. Mackey, Marilyn M. Tremaine, and Maria M. Klawe. 2006. Participatory Design with Proxies: Developing a Desktop-PDA System to Support People with Aphasia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. Association for Computing Machinery, New York, NY, USA, 151–160. https://doi.org/10.1145/1124772.1124797

[8] Ingrid Burkett. 2016. An Introduction to Co-Design. (2016).

[9] Danton S. Char, Michael D. Abràmoff, and Chris Feudtner. 2020. Identifying Ethical Considerations for Machine Learning Healthcare Applications. *Am J Bioeth* 20, 11 (Nov. 2020), 7–17. https://doi.org/10.1080/15265161.2020.1819469

[10] Jennifer A. Cleland, Keiko Abe, and Jan-Joost Rethans. 2009. The Use of Simulated Patients in Medical Education: AMEE Guide No 42. *Med Teach* 31, 6 (June 2009), 477–486. https://doi.org/10.1080/01421590903002821

[11] Google Cloud. n.d.. Speech-to-Text. https://cloud.google.com/speech-to-text/?hl=en Accessed: 2024-09-12.

[12] Google Cloud. n.d.. Text-to-Speech. https://cloud.google.com/text-to-speech?hl=en Accessed: 2024-09-12.

[13] Berengere Couturier, Fabrice Carrat, Gilles Hejblum, and SENTIPAT Study Group. 2015. Comparing Patients' Opinions on the Hospital Discharge Process Collected With a Self-Reported Questionnaire Completed Via the Internet or Through a Telephone Survey: An Ancillary Study of the SENTIPAT Randomized Controlled Trial. *J Med Internet Res* 17, 6 (June 2015), e158. https://doi.org/10.2196/jmir.4379

[14] Karl Daher, Jacky Casas, Omar Abou Khaled, and Elena Mugellini. 2020. Empathic Chatbot Response for Medical Assistance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20)*. Association for Computing Machinery, New York, NY, USA, 1–3. https://doi.org/10.1145/3383652.3423864

[15] Charles Darby, Ron D Hays, and Phillip Kletke. 2005. Development and Evaluation of the CAHPS® Hospital Survey. *Health Serv Res* 40, 6 Pt 2 (Dec. 2005), 1973–1976. https://doi.org/10.1111/j.1475-6773.2005.00490.x

[16] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2024. Security and Privacy Challenges of Large Language Models: A Survey. https://doi.org/10.48550/arXiv.2402.00888 arXiv:2402.00888 [cs]

[17] Zijian Ding, Jiawen Kang, Tinky Oi Ting HO, Ka Ho Wong, Helene H Fung, Helen Meng, and Xiaojuan Ma. 2022. TalkTive: A Conversational Agent Using Backchannels to Engage Older Adults in Neurocognitive Disorders Screening. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3502005

[18] Nathan Dowlin, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. 2016. CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16)*. JMLR.org, New York, NY, USA, 201–210.

[19] Dagrunn Nåden Dyrstad, Ingelin Testad, and Marianne Storm. 2015. Older Patients' Participation in Hospital Admissions through the Emergency Department: An Interview Study of Healthcare Professionals. *BMC Health Serv Res* 15 (Oct. 2015), 475. https://doi.org/10.1186/s12913-015-1136-1

[20] Hugging Face. n.d.. Qwen-2-7B. https://huggingface.co/Qwen/Qwen2-7B Accessed: 2024-09-12.

[21] Xiangmin Fan, Daren Chao, Zhan Zhang, Dakuo Wang, Xiaohua Li, and Feng Tian. 2021. Utilization of Self-Diagnosis Health Chatbots in Real-World Settings: Case Study. *J Med Internet Res* 23, 1 (Jan. 2021), e19928. https://doi.org/10.2196/19928

[22] Arthur W. Frank. 1995. *The Wounded Storyteller: Body, Illness, and Ethics*. University of Chicago Press, Chicago.

[23] Misbah N. Ghazanfar, Per Hartvig Honoré, Trine R. H. Nielsen, Stig E. Andersen, and Mette Rasmussen. 2012. Hospital Admission Interviews Are Time-Consuming with Several Interruptions. *Dan Med J* 59, 12 (Dec. 2012), A4534.

[24] HP Grice. 1975. Logic and Conversation. *Cole and Morgarn* (1975).

[25] Jonathan Grudin and Richard Jacques. 2019. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300439

[26] Greg Guest, Kathleen M. MacQueen, and Emily E. Namey. 2012. *Applied Thematic Analysis*. SAGE.

[27] Pamela Herd and Donald Moynihan. 2021. Health Care Administrative Burdens: Centering Patient Experiences. *Health Serv Res* 56, 5 (Oct. 2021), 751–754. https://doi.org/10.1111/1475-6773.13858

[28] Ida E. Højskov and Stinne Glasdam. 2014. Transformation of Admission Interview to Documentation for Nursing Practice. *Scandinavian Journal of Caring Sciences* 28, 3 (2014), 478–485. https://doi.org/10.1111/scs.12071

[29] Jiaxiong Hu, Junze Li, Yuhang Zeng, Dongjie Yang, Danxuan Liang, Helen Meng, and Xiaojuan Ma. 2024. Designing Scaffolding Strategies for Conversational Agents in Dialog Task of Neurocognitive Disorders Screening. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3613904.3642960

[30] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. https://doi.org/10.48550/arXiv.2311.05232 arXiv:2311.05232 [cs]

[31] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in Building Intelligent Open-domain Dialog Systems. arXiv:1905.05709 [cs]

[32] Eui Jun Hwang, Bruce A. Macdonald, and Ho Seok Ahn. 2019. End-to-End Dialogue System with Multi Languages for Hospital Receptionist Robot. In *2019 16th International Conference on Ubiquitous Robots (UR)*. 278–283. https://doi.org/10.1109/URAI.2019.8768694

[33] Apple Inc. n.d.. Siri. https://www.apple.com/siri/ Accessed: 2024-09-12.

[34] Azra Ismail and Neha Kumar. 2021. AI in Global Health: The View from the Front Lines. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–21. https://doi.org/10.1145/3411764.3445130

[35] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445385

[36] Inger Jansson, Ewa Pilhammar, and Anna Forsberg. 2009. Obtaining a Foundation for Nursing Care at the Time of Patient Admission: A Grounded Theory Study. *Open Nurs J* 3 (Aug. 2009), 56–64. https://doi.org/10.2174/1874434600903010056

[37] Yucheng Jin, Li Chen, Xianglin Zhao, and Wanling Cai. 2024. The Way You Assess Matters: User Interaction Design of Survey Chatbots for Mental Health. *International Journal of Human-Computer Studies* 189 (Sept. 2024), 103290. https://doi.org/10.1016/j.ijhcs.2024.103290

[38] 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 (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3581503

[39] Eunkyung Jo, Yuin Jeong, Sohyun Park, Daniel A. Epstein, and Young-Ho Kim. 2024. Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3613904.3642420

[40] Doerte U. Junghaenel, Joan E. Broderick, Stefan Schneider, Cheng K. F. Wen, Hio Wa Mak, Sarah Goldstein, Marilyn Mendez, and Arthur A. Stone. 2021. Explaining Age Differences in the Memory-Experience Gap. *Psychol Aging* 36, 6 (Sept. 2021), 679–693. https://doi.org/10.1037/pag0000628

[41] Vera Kalitzkus and Peter F Matthiessen. 2009. Narrative-Based Medicine: Potential, Pitfalls, and Practice. *Perm J* 13, 1 (2009), 80–86.

[42] Fatemeh Karami, Alireza Nikbakht Nasrabadi, Camellia Torabizadeh, Monir Mazaheri, and Leila Sayadi. 2024. The Challenges of Voluntary Care Provision for Hospitalized Patients with COVID-19: A Qualitative Study of the Public Volunteers' Experiences. *Health Expect* 27, 2 (April 2024), e13998. https://doi.org/10.1111/hex.13998

[43] Eleni Karasouli, Daniel Munday, Cara Bailey, Sophie Staniszewska, Alistair Hewison, and Frances Griffiths. 2016. Qualitative Critical Incident Study of Patients' Experiences Leading to Emergency Hospital Admission with Advanced Respiratory Illness. *BMJ Open* 6, 2 (Feb. 2016), e009030. https://doi.org/10.1136/bmjopen-2015-009030

[44] Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23)*. Association for Computing Machinery, New York, NY, USA, 115–135. https://doi.org/10.1145/3563657.3595996

[45] 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 (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300316

[46] Arthur Kleinman. 1988. *The Illness Narratives: Suffering, Healing, and the Human Condition*. Basic Books, New York, NY, US. xviii, 284 pages.

[47] Rafal Kocielnik, Elena Agapie, Alexander Argyle, Dennis T Hsieh, Kabir Yadav, Breena Taira, and Gary Hsieh. 2020. HarborBot: A Chatbot for Social Needs Screening. *AMIA Annu Symp Proc* 2019 (March 2020), 552–561.

[48] Irma P.M. Kruijver, Ada Kerkstra, Jozien M. Bensing, and Harry B.M. Van De Wiel. 2001. Communication Skills of Nurses during Interactions with Simulated Cancer Patients. *Journal of Advanced Nursing* 34, 6 (2001), 772–779. https://doi.org/10.1046/j.1365-2648.2001.01807.x

[49] H. S. Lau, C. Florax, A. J. Porsius, and A. De Boer. 2000. The Completeness of Medication Histories in Hospital Medical Records of Patients Admitted to

General Internal Medicine Wards. *Br J Clin Pharmacol* 49, 6 (June 2000), 597–603. https://doi.org/10.1046/j.1365-2125.2000.00204.x

[50] Kristin Laugaland, Karina Aase, and Paul Barach. 2012. Interventions to Improve Patient Safety in Transitional Care–a Review of the Evidence. *Work* 41 Suppl 1 (2012), 2915–2924. https://doi.org/10.3233/WOR-2012-0544-2915

[51] Colleen Doherty Lauster and Sneha Baxi Srivastava. 2013. *Fundamental Skills for Patient Care in Pharmacy Practice*. Jones & Bartlett Publishers.

[52] Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med* 388, 13 (March 2023), 1233–1239. https://doi.org/10.1056/NEJMsr2214184

[53] Sunghee Lee and David Grant. 2009. The Effect of Question Order on Self-rated General Health Status in a Multilingual Survey Context. *American Journal of Epidemiology* 169, 12 (June 2009), 1525–1530. https://doi.org/10.1093/aje/kwp070

[54] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376175

[55] 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 (CHI '24)*. ACM, Honolulu HI USA, 1–24. https://doi.org/10.1145/3613904.3641913

[56] 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 (CHI EA '24)*. ACM, Honolulu HI USA, 1–10. https://doi.org/10.1145/3613905.3651891

[57] Shuya Lin, Lingfeng Lin, Cuiqin Hou, Baijun Chen, Jianfeng Li, and Shiguang Ni. 2023. Empathy-Based Communication Framework for Chatbots: A Mental Health Chatbot Application and Evaluation. In *International Conference on Human-Agent Interaction*. ACM, Gothenburg Sweden, 264–272. https://doi.org/10.1145/3623809.3623865

[58] Bingjie Liu and S. Shyam Sundar. 2018. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot. *Cyberpsychol Behav Soc Netw* 21, 10 (Oct. 2018), 625–636. https://doi.org/10.1089/cyber.2018.0110

[59] Dingdong Liu, Sensen Gao, Zixin Chen, Yifan Shen, Chuhan Shi, Bertram E. Shi, and Xiaojuan Ma. 2024. Exploring Scaffolding Techniques for Agent-Administered Brief Cognitive Screening in Hospital Settings. In *Companion Publication of the 2024 ACM Designing Interactive Systems Conference (DIS '24 Companion)*. Association for Computing Machinery, New York, NY, USA, 185–189. https://doi.org/10.1145/3656156.3663697

[60] Awakening Health Ltd. n.d.. Awakening Health. https://awakening.health Accessed: 2024-09-12.

[61] Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2025. User Interaction Patterns and Breakdowns in Conversing with LLM-Powered Voice Assistants. *International Journal of Human-Computer Studies* 195 (Jan. 2025), 103406. https://doi.org/10.1016/j.ijhcs.2024.103406 arXiv:2309.13879 [cs]

[62] 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 (ICIET 2019)*. Association for Computing Machinery, New York, NY, USA, 111–119. https://doi.org/10.1145/3323771.3323824

[63] Thorsten Meyer, Ruth Deck, and Heiner Raspe. 2007. Problems Completing Questionnaires on Health Status in Medical Rehabilitation Patients. *J Rehabil Med* 39, 8 (Oct. 2007), 633–639. https://doi.org/10.2340/16501977-0098

[64] Lisa Mikesell. 2013. Medicinal Relationships: Caring Conversation. *Med Educ* 47, 5 (May 2013), 443–452. https://doi.org/10.1111/medu.12104

[65] Michael J. Muller and Sarah Kuhn. 1993. Participatory Design. *Commun. ACM* 36, 6 (June 1993), 24–28. https://doi.org/10.1145/153571.255960

[66] Öncel Naldemirci, Nicky Britten, Helen Lloyd, and Axel Wolf. 2020. The Potential and Pitfalls of Narrative Elicitation in Person-Centred Care. *Health Expectations* 23, 1 (2020), 238–246. https://doi.org/10.1111/hex.12998

[67] NEJM Catalyst. 2017. What Is Patient-Centered Care? *Catalyst Carryover* 3, 1 (Jan. 2017). https://doi.org/10.1056/CAT.17.0559

[68] OpenAI. 2024. GPT-4O Language Model. https://openai.com/index/hello-gpt-4o/ Model version: GPT-4O-2024-05-13.

[69] OpenAI. n.d.. GPT-4. https://openai.com/index/gpt-4/ Accessed: 2024-09-12.

[70] Venkataraman Palabindala, Amaleswari Pamarthy, and Nageshwar Reddy Jonnalagadda. 2016. Adoption of Electronic Health Records and Barriers. *J Community Hosp Intern Med Perspect* 6, 5 (Oct. 2016), 10.3402/jchimp.v6.32643. https://doi.org/10.3402/jchimp.v6.32643

[71] Pallets Projects. n.d.. Flask Documentation. https://flask.palletsprojects.com/en/3.0.x/ Accessed: 2024-09-12.

[72] Wendy Pugh and Alison M. Porter. 2011. How Sharp Can a Screening Tool Be? A Qualitative Study of Patients' Experience of Completing a Bowel Cancer Screening Questionnaire. *Health Expectations* 14, 2 (2011), 170–177. https://doi.org/10.1111/j.1369-7625.2010.00629.x

[73] Denise D. Quigley, Steven C. Martino, Julie A. Brown, and Ron D. Hays. 2013. Evaluating the Content of the Communication Items in the CAHPS® Clinician and Group Survey and Supplemental Items with What High-Performing Physicians Say They Do. *The Patient - Patient-Centered Outcomes Research* 6, 3 (Sept. 2013), 169–177. https://doi.org/10.1007/s40271-013-0016-1

[74] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. ACM, Oslo Norway, 117–126. https://doi.org/10.1145/3020165.3020183

[75] Ellen Rosenberg. 2003. Narrative-Based Primary Care: A Practical Guide. *BMJ* 326, 7379 (Jan. 2003), 56.

[76] Tetsuya Sakai. 2017. The Probability That Your Hypothesis Is Correct, Credible Intervals, and Effect Sizes for IR Evaluation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 25–34. https://doi.org/10.1145/3077136.3080766

[77] Anna Schneider, Markus Wehler, and Matthias Weigl. 2019. Provider Interruptions and Patient Perceptions of Care: An Observational Study in the Emergency Department. *BMJ Qual Saf* 28, 4 (April 2019), 296–304. https://doi.org/10.1136/bmjqs-2018-007811

[78] Norbert Schwarz and Seymour Sudman. 1992. *Context Effects in Social and Psychological Research*. https://doi.org/10.1007/978-1-4612-2848-6

[79] Maria Seehausen, Philipp Kazzer, Malek Bajbouj, Hauke R. Heekeren, Arthur M. Jacobs, Gisela Klann-Delius, Winfried Menninghaus, and Kristin Prehn. 2016. Effects of Empathic Social Responses on the Emotions of the Recipient. *Brain and Cognition* 103 (March 2016), 50–61. https://doi.org/10.1016/j.bandc.2015.11.004

[80] Johanna Shapiro. 1993. The Use of Narrative in the Doctor-Patient Encounter. *Family Systems Medicine* 11, 1 (1993), 47–53. https://doi.org/10.1037/h0089128

[81] Nelson Shen, Thérèse Bernier, Lydia Sequeira, John Strauss, Michelle Pannor Silver, Abigail Carter-Langford, and David Wiljer. 2019. Understanding the Patient Privacy Perspective on Health Information Exchange: A Systematic Review. *International Journal of Medical Informatics* 125 (May 2019), 1–12. https://doi.org/10.1016/j.ijmedinf.2019.01.014

[82] Yifan Shen, Dingdong Liu, Yejin Bang, Ho Shu Chan, Rita Frieske, Hoo Choun Chung, Jay Nieles, Tianjia Zhang, Kien T. Pham, Wai Yi Rosita Cheng, Yini Fang, Qifeng Chen, Pascale Fung, Xiaojuan Ma, and Bertram E. Shi. 2024. A Humanoid Robot Dialogue System Architecture Targeting Patient Interview Tasks. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. 1394–1401. https://doi.org/10.1109/RO-MAN60168.2024.10731285

[83] Donghoon Shin, Soomin Kim, Ruoxi Shang, Joonhwan Lee, and Gary Hsieh. 2023. IntroBot: Exploring the Use of Chatbot-assisted Familiarization in Online Collaborative Groups. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3544548.3580930

[84] M. A. Sprangers and N. K. Aaronson. 1992. The Role of Health Care Providers and Significant Others in Evaluating the Quality of Life of Patients with Chronic Disease: A Review. *J Clin Epidemiol* 45, 7 (July 1992), 743–760. https://doi.org/10.1016/0895-4356(92)90052-o

[85] Gail M. Sullivan and Richard Feinn. 2012. Using Effect Size—or Why the P Value Is Not Enough. *J Grad Med Educ* 4, 3 (Sept. 2012), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

[86] Mariska E. te Pas, Werner G. M. M. 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 (Dec. 2020), e21982. https://doi.org/10.2196/21982

[87] Diana M. Tisnado, John L. Adams, Honghu Liu, Cheryl L. Damberg, Fang Ashlee Hu, Wen-Pin Chen, David M. Carlisle, Carol M. Mangione, and Katherine L. Kahn. 2006. Does the Concordance between Medical Records and Patient Self-Report Vary with Patient Characteristics? *Health Serv Outcomes Res Method* 6, 3 (Dec. 2006), 157–175. https://doi.org/10.1007/s10742-006-0012-1

[88] Maciej Tomczak and Ewa Tomczak. 2014. The Need to Report Effect Size Estimates Revisited. An Overview of Some Recommended Measures of Effect Size. (2014).

[89] Thiemo Wambsganss, Rainer Winkler, Matthias Söllner, and Jan Marco Leimeister. 2020. A Conversational Agent to Improve Response Quality in Course Evaluations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3382805

[90] Ziang Xiao, Michelle X. Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. 2020. If I Hear You Correctly: Building and Evaluating Interview Chatbots with Active Listening Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376131

[91] 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 (June 2020), 15:1–15:37. https://doi.org/10.1145/3381804

[92] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2 (May 2024), 73:1–73:35. https://doi.org/10.1145/3659625

[93] Itamar Zimerman, Moran Baruch, Nir Drucker, Gilad Ezov, Omri Soceanu, and Lior Wolf. 2023. Converting Transformers to Polynomial Form for Secure Inference Over Homomorphic Encryption. https://doi.org/10.48550/arXiv.2311.08610 arXiv:2311.08610 [cs]