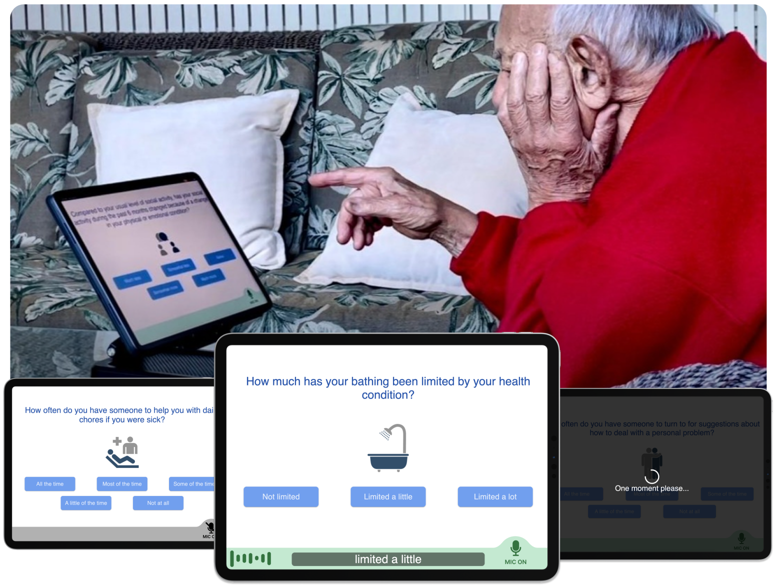
Digital Forms for All: A Holistic Multimodal Large Language Model Agent for Health Data Entry

ANDREA CUADRA, Stanford University, USA
JUSTINE BREUCH, Stanford University, USA
SAMANTHA ESTRADA, Stanford University, USA
DAVID IHIM, Stanford University, USA
ISABELLE HUNG, Monta Vista High School, USA
DEREK ASKARYAR, Stanford University, USA
MARWAN HASSANIEN, Stanford University, USA
KRISTEN L. FESSELE, Memorial Sloan Kettering Cancer Center, USA
JAMES A. LANDAY, Stanford University, USA



How often do you have someone to help you with dai chores if you were sick?

How much has your bathing been limited by your health condition?

often do you have someone to turn to for suggestions about how to deal with a personal problem?

Fig. 1. Our system, *My Care Questionnaire*, is a holistic multimodal large language model (LLM) agent for health data entry. In the large image, a participant interacts with an iPad on a stand running our system, *My Care Questionnaire*. The bottom images display different possible states of our system as the participant proceeds through the questionnaire, including a microphone muted state (bottom left), a listening state (bottom center), and a response processing state (bottom right).

Authors' addresses: Andrea Cuadra, Stanford University, Stanford, CA, USA; Justine Breuch, Stanford University, Stanford, CA, USA; Samantha Estrada, Stanford University, Stanford, CA, USA; David Ihim, Stanford University, Stanford, CA, USA; Isabelle Hung, Monta Vista High School, Cupertino, CA, USA; Derek Askaryar, Stanford University, Stanford, CA, USA; Marwan Hassanien, Stanford University, Stanford, CA, USA; Kristen L. Fessele, Memorial Sloan Kettering Cancer Center, New York, NY, USA; James A. Landay, Stanford University, Stanford, CA, USA.

Digital forms help us access services and opportunities, but they are not equally accessible to everyone, such as older adults or those with sensory impairments. Large language models (LLMs) and multimodal interfaces offer a unique opportunity to increase form accessibility. Informed by prior literature and needfinding, we built a holistic multimodal LLM agent for health data entry. We describe the process of designing and building our system, and the results of a study with older adults (N=10). All participants, regardless of age or disability status, were able to complete a standard 47-question form independently using our system—one blind participant said it was "a prayer answered." Our video analysis revealed how different modalities provided alternative interaction paths in complementary ways (e.g., the buttons helped resolve transcription errors and speech helped provide more options when the pre-canned answer choices were insufficient). We highlight key design guidelines, such as designing systems that dynamically adapt to individual needs.

Additional Key Words and Phrases: Accessibility; Health - Clinical; Input Techniques; Text/Speech/Language; User Experience Design; Older Adults; Mobile Devices: Phones/Tablets; Artifact or System; Field Study; Interaction Design; Prototyping/Implementation; Qualitative Methods

ACM Reference Format:
Andrea Cuadra, Justine Breuch, Samantha Estrada, David Ihim, Isabelle Hung, Derek Askaryar, Marwan Hassanien, Kristen L. Fessele, and James A. Landay. 2024. Digital Forms for All: A Holistic Multimodal Large Language Model Agent for Health Data Entry. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 72 (May 2024), 39 pages. https://doi.org/10.1145/3659624

1 INTRODUCTION

Forms serve as the entry point for accessing many services and opportunities, from opening bank accounts to applying for employment or social services. With the onset of paperless initiatives and online information systems, forms have become increasingly digitized, helping to mitigate human error, increase efficiency, and expand access to a broader community. However, despite their promise, not all groups can complete these forms independently. Within the domain of healthcare, specifically, this creates a critical challenge. The widespread adoption of electronic health record systems and patient portals now demands that patients complete multiple digital registration and health questionnaires before, during, and after many medical interventions.

This trend leaves many people behind and may prevent better health outcomes. For example, prior research found that despite the geriatric assessment's (GA) importance for older adults, its accessibility remains limited. In a study of 3,456 cancer patients taking a web-based GA, only 58% completed the survey independently [25]. Voice may be a more inclusive modality for assisting older adults [102]; however, older adults' needs have been largely overlooked during the design of voice-based systems [72]. Existing voice-based technologies, such as smart speakers, present several barriers for older adults [90]. Off-the-shelf smart speaker platforms require a high degree of hard-coded scripting, making it nearly impossible to develop interfaces that appropriately handle the unpredictable nature of human dialogue. As a result, voice-based surveys become unnatural and error-prone. Additionally, commercial smart speakers lack conversational grounding to appropriately perform conversational error repair. Instead, they defer to generic error messages, resulting in many breakdowns in communication [8, 18] that can make interactions with smart speakers frustrating.

Recent breakthroughs in large language models (LLMs), also known as foundation models [11], allow us to handle these issues in previously impossible ways, providing opportunities to build a more adaptive system. LLMs can help a system respond to natural language patterns and establish conversational grounding with far greater success than their predecessors. In this work, we prioritize the needs of those who could benefit from this

2474-9567/2024/5-ART72 $15.00
https://doi.org/10.1145/3659624

technology in potentially life-altering ways. Per Chan et al. [17]'s Holistic Multimodal Interaction and Design (HMID) framework, we focus on building a holistic system that: 1) **integrates** different modalities as one package; 2) **seamlessly transitions** from one modality to another at any time, even within a task, without much effort or loss of data; and 3) **is consistent** by using different modalities towards the same objective and to communicate the same message, while complementing each other by providing different advantages. Based on prior literature and needfinding, we built a holistic LLM-based system for health data entry. We placed the unique requirements of older adults with high degrees of frailty at the forefront of our design process to increase inclusion. We utilized a pre-trained LLM, OpenAI's GPT-4, which offers both text completion and chat capabilities. To mitigate the challenges that LLMs present for conversational controllability [93], we employed prompt-chaining strategies to guide a fluid conversation along survey questions and extract appropriate answer choices from user responses. We then ran a study with a wide range of participants, including one completely blind participant, one with poor hearing and fair vision, one 99 years old, several who speak English as a second language (ESL), and some older adults with low degrees of frailty. Our study had three main components: interacting with our system to complete a 47-question geriatric assessment, a post-interaction questionnaire, and a semi-structured interview. Through our design and development process and our study, we address the following research questions:

RQ1: What opportunities and challenges does a holistic multimodal LLM-based agent for health data entry uncover? How is such a system received by older adult participants?
RQ2: For older adults specifically: How do the different modalities work towards the same objective and to communicate the same message? How do they complement each other by providing different advantages?

We make three major contributions to the Ubiquitous Computing literature. First, we contribute a holistic, LLM-based multimodal system for health data entry that is usable by people of varying abilities and successfully classifies unstructured natural language data. Second, we provide qualitative and quantitative results from a field study investigating how a wide range of participants engage with this system. We find that it provides an alternative path for those marginalized by existing digital forms to participate in aspects of daily life from which they are otherwise excluded. Our participants, regardless of age or disability status, independently completed the geriatric assessment from start to finish. One participant who requires assistance to complete all digital forms called our system, "*a prayer answered.*" Our quantitative findings show low cognitive load levels, and above-average usability scores. Third, we offer design guidelines for creating holistic LLM-based multimodal systems to increase the inclusion of marginalized groups in the activities of everyday life, such as creating scientifically validated medical questionnaires for the voice modality, and designing systems that dynamically adapt to individual needs.

As a whole, our system and findings set a foundation for creating inclusive and scalable digital forms without overburdening other stakeholders.

1.1 Positionality

We are a multidisciplinary group of researchers with backgrounds in design, computer science, and geriatric oncology. The idea for the *Care Questionnaire* originated from medical practitioners who first-hand saw their patients struggle to independently complete the geriatric assessment. One author is a practicing clinician from the institution where the idea originated, and where user feedback from clinicians and patients with cancer has been positive.

2 RELATED WORK

In this section, we will contextualize our research and our motivations for conducting this study by situating it within the Ubiquitous Computing and Human-Computer Interaction (HCI) literature for multimodal systems, voice assistant use by older adults, and medical literature about the GA. We then describe work related to translating unstructured natural language into health records.

2.1 Multimodal systems in Ubiquitous Computing and HCI

Ubiquitous Computing and HCI have long concerned themselves with multimodal systems [7, 32, 51, 57, 69, 75, 91, 97, 98]. Multimodality conveys two salient features: the fusion of different types of data from/to different input/output devices, and the temporal constraints imposed on information processing from/to input/output devices [68]. Bolt [10]'s *Put That There* is a famous example of a multimodal system in which gestures (pointing) and speech were fused to place items on a projected display. We will now describe two lines of research concerning voice and touch/visual modality fusions, the modalities our system integrates—one in which the modalities complement each other in parallel by leveraging each other's strengths, and another in which they are sequentially used to perform different functions for accessibility.

The first line of research has examined processing these modalities simultaneously, or in parallel. For example, in 2002, Tsang et al. [91] developed *Boom Chameleon*, an input/output device consisting of a flat-panel display mounted on a tracked mechanical boom, that like ours, fused voice and touch input towards creating an enhanced 3D design review experience. Similarly, Hoque et al. [41] developed *MACH*, a multimodal virtual agent (embodied visually as an animated human character) that reads facial expressions, speech, and prosody and responds with verbal and nonverbal behaviors in real-time. Srinivasan et al. [81] developed a prototype system with in-situ, visual command suggestions that promoted discovery and encouraged the use of speech input during image editing. More recently, Srinivasan et al. [82] developed *inChorus*, a tablet-based system that supports pen, touch, and speech to investigate how multimodal interactions may function consistently across different visualizations to enhance visual data analysis. Srinivasan et al. [83] also created *DataBreeze*, a system employing the same three modalities as *inChorus* but on a digital whiteboard, which similarly leverages the strengths of one modality to complement the weaknesses of others. Similarly, Kim et al. [51] fused speech and touch interaction to foster the visual exploration of personal data, and found that their participants successfully adopted multimodal interactions for convenient and fluid data exploration.

The second line of research has investigated how these modalities can be used in place of one another for accessibility. For example, Zhang et al. [98] looked at how speech input could facilitate emoji (visual) use for people with visual impairments. Another example is Piper et al. [71]'s multimodal paper-digital interfaces for speech-language therapy, in which a client could use a pen to touch a specific item on an image and the pen would play back the audio corresponding to that item.

While the first line of research explores multimodal inputs, it does not explicitly focus on how to make these systems inclusive of people who may not be able to use all modalities. Similarly, the second line of research focuses explicitly on accessibility—not necessarily on how to serve a wide range of users with a wide range of abilities (inclusion). In our study, we extend this work by investigating a holistic multimodal interface that carefully finds a balance of using both speech and touch modalities to carry out the same objectives and communicate the same messages, while also using the modalities in a complementary manner.

2.2 Voice-assistant use by older adults

The predecessors to LLM-based multimodal systems like ours are voice assistants with touchscreens, such as the Amazon Echo Show or the Google Next Hub. Prior research suggests that patients with high degrees of frailty regard them positively and feel that they augment their independence [56, 61, 76]. In fact, some work suggests that people with disabilities and older adults actually prefer voice over "remote, smartphone, smartwatch, in-air gestures, or direct touch" for its simplicity, convenience, and "digital companionship" [47, 73]. Kocielnik et al. [54] found users especially prefer conversational user interfaces over non-conversational electronic form entry based on the added user engagement and understandability. However, Arnold et al. [5] warn that there are inconclusive results regarding the impact of voice assistants on older adults [5]. He cites three primary areas of concern around these systems: the pervasiveness of technical bugs, older adults' lack of familiarity with their functionality, and overall inaccessibility to marginalized communities. In our own explorations with the Alexa developer platform, we have found that even when extensively programmed, these systems can neither effectively direct the conversation, dynamically respond, nor provide advanced conversational self-repair that many users may require for successful and delightful interactions. While older adults were not necessarily prioritized in the design and development of smart speakers [73], recent work has highlighted important design guidelines to make them more inclusive: clear and concise instructions, minimizing complex menus and user interface interactions, and supporting error recovery [22, 38, 72]. In the context of voice-based systems, specifically, Pradhan et al. [72] warns about idiosyncratic patterns, accents, or pronunciation among older adults, which can lead to misinterpretation by existing voice assistants. For example, Kim et al. [49] found that natural language enabled both more natural and information-dense data inputs than graphical user interface inputs on a smartwatch, but they discuss that their system would likely present limitations for dysarthric, deaf, and accented people. Our work contributes to this literature by exploring and evaluating how LLMs may be used to address many of these usability issues that may create barriers to inclusion.

2.3 Geriatric assessment (GA)

The inaccessibility of the web-based GA served as the primary motivating use case for our system. For older adults, the GA is one of the more critical medical forms, a series of health questionnaires used to evaluate their "functional ability, physical health, cognition and mental health, and socio-environmental circumstances" [29]. Given that LLMs will inevitably play an important role in medical surveys [3], it is imperative that we gather empirical data specifically about the GA. Clinicians depend on the GA as a decision-making tool during critical periods of care: pre-operatively or during oncology treatment to predict mortality, surgical complications, the ability to successfully tolerate intensive treatment, and to live independently [31, 66, 70]. The GA extensively covers Activities of Daily Living (ADLs)—bathing, dressing, grooming, feeding, walking in and outside the home, and toileting [45]—as well as Instrumental ADLs (iADLs)—the ability to use the telephone, do laundry and other housework, shop, prepare meals, handle money and medications, and travel to the doctor's office [59]—among other measures.

Despite the GA's importance for older adults, its accessibility remains limited. In a study of 3,456 cancer patients taking a web-based GA, only 58% completed the survey independently [25]. As previously noted, the same study found a correlation between patients' frailty and their inability to complete the assessment without assistance (dependency). Though another web-based GA showed limitations in serving patients with higher frailty, 67% of participants claimed that they preferred the web-based GA to the pen-and-paper version [42]. Moreover, limited access to specialized services in many communities has heightened interest in patient care co-management models between geriatrics, primary care, and other medical specialties [63, 77, 89]. The success of these models, in part, depends upon the accurate completion of the GA; without an accurate frailty evaluation, providers risk overlooking critical information for care and exacerbating health disparities. Our system builds upon this preference for a web-based assessment, which may also facilitate care co-management, while expanding its accessibility to this important group of adults unable to complete the web-based GA independently.

2.4 Unstructured natural language for healthcare interfaces

In this section, we will first describe related work about conversational agents in healthcare, and then focus on technological advances regarding translating unstructured natural language into health records.

2.4.1 Conversational agents in healthcare. Conversational agents have gained popularity in the healthcare space, increasing the need for technology that can appropriately process unstructured natural language. Commercial artificial intelligence (AI) chatbots have flooded the market to intake patient data, aid in diagnosis, triage urgency, support patient care, and provide counseling [4, 12, 87]. These products are largely chat-based and serve as a foundation for conversational data input; however, publicly available research around these systems and their patient outcomes has been limited. Promisingly, one recent study found that an AI-based conversational agent for type-2 diabetics to manage their data resulted in improvements in time to optimal insulin dose, insulin adherence, glycemic control, and diabetes-related emotional distress [67]. Additionally, Dwaraghanath et al. [28] developed a virtual-assistant framework to anonymously gather mental health data that outperformed a human researcher control group on various measures (N=176).

Jo et al. [43] found similarly encouraging outcomes through *CareCall*, a patient assessment conducted over the phone to socially isolated patients in South Korea. *CareCall* extracts five health metrics, such as meals, sleep, general health, going out, and exercise, classifying them as positive, negative, or unknown based on a five-minute, open-ended dialogue. According to Jo et al. [43], *CareCall* offered a holistic understanding of each individual's condition while decreasing public health workload and helped mitigate loneliness and emotional burdens. While exciting, *CareCall* focused on an open-ended conversation without a controlled question schema, something crucial to the GA, and CareCall did not employ multiple modalities to increase inclusion.

Conversational agents also introduce new challenges. For example, as noted by Kim et al. [48], an increase in expressiveness afforded by voice inputs often requires added refinement to clarify user intentions. Moreover, converting naturalistic conversation into self-reported user data often presents prompting challenges of populating information slots [50, 93]. In our study, we address these challenges by building error-recovery mechanisms into our system, and using alternative prompting strategies beyond few-shot synthetic data [50].

2.4.2 Translating unstructured natural language into health records. Exciting technological advances in LLMs have dramatically improved our ability to translate unstructured natural language into health records. For example, Kjell et al. [52] demonstrated how pre-trained word embedding from LLM models (BERT) could be fine-tuned to quantify user responses to life satisfaction surveys and predict life satisfaction rating scales. When comparing validated rating scales and LLM classifications for patient free-text responses, the LLM accuracy converged with those of validated rating scales. Though it is worth noting patients preferred the natural expressiveness of open-ended free responses over rating scales [52, 53, 79], the paper's authors also suggest that quantitative scores about psychological wellbeing can be derived from natural language responses without sacrificing accuracy as measured by rating scales. While the Kjell et al. [53] fine-tuned these models to classify patient responses as clinical ratings, we used a more advanced, off-the-shelf LLM (OpenAI's GPT-4) that did not require fine-tuning. This said, fine-tuned LLMs may address concerns entailing potentially inaccurate health information provided by conversational agents [9]. Google's state-of-the-art MedPalm 2 question answering model, for example, achieves 86.5% accuracy on U.S. Medical Licensing Examination-style questions (MedQA dataset), and a comparison indicated preference for the system's answers to consumer health questions over physician responses [80]. MedPalm 2 has yet to be released beyond a select number of Google Cloud users; however, future iterations of our system could eventually benefit from this model, as the GA is typically administered by a clinician. Together, these technological advancements highlight the relevance and promising future of our system for processing various kinds of unstructured conversational data to be used in high-stakes settings, such as healthcare.

3 DESIGN PROCESS

In this section, we describe the process of adapting content from written words to spoken ones, designing a system backed by an LLM, and iterating on the interaction design. First, we identified a set of core requirements that we wanted our system to support with the goal of increasing the proportion of patients who could complete the GA independently. These included all questions from the original GA with voice and visual/touch interactivity. Then, we developed a version of this system using Alexa, but found the software development kit limiting and error-prone. We ultimately rebuilt the system outside the Alexa platform as a web application, which we refined as we learned more about the strengths and limitations of the tools we employed, and how people interacted with our system.

3.1 Conversion from written questions to spoken ones

Though the items included in the GA are derived from validated health questionnaires, some were easier to convert from a written to voice format than others. Questions with simple terms such as "how limited has your bathing been" and a limited number of single-word or short answer choices are the ones that can more easily be implemented in a multimodal system without demanding too much effort from a potential user. We omitted using some standard instruments that include complex, multi-part question construction. For example, after exploring some possibilities for how to adapt the Karnofsky Performance Scale [21] to voice, we determined that the task was too difficult to fit within the scope of this study. This scale asks patients to rate their functional status from 30% "severely disabled; hospital admission is indicated although death not imminent" to 100% "normal no complaints; no evidence of disease." Each step is ten percentiles, and includes about eight descriptive words. In written form, it is easier to see the scale and scan the form before reading the specific words next to a percentage value. However, it is unclear how best to administer this survey using a speech-based approach without overloading a participant with information. Moreover, this question was created for medical providers and the language was not validated for patient response, supporting our choice to exclude it. Prior to removal, we found that adapting slider-style questions with a large amount of information for each tick was a meaningful challenge.

Using an LLM allowed us to mitigate some of the redundancy in another style of question that was difficult to adapt to voice without creating too much repetition: select-all-that-apply questions. For example, a select-all-that-apply question about assistive devices, which required a "yes" or "no" question for each item in the Alexa version of our system, is now a single question in our LLM-based system with follow-ups. Despite having designed and built some interactions based on existing digital interaction patterns, such as dropdown menus, number pads, and sliders, we ultimately excluded them from our system. We found through our preliminary feedback that doing so would make the system more intuitive for more people, in particular those who may not be as familiar with many digital interaction patterns. This was especially important for the various modalities to complement each other while also being able to function independently from each other.

3.1.1 Textual and visual choices. The textual choices were made by a multidisciplinary team that included a geriatric oncologist, a nurse practitioner, and user experience designers. Conflicts were resolved through conversations, expert consultations, and user testing. We chose words as close as possible to the existing instruments, while having them still work in speech and visual formats. For example, we ensured that the text on the screen was not too verbose to avoid crowding the space and overwhelming users.

The visual choices were made in a similar, iterative manner, led by a user experience designer. The color blue was chosen to indicate trust, and white to represent a clean clinic. We chose a sans serif typeface for ease of readability on an iPad. The illustrations, such as the bathtub displayed in the teaser image, were purchased via subscription from Noun Project[1] under their royalty-free license and slightly modified. We specifically chose straightforward images that would not create much visual clutter. The captions followed common visual patterns also found in other popular captioned products, such as YouTube. We added visual separation between the question and answer option buttons and the status and control bar to indicate their different functionalities. We chose the color green for the bottom bar to indicate that the system is actively listening and working, and because in user testing participants thought that red meant "stop." A voice modulation graphic uses movement

[1] https://thenounproject.com/

to signal whether the device is listening and detects sound. The microphone icon with a "mic on" or "mic off" label uses a similar design pattern to Zoom, an interface many people have become familiar with through the COVID pandemic. We also chose it and included a label to avoid abstractions that could confuse users, such as just a circle (e.g., Siri) or a dot at the top of the screen (e.g., the orange dot on iPhones used to denote that the microphone is active). We made the typeface as large as possible to guarantee size consistency between questions while fitting the largest question and answer options on the device's screen. The question buttons and text are centered for visual balance, with sufficient surrounding space and button size to facilitate accurate tapping of the desired response. We added a full-screen that said "one moment please" with a typing sound to simultaneously signal that the machine is busy and prevent prevent users from tapping buttons in that moment.

3.2 LLM integration

The *Care Questionnaire* builds upon a voice assistant-based GA that did not have touch input [24]. We replicated this Alexa-based system and tried to leverage the Echo Show's built-in features by adding buttons to its touchscreen. However, the process was impractically slow due to a lack of up-to-date documentation and developer support from Amazon. Moreover, Alexa's requirement for pre-programmed rules prevented us from being able to support the type of unpredictable, open-ended answers that the voice modality elicited. These barriers motivated us to create a new web-based system that we could have more control over, including the ability to integrate an LLM.

We first built a rule-based agent, with hard-coded statements for each question to avoid LLM hallucinations. This meant injecting statements with filler words like "alright, now" and "for the next question," to sound conversational. We only generated text through an LLM when the user responded with an answer outside the predefined choices. When the model could not produce an answer classification, we reiterated the original question. This version of the system maintained tighter control over how the GA was administered; however, the benefits did not outweigh the need for more conversational fluency. This version also required that we explicitly map dependencies between questions and manage this logic in the system's front-end, which increased the possibility for programmatic errors. For example, the question "when did you last fall?" depends on the user responding affirmatively to "have you fallen in the last 6 months?" Thus, for questions with dependencies, we treated the dependencies as a graph problem, marking questions as children of their predicates.

In contrast to our first version, our latest version of the system delegates both the answer classification and response generation to GPT-4, having the system fully drive the survey. To successfully complete both tasks, we split up the work into two chained tasks, described in greater detail in Section 4. **Task 1** classifies the user's responses into valid answer choices—the list of choices is provided by our survey schema. **Task 2** generates a reply back to the user based on this classification. Task 2's output might include a response to a user's question about what they were asked (here, the system *provides* clarification), a clarifying question to get a more accurate answer from the user (here, the system *seeks* clarification), or the next unanswered question from the GA.

At every conversational turn, for the system to return an optimal response back to the user, it has to meet two requirements: A) Task 1 cannot produce a false-positive classification for the last question asked (i.e. classifying it as "Not limited" when the user has not given a sufficient answer), and B) Task 2 cannot advance to the next question without a sufficient answer to the last question. If not addressed, these requirements result in system errors. For meeting the first requirement, we gave succinct warnings to the model to avoid incorrectly classifying ambiguous answers; see the meta-prompt in Section 4.4.3, and Table 1 for examples. For meeting the second requirement, we provided copious information in the prompt. Specifically, we had to provide the status of the last question asked: either "(question (id: [id]) has not gotten a sufficient answer from the patient" or "question (id: [id]) has been successfully recorded as: [answer option])". We also had to add explicit instructions to remain on a question until a sufficient answer was received. Without all of these instructions, GPT-4 frequently inferred

State of the last question asked	question (id: 20) has not gotten a sufficient answer from the patient
Next questions	(id: 20) How physically active are you? Answer choices: very active, moderately active, not active (id: 21) How much water do you drink a day Answer choices: 1-5 cups, 5-10 cups, 10+ cups
Conversation with an exemplar response	**System**: So how physically active are you? The answer choices are very active, moderately active, not active. (id: 20) **User**: I play tennis. **System**: So would you say you're very active or moderately active? (id: 20)

Table 1. Example given to the LLM for conversations requiring clarification.

incorrect answers, and advanced to the next question prematurely. We now provide more details about how we prompted our system to meet these requirements.

A) Asking for clarification. Our preliminary testing revealed that the speech modality elicited responses that often deviated from the provided answer choices in unpredictable ways. Thus, we designed the system so that Task 2 (Section 4.4.3) ensures the model continues to ask follow-up questions to arrive at a provided answer choice when Task 1 (Section 4.4.2) cannot produce an answer classification. Here, again, we provided examples in the prompt, because we found that doing so was more effective than in-context instructions alone (see Table 1 for an example).

B) Determining which question to ask. Determining the next question required more complexity than simply advancing to the next question in the survey. Some user responses might obviate subsequent questions in the survey. If users, for example, indicate they do not smoke, our system needed to skip subsequent follow-ups around the quantity or history of smoking. Ceding this kind of question navigation to the model was difficult, because instruction alone (telling the model to disregard irrelevant follow-up questions) did not prevent redundant questions. Therefore, we added several examples, such as the example in Table 2. These examples needed to differ from the ones in the GA; otherwise, GPT-4 would repeat those examples instead of generating a contextually appropriate response. We also found that the order in the output structure impacted the accuracy of LLM response generation [92]. Asking the model to prefix its response to the user with the current question identifier (id: [id]) often resulted in inaccurate question identifiers. However, having the model append the current question identifier to the end of the response improved the overall output.

In the latest version of our system, meeting the first requirement (A) continues to be a larger challenge than meeting the second one (B). For example, using the schema in Figure 2, the LLM produces a classification of "WITH_SOME_HELP", despite the ambiguity that the grandson may merely accompany the user for companionship. The system will generate a variant of "That's lovely, I'll note that as 'with some help'." Whereas a human interviewer would likely follow-up for more refinement, GPT-4's completion model presumptuously infers an

State of the last question asked	question (id: 13) has been successfully recorded as: "No"
Next questions	(id: 13) Have you had any heart attacks? Answer choices: Yes, No (id: 14) How many heart attacks? Answer choices: 1, 2, 3 or more (id: 15) Have you recovered from the heart attack? Answer choices: Yes, No (id: 16) Are you able to drive independently? Answer choices: Yes, No
Conversation with an exemplar response	**System**: Have you ever had a heart attack? (id: 13) **User**: Never. **System**: Alright, now I'd like to ask if you're able to drive independently (id: 16).

Table 2. Example given to the LLM for conversations requiring skipping follow-ups.

answer choice. This said, as conveyed in Section 2.4.2, models for classifying unstructured natural language are rapidly advancing, which may soon improve our system's ability to address the first requirement more gracefully.

3.3 Interaction design

We worked on the core functionality and interaction design of our system in parallel. For the interaction design, we rapidly generated many ideas in Figma, bringing them to life via short demo videos to obtain feedback. In addition to this feedback, we conducted a heuristic evaluations using heuristics developed for speech-based smart devices [94] to evaluate our system. Our interaction videos underwent a total of seven iterations, based on feedback from the authors and from other experts in the field of ambient intelligence [20]. Based on our first feedback session, we decided to 1) design for tablets instead of phones to maximize visual and touch accessibility, 2) increase the contrast between selected and unselected button states, and 3) create more designs for communicating the state of the system. There were four main interaction states that were challenging to figure out how to best communicate to older adults, when the system is: listening, processing information, speaking, or idle. Ideas to communicate these varied, such as explicitly spelling the state out on the screen, implicitly displaying some elements (e.g., by using captions), or combinations of implicit and explicit design signifiers. In the second feedback session, with eight experts, we obtained feedback on revised designs and new concepts from the ideas in the prior session. For example, we considered a chatbot design paradigm with bubbles for each speaker (as shown in Figure 3 middle right bottom). We also discussed the challenges of creating voice-based parallel interactions for the navigation buttons and review buttons on the header of the interface. Ultimately, we decided to communicate the state of the system implicitly through captions, a "one moment please..." screen with a spinning GIF accompanied by typing sounds, and a microphone "on/off" button. We also de-prioritized the navigation header for the first study of the system. Iteratively, we arrived at the final version shown in Figures 1 and 3 (rightmost).

```
{
  "json": {
    // How independently can you do your own shopping? (id: 23)
    "shopping": enum,
    // How independently can you cook your own meals? (id: 24)
    "meals": enum,
    // How independently can you do your own housework? (id: 25)
    "housework": enum,
  },
  "schema": {
    "shopping": {
      "Enum values": [["WITHOUT_HELP", "WITH_SOME_HELP", "UNABLE_TO_DO_SO"]
      // null if not known
    },
    "meals": {
      "Enum values": ["WITHOUT_HELP", "WITH_SOME_HELP", "UNABLE_TO_DO_SO"]
      // null if not known
    },
    "housework": {
      "Enum values": ["WITHOUT_HELP", "WITH_SOME_HELP", "UNABLE_TO_DO_SO"]
      // null if not known
    }
  },
  "conversation": "
       Human: Not limited.
       System: How independently can you use the telephone? The answer
       choices are without help, with some help, unable to do so.
       Human: I call my sister every day by myself.
       System: How about laundry?
       Human: It's easy for me to do.
       System: That's great How independently can you go shopping?
       Human: Sometimes my grandson comes with me"
 }
}
```

Fig. 2. Schema parsing with ambiguous response.

Additionally, we found strategies to reduce cognitive load when moving to a voice modality. For example, the written version of the GA often provided patients with an overwhelming number of options (> 5), creating a need for the user to remember all the items by the time they had to respond. To avoid doing this, we split these into hierarchical questions. For example, instead of asking patients to choose a range of pounds lost or gained, we asked (1) "Have you gained or lost weight in the past 6 months?" and (2) "How much weight have you lost (or gained)?" If users experienced no weight change, they skipped the follow-up entirely. We considered other

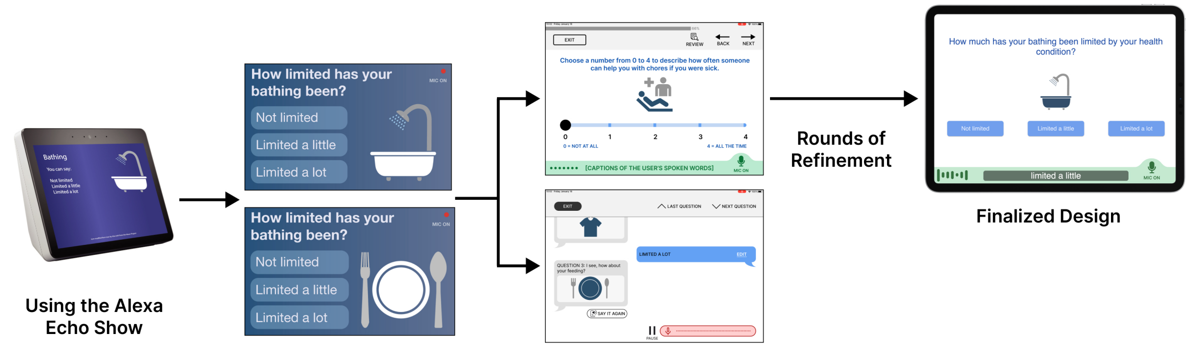

Using the Alexa Echo Show
How limited has your bathing been?
Not limited
Limited a little
Limited a lot
How limited has your bathing been?
Not limited
Limited a little
Limited a lot
Choose a number from 0 to 4 to describe how often someone can help you with chores if you were sick.
0 = NOT AT ALL
4 = ALL THE TIME
[CAPTIONS OF THE USER'S SPOKEN WORDS]
LIMITED A LOT
SAY IT AGAIN
Rounds of Refinement
How much has your bathing been limited by your health condition?
Not limited
Limited a little
Limited a lot
limited a little
MIC ON
Finalized Design

Fig. 3. Alexa Echo Show (leftmost) was redesigned in a web-based format (middle left), iterated upon using mocks and videos (middle right), and refined to our current version (rightmost).

options such as using a number pad, which would make the speech interaction more efficient. However, doing so would require introducing a new interaction mechanism, which could increase the difficulty of the touch input.

4 SYSTEM DESCRIPTION

We now describe the key components of our system, including its multimodal interface, software design, questionnaire structure, and prompts. Figure 4 provides an overview of our system design.

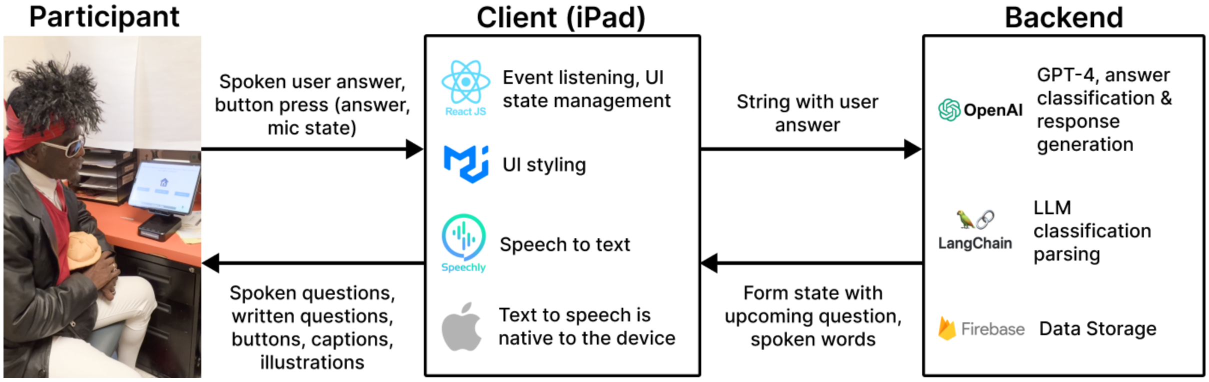

Participant
Client (iPad)
Backend
Spoken user answer, button press (answer, mic state)
Spoken questions, written questions, buttons, captions, illustrations
React JS
Event listening, UI state management
UI styling
Speechly
Speech to text
Text to speech is native to the device
String with user answer
Form state with upcoming question, spoken words
OpenAI
GPT-4, answer classification & response generation
LangChain
LLM classification parsing
Firebase
Data Storage

Fig. 4. Overview of our system design. The participant interacts with our system on an iPad. The iPad handles the UI state management and styling, and the speech-to-text and text-to-speech. The answer classification and response generation is performed on a server in the backend of our system.

4.1 Multimodal interface

Our system's multimodal components include the hardware used, the design of the speech-based interactions, and the design of the visual and touch-based interactions.

4.1.1 Hardware. Our system is designed for an iPad. For our study, the iPad sits on a speaker stand that props the screen up at an angle facing the participant. We use the Safari web browser, and emulate a native app by

Synonymous Answer	Implied Answer	Answer with Context	Several Back and Forths
System: How much has your bathing been limited by your health condition? **User:** It's been a bit challenging.	**System:** How is your vision? **User:** I'm blind.	**System:** How much has your walking outside the home been limited by your health condition? **User:** I used to go hiking, walk on the beach, or go on long strolls. Now I can't do any of that anymore.	**System:** How much has your grooming been limited by your health condition? **User:** What do you mean by grooming? **System:** Ability to brush your teeth, shave, or maintain personal hygiene. **User:** Oh, it's totally been fine.
Classification			
Not limited, **Limited a little**, Limited a lot	**Poor**, Fair, Good, Excellent	Not limited, Limited a little, **Limited a lot**	**Not limited**, Limited a little, Limited a lot

Table 3. Examples of open-ended answers.

saving our web app to the iPad home screen. The iOS Guided Access feature supports "kiosk mode" to prevent the user from accidentally exiting the questionnaire and needing to restart.

4.1.2 Speech-based interactions. The speech-based interactions are designed such that a user can complete the questionnaire without looking at the screen or touching the iPad. As presented in Table 3, our system supports answers that are synonyms, implied, give a large amount of context, or include multiple conversational turns. It also states the answer choices that the user may select from after a new question is asked.

4.1.3 Visual and touch-based interactions. The visual display has three major components. First, it displays the written question. Second, it provides buttons with the potential answers to the question. Last, it has a footer with several elements: a voice modulation animation, captions, and a microphone icon. The microphone can be enabled or disabled through a tap to allow the user to make the system stop listening.

4.2 Software Design

The system is comprised of two primary components: a frontend built in React[2], and a backend built with Firebase[3]. This system architecture allowed us to avoid the infrastructure requirements of a full backend server. End to end, the interactions begin on the client side (the iPad), which welcomes the user, requests microphone permissions, and awaits confirmation to continue. To support the microphone, text-to-speech, and transcription

[2] React is an open-source, frontend JavaScript library for building user interfaces. We use React in Typescript, a strongly typed programming language built on top of Javascript.

[3] Firebase is an app development platform that provides developers with a suite of tools and services. Within Firebase, we use the services Cloud Functions to house backend logic code that communicates with the LLM, and Firestore as a NoSQL database to store responses.

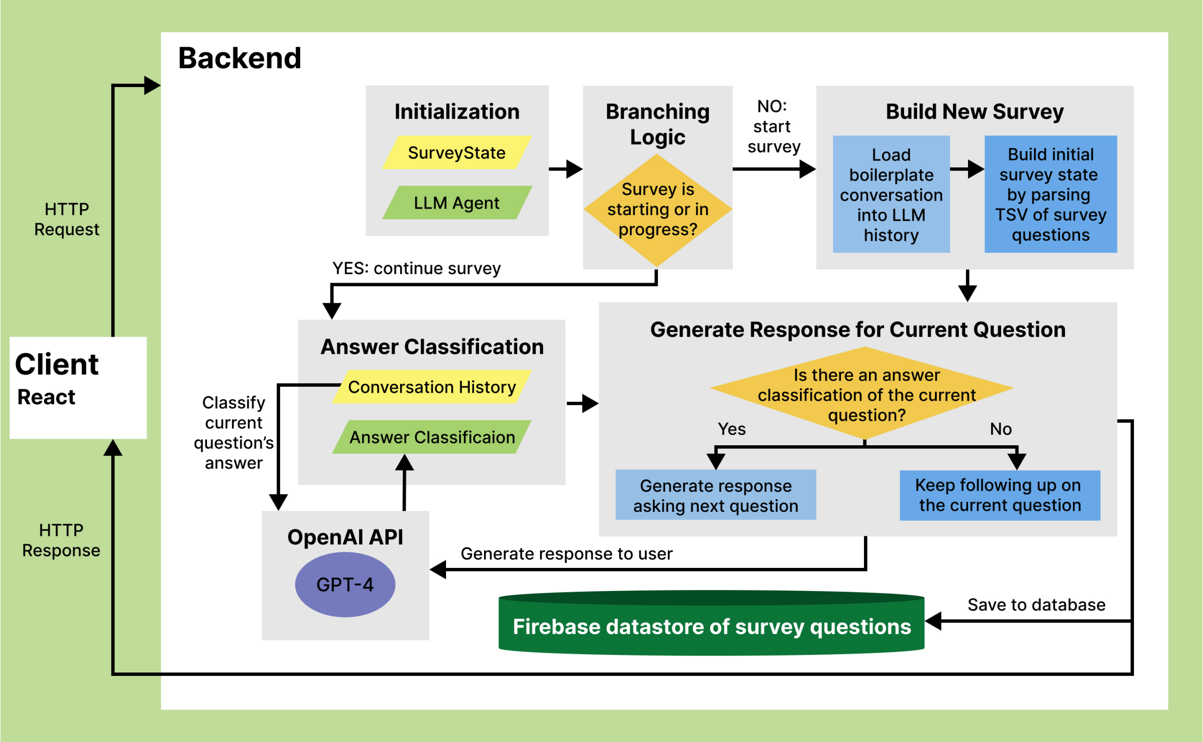

Backend
Initialization
SurveyState
LLM Agent
Branching Logic
Survey is starting or in progress?
NO: start survey
Build New Survey
Load boilerplate conversation into LLM history
Build initial survey state by parsing TSV of survey questions
YES: continue survey
Answer Classification
Conversation History
Answer Classificaion
Generate Response for Current Question
Is there an answer classification of the current question?
Yes
No
Generate response asking next question
Keep following up on the current question
Classify current question's answer
OpenAI API
GPT-4
Generate response to user
Save to database
Firebase datastore of survey questions
HTTP Request
Client
React
HTTP Response

Fig. 5. System architecture. The React frontend makes requests to the backend Firebase server. A SurveyState is built based on questions and answer choices in a provided .tsv file. Once the survey is in progress, the conversation between the system and the user is parsed for answer choices. Once they are extracted, the system generates a response to the user through the OpenAI Chat Completions API endpoint.

functionalities, we deployed Speechly, an on-device Typescript library[4]. However, for U.S. Health Insurance Portability and Accountability Act (HIPAA) compliance, future iterations will require a qualifying platform, such as Microsoft Azure. On the initial request to the backend, the API returns a blank SurveyState, complete with all questions to ask in the survey, based on a .tsv file of questions and respective answer choices. We maintain a conversation history between the agent and user to provide to the LLM when classifying answers and generating responses. On the front end, React UI event hooks constantly poll for verbal input from the user. When no new speech is processed for three seconds, a Speechly-generated transcript is sent to the backend for processing along with the SurveyState. During each turn, the backend conducts two tasks: 1) parsing the user transcript for answers to questions and update the survey state accordingly, and 2) generating a response back to the user based on the current question being asked, see Figure 5 for more details.

4.3 Questionnaire Structure and Scoring System

The system sequentially steps through the GA's items, see Table 7 in Appendix B for the list of all questions and their respective answer options. Once users complete the GA, we compute risk scores based on their responses.

[4] Speechly is a voice technology that offers Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) tools and APIs. We only used its Speech Recognition API to perform speech-to-text conversions.

Within each section, there is a threshold for "abnormal scores" calculated using guidelines from Shahrokni et al. [78]. A final report is generated upon completion and uploaded to a database from where it can be sent via secure email to a patient's care team or printed.

4.4 Prompts

In this section, we describe the high-level design of our prompts, and two tasks, chained to one another, that are passed to GPT-4.

4.4.1 Prompt design. There were three primary pieces of information we delegated to GPT-4: 1) the classification of user responses into appropriate answer choices (i.e., "Not limited", "Limited a little", "Limited a lot"), 2) a reply to the users' prior response, usually containing the registered response classification as implicit conversational grounding, and 3) the ID of the current question being asked. We then chain two sub-tasks—Task 1 passes LLM outputs to Task 2 (following a least-to-most prompting approach [101]). We now describe each task in more detail.

4.4.2 Task 1: Parsing the user response into a structured output. With each user reply, we prompted the model to parse the most recent six conversational turns—ending with the last system utterance—to create a memory buffer. The script below shows the six turns forming the memory buffer, in addition to the last user utterance (a clarification question), which is appended to the prompt.

User: It's been harder to get into the tub lately.

System: I see, thank you for sharing that. Would you say it's been limited a lot or a little?

User: Only a bit

System: Understood. Now, could you tell me how much your health condition has limited your ability to dress yourself? (id: 2)

User: Limited a little

System: Thanks. Moving on, how about grooming? (id: 3)

User: What do you mean by grooming?

Our system architecture diagram in Figure 5 illustrates this process. Given that the survey has a specific schema of questions and answer choices, we provide it to the LLM along with the expected JSON output format. Then, we can parse it to update our survey state. Figure 6 shows the resulting Task 1 prompt for the exchanges above.

4.4.3 Task 2: Providing the next statement to the user. We then relay the classified answer from Task 1 to the model as a follow-up prompt, Task 2. This generates the response back to the user. Informed by the work of Zhang et al. [99], we give the system a delineation of the reasoning process so that it generates a contextually-aware response. This delineation starts with a metaprompt, which is described by Reynolds and McDonell [74] as a "more general intention that will unfold into a specific prompt when combined with additional information." This is our metaprompt:

> *At every step, I will provide two things: 1) The next few questions to ask, along with their ID, and possible answer choices. You should end your entire response with the ID of the question you are asking, and 2) The recorded answer for the last question. Do not move onto the next question if the current question is not sufficiently answered.*

```
{
  "json": {
    // How limited has your grooming been? (id: 3)
    "grooming": enum,
    // How limited has your feeding been? (id: 4)
    "feeding": enum,
    // How limited has walking inside the home been? (id: 5)
    "walking_inside_home": enum,
  },
  "schema": {
    "grooming": {
      "Enum values": ["NOT_LIMITED", "LIMITED_A_LITTLE", "LIMITED_A_LOT"]
      // null if not known
    },
    "feeding": {
      "Enum values": ["NOT_LIMITED", "LIMITED_A_LITTLE", "LIMITED_A_LOT"]
      // null if not known
    },
    "walking_inside_home": {
      "Enum values": ["NOT_LIMITED", "LIMITED_A_LITTLE", "LIMITED_A_LOT"]
      // null if not known
    }
  },
  "conversation": "
       Human:  It’s been harder to get into the tub lately.
       System: I see, thank you for sharing that. Would you say it’s
       been limited a lot or a little?
       Human:  Only a bit
       System: Understood. Now, could you tell me how much your health
       condition has limited
       your ability to dress yourself? (id: 2)
       Human: Limited a little
       System: Thanks. Moving on, how about grooming?
       Human: What do you mean by grooming?"
 }
}
```

Fig. 6. Overview of Task 1.

We then give a few short examples as part of the prompt [15, 55], as shown in Tables 1 and 2 and provide rules, such as stating the answer choice the system selected aloud when the user's answer was ambiguous, for conversational grounding. Our full prompts are available in the Supplementary Material[5].

[5] https://github.com/StanfordHCI/CareQuestionnaire

5 FIELD STUDY METHOD

We conducted an IRB-approved field study with older adults (N=10) to evaluate our system's performance, and video-recorded their interactions with our system. We now provide a description of our participants, procedures, and measures.

5.1 Participants

We recruited 10 participants (six women, three men, and one preferred to self-describe) who were on average 80 years old through email, word of mouth, and physical flyers. When asked about their current way of filling out medical forms, three participants stated needing assistance from another person, four used paper-based forms, and three filled them out digitally without help. When asked what type of work they had done most of their lives, participants reported a wide range of professions, including: artist, fashion, dentistry, casino, technology industry, teacher, music composer, homemaker, professor, professional interviewer, veteran, and massage therapist. Nine participants identified as white, and one as Black. None identified as Latinx. Five participants were never married, three were widowed, and two were married. Seven participants lived alone, and three did not. All participants reported owning and using at least one computing device. Eight participants reported using a computing device to go online every day and two did not provide frequency of use information. Nine participants had WiFi at home. Eight participants felt very confident reading and writing, one felt neutral, and one not confident at all. Table 4 shows information by participant that may aid in contextualizing some of the findings, including: age and gender demographics, whether they spoke ESL, their schooling and income information, their confidence using computing devices, and their confidence using speech-based computing devices specifically.

We conservatively determined our sample size based on prior research suggesting that for nonprobabilistic sampling, basic elements for meta-themes are present in as early as six interviews [36], and usability guidelines that suggest that 85% of usability problems are found in the first study with five participants[6].

5.2 Procedure

We conducted the study at a location convenient to each participant, either at their home, at their local senior center, or at our institution. The location had to be a relatively quiet room, in which only the participant and one to two researchers were present. We set up an iPad (iPad Pro, 12.9-inch, fifth generation) on a stand to mimic a smart speaker, and placed it on a table next to a chair where participants sat. We used the iPad's native, female-sounding voice with an American accent. The video was recorded using a smartphone on a tripod at 30 frames/second with a resolution of 1920 x 1080 DPI. Study sessions lasted approximately 90 minutes, and participants were compensated with $50 gift cards.

We first obtained consent, including permission to audio and video record, and then gave participants the following scenario: "*Imagine your doctor sent you this questionnaire. Please answer the questions that the machine asks you. We are not evaluating the answers to the questions, please feel free to lie. We are only evaluating the machine's ability to ask questions in a natural way that feels intuitive to older adults.*" They received minimal to no training. For the first half of participants, we gave them no indications about the different available modalities. However, in two occasions, we realized some did not know about the existence of an alternate modality, and we interrupted the session to let them know about the other available option. In subsequent sessions, we mentioned the two modalities before they started their interaction, and instructed them to interact with the system in whichever manner felt intuitive to them. Nevertheless, P9 asked for extra clarification during the interaction about what modalities she could use. Additionally, P6 used the voice modality for the majority of the interaction, though the researcher reminded her that the touch modality was available when her body language conveyed frustration with the slow pace. Participants were then asked to interact with our system to complete the geriatric assessment.

[6] https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

P#	Age	Gdr	ESL	Highest degree or level of school completed	Gross income ($)	Confidence using computing device	Confidence using speech-based computing device
P1	56	M	No	Some college	10,000–14,999	Neutral	Somewhat confident
P2	87	F	Yes	Bachelor's degree	25,000–34,999	Somewhat confident	Not at all confident
P3	93	M	No	Doctorate	75,000 or greater	Somewhat confident	Somewhat confident
P4	76	F	No	High school diploma or GED	25,000–34,999	Unreported	Somewhat confident
P5	83	M	No	Some college	25,000–34,999	Very confident	Very confident
P6	64	F	No	Bachelor's degree	35,000–49,999	Very confident	Not at all confident
P7	99	M	No	High school diploma or GED	Unreported	Not at all confident	Not at all confident
P8	83	N/A	No	Doctorate	Don't know / decline to answer	Very confident	Only a little confident
P9	N/A	F	No	Some college	25,000–34,999	Neutral	Somewhat confident
P10	81	F	Yes	Associate's Degree	Unreported	Neutral	Not at all confident

Table 4. Demographic details by participant.

This portion of the study took 16 minutes and 19 seconds on average—with the shortest at 10:03, and longest at 23:03. Once the interactions concluded, participants completed, in chronological order: a questionnaire about their interactions, a semi-structured interview, and a demographics questionnaire. All participants completed these components of the study immediately after their interaction task, except for P9, who asked to stop after the interaction with the *Care Questionnaire* ended[7]. P9 completed the post-interaction parts of the study on a different day.

5.3 Measures

In this section, we describe our measures and analysis approaches, including the video analysis of interactions with our system, the interview, and our quantitative measures from the post-interaction questionnaire and logged responses.

5.3.1 Thematic analysis of interaction videos and follow-up interviews. We used an automated transcription service to transcribe the interactions with our system and follow-up interviews that took place during the study.

[7] She had gone through hardships that generated difficult emotions, which were triggered by some of the questions.

We then analyzed the transcripts while using the videos for reference using thematic analysis [13]. We initially open-coded [46] a portion of the transcripts. For the interviews, this process resulted in over 65 different codes (e.g., too slow, intrusive, responsive, and self reflection), which we then clustered, roughly by research question, and refined to create a codebook with about a fifth of the initial number of codes. For the interaction transcripts, this resulted in a set of codes that we deemed most important (e.g., researcher intervention, conversational repair, touch input, voice input, and input error). These refined codes were used used to continue labeling the remaining transcripts, and were iterated upon as appropriate. As discussed by McDonald et al. [64], the primary goal of this process was not to find agreement, but to identify recurrent topics or meanings that represent a phenomena. At least two researchers reviewed each video and transcript, and one researcher that was present in all study sessions ensured that there was consistency between all of the coded transcripts, often revisiting already-coded ones. We now discuss more specific details about the thematic analysis of the interactions and interviews, separately.

Video analysis of interactions with our system. We used video analysis [22, 44, 62, 84, 85, 88, 95] to capture patterns from this primary empirical data that would not be visible without video (e.g., by changing playback speed, or measuring the duration of certain behaviors), resulting in more consistency and reliability in our observations. We annotated the transcriptions with our observations from the videos, taking time to pause, slow down, or speed up certain aspects of the interactions to observe behaviors in different ways.

We also used these videos to create the interaction sequence shown in Figures 7 and 8. For this step, three authors met to discuss the important interactions to record (i.e., an early version of the key in Figure 7). Two authors then made interaction sequences independently, and then received feedback on them from the first author. We met to create a final plan for generating the diagrams based on the initial feedback, and this plan was used to complete the remaining sequences and further refined as the interaction sequences were completed. This task was split between two coders, such that only one coder fully reviewed each video for these sequences. The coders maintained frequent and constant communication to address any ambiguity encountered.

Interview. Following completion of the post-test questionnaire, we conducted a 30-minute, semi-structured interview about participants' experience with our system. For example, we asked, "How did that go?" and "What do you currently do to fill out health forms? How does this compare?" We then grouped the codes into several major themes, including: error recovery, open-endedness of speech, accessibility, cumbersome interactions, inadequate pace, transcription errors, problem with the question itself, and other applications based on lived experiences.

5.3.2 *NASA Task Load Index (TLX).*

The NASA TLX is a multidimensional scale used to obtain estimated workload from users related to a specific task. Originally developed for use in engineering and aviation [39], it has been validated in use with medical personnel and older adults, among other populations [27, 40, 100]. The user ranks six domains (mental, physical and temporal demand, as well as perceptions of their performance, effort, and frustration with the task) on a visual analog scale from zero (Low) to 20 (High), each of which is then multiplied by five. We report the subscale means and unweighted raw TLX score for each participant [16].

5.3.3 *System Usability Scale (SUS).*

The SUS is a ten-item scale created to capture the user's subjective global view of the usability of an electronic system. After exposure to a new system, the user ranks each item on a Likert scale from "Strongly disagree" to "Strongly agree," and the scores are transformed according to their negative or positive polarity, then summed and multiplied by 2.5 to result in an overall score from zero to 100 [14]. The SUS has been used extensively with varied populations, and demonstrates excellent reliability with an alpha coefficient of 0.91 [37, 60].

5.3.4 *Self-Consciousness Scale.*

Use of a voice interface, especially when others are present, may induce feelings of exposure and self-awareness. We used the public self-consciousness nine-item subscale of this instrument, which examines participants' feelings about themselves and how they are perceived by others, e.g. "I am concerned about what other people think of me." Each of the nine items is rated on a Likert scale from "Strongly disagree" to "Strongly agree" [30].

5.3.5 Transportation. Psychological transportation is defined as absorption into a story or experience that involves imagery, attentional focus, and emotional involvement [34]. We used an eight-item version of the Transportation Scale, the same selection of items employed by Wenzel et al. [96] in their voice assistant study, to measure perceptions of task immersion and relevance, with items such as "I was distracted by my own thoughts and feelings," and "I felt the tasks were relevant to ones I'd do in my everyday life" [33].

5.3.6 Technology evaluation. Based on the Technology Acceptance Model [26] and Reasoned Action Approach [2], an 11-item series of semantically paired questions asked users to rate dimensions about specific qualities of the technology on a scale from one to seven, e.g., "Useful-Useless" and "Pleasant-Unpleasant." We used the same items that Wenzel et al. [96] selected for their study to enable the use of their results as a benchmark. All the survey items and aggregated participant responses are provided in the Supplementary Material.

5.3.7 System performance. A nurse practitioner examined the videos of participant interactions and reported the responses she would have registered in a clinical setting. These responses were compared to those logged by the system in two ways. First, the nurse's answers were treated as the ground truth and we measured the percentage of agreement between the nurse and the system. Second, we treated the nurse practitioner and the system as separate raters and calculated inter-rater reliability using Cohen's Kappa for each question, as percentages do not correct for how often raters may agree by chance [65].

6 FIELD STUDY FINDINGS

Participants highly differed in their preferences, both in which aspects of our system they benefited from or were bothered by. In this section, we describe our findings surrounding diverse conversational styles, the unstructured, or unscripted, content elicited by the speech modality, how the different modalities complemented each other, the benefits of our system for inclusion, and challenges that arose. We also share ideas our participants offered for other uses for a system like ours as well as quantitative findings.

6.1 Participants interaction styles widely varied

We observed a wide variety of interaction styles among our participants, ranging from frequently unscripted (e.g., P2, and P7)—using language that deviated from the provided answer options—to mostly scripted (e.g., P1, and P8)—strictly adhering to the answer options that our system provided. As can be observed by the mixture of colors in the interaction sequence illustrated in Figures 7 and 8, most participants alternated between scripted and unscripted speech responses and the speech and touch input modes, with eight of the ten participants using each of the two main modalities at some point during their interaction. Additionally, all participants used speech and gave or attempted to give an unscripted response at least once, indicated by the presence of blue squares. Unscripted responses were mediated by the LLM capabilities of our system in ways that previous systems, such as the Alexa version of this questionnaire, would not be able to handle (see Section 6.2 for more detail). Some participants' interactions were not just unscripted, but also involved both speech and touch input simultaneously, creating timing challenges, as illustrated by attached squares representing different modalities with one only partially filled. For example, when asked about education level, P3 tapped the "advanced degree" button as he verbally expressed, *"I'm smart as hell"*. P7 spoke similarly, offering additional context as she pressed buttons on the touchscreen. Additionally, some reactions to our system or responses involved different forms of non-verbal expressions, such as body language (e.g., shaking head, laughing and leaning towards and away from the device)

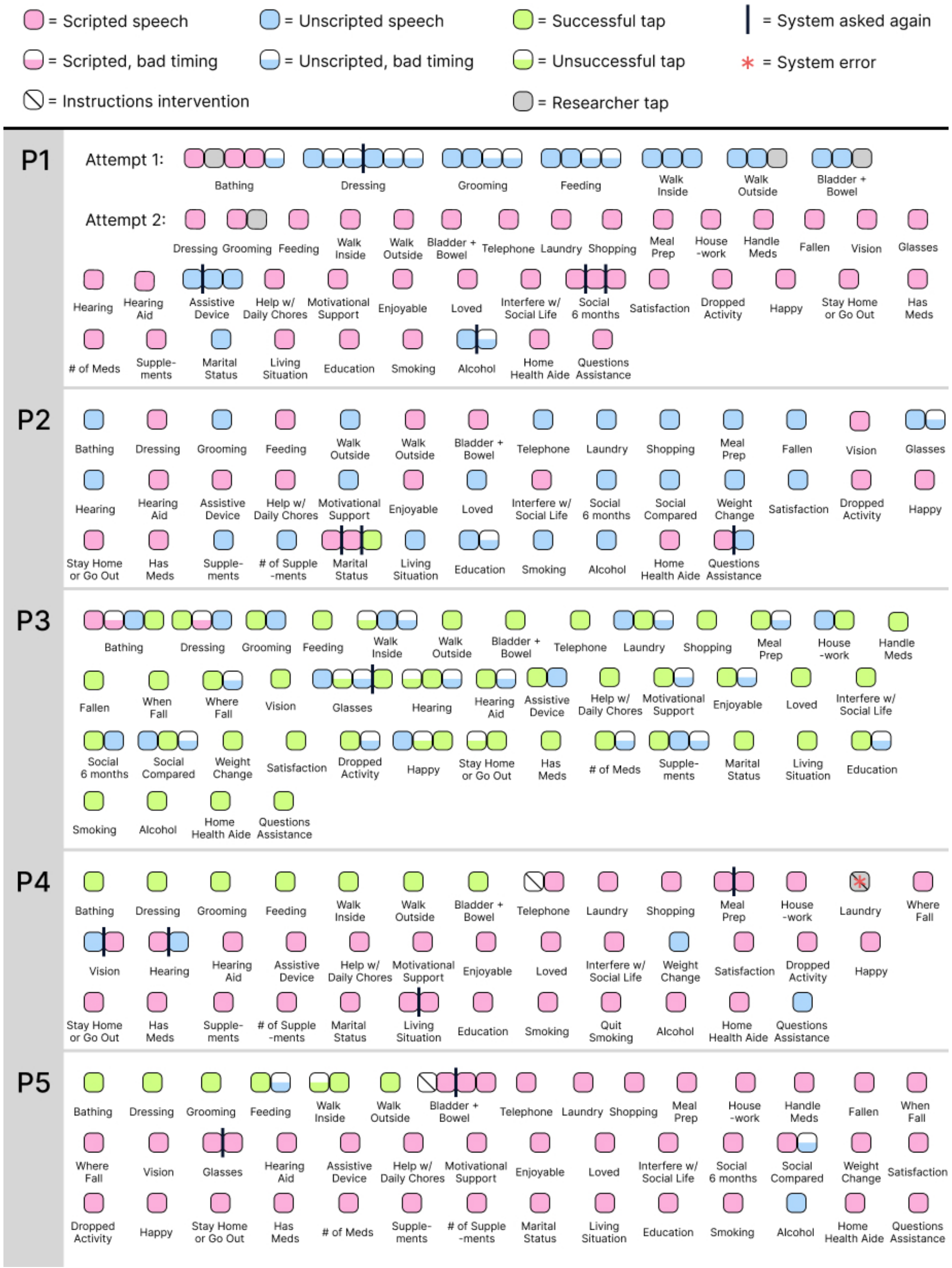


Fig. 7. Interaction sequence part I.

or intonation variations. As a whole, we observed a wide range of interaction styles and modality preferences in participants' interactions with our system.

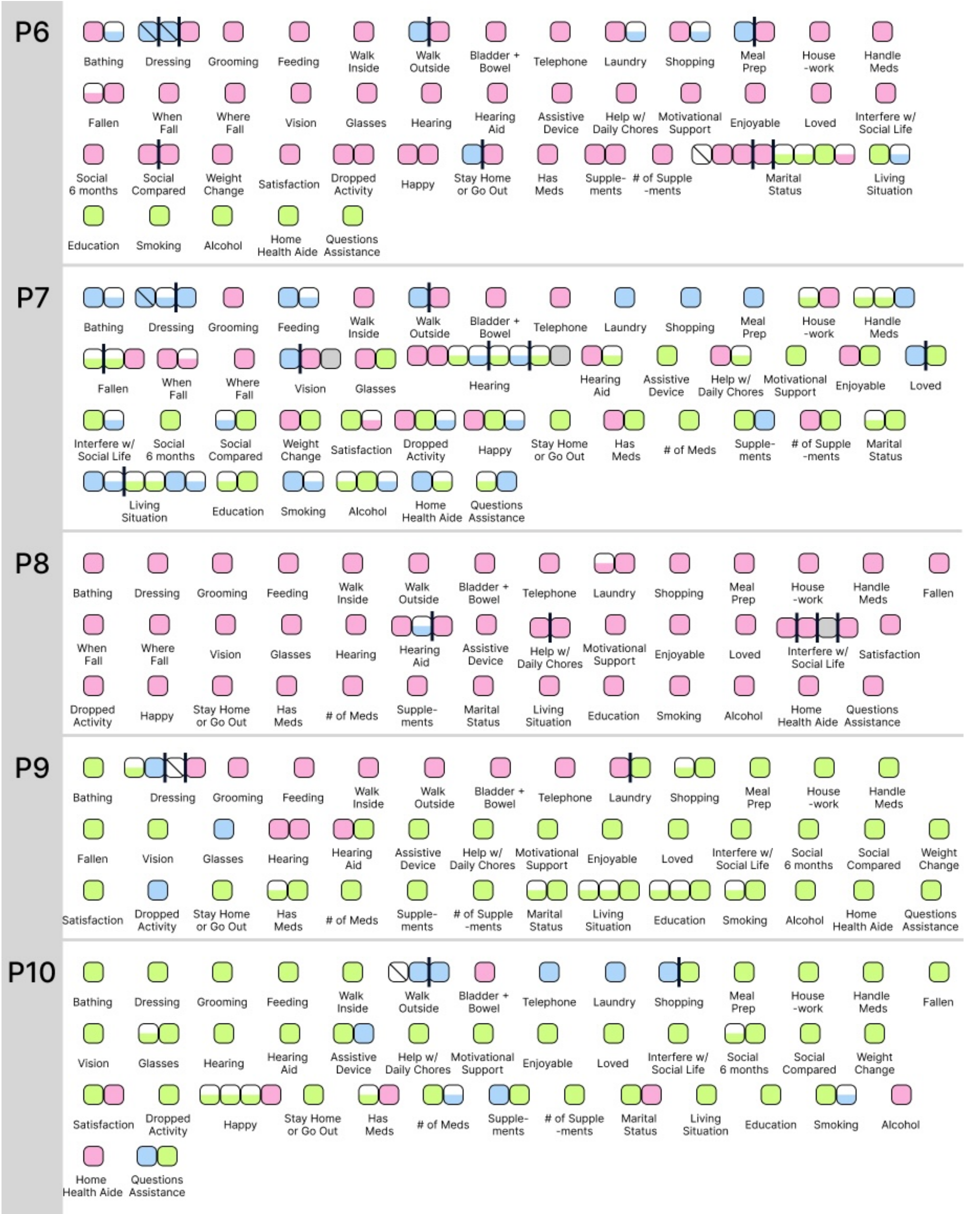


Fig. 8. Interaction sequence part II.

6.2 Voice modality elicited unstructured, conversational responses, which our system was able to handle

As introduced in the previous section, the speech modality elicited unstructured, conversational responses from some participants. Instances of conversational responses were typically highly context dependent or implied in the response (see Table 3 for an overview of open-ended answer categories). P2 and P7 (87 and 99 years old respectively) would give especially context-rich answers. For example when asked about how limited her dressing has been, P7 answered: *"Well, I do have a problem hooking my bra at times"*, to which the system correctly categorized her response as "*limited a little*". In a different question, asking, *"how often do you have someone to love and make you feel wanted?"* she responded, *"my daughter calls me every day, so that makes me feel loved"*. Only catching the first part of her response, our system was able to recognize that a positive answer was implied and attempted to verify this assumption by responding, *"that's wonderful to hear about your daughter. So would you feel loved and wanted most of the time?"* P2's answer to the same question contrasted P7's, stating *"oh, no, I do it by myself. I don't need anybody to tell me that I am wanted. Sorry."* To which our system responded, *"I see, it is important to have self love"* before asking the next question, which appropriately addressed the additional context provided by P7.

In some semi-structured interactions, we observed participants make use of the voice modality to give responses that were more specific than the provided options. For example when asked, *"how limited has your dressing been?"* P1 responded *"75 percent"*, expressing more nuance that made the task of categorizing his response as limited "a little" or "a lot" more difficult. Our system appropriately inferred that it was one of the two options with limitations, and asked for further clarification about which, *"I understand you are experiencing some limitations. Would you say that your dressing is limited a lot or just limited a little?"* Similarly, the voice modality allowed P10 to answer *"I never smoke"* to a smoking question where the answer choices were *"not at all"*, *"at least once"*, and *"every day"*, and then appropriately categorize it as *"not at all"*. As a whole, our system's LLM-backed speech modality expanded interaction possibilities without creating conversational breakdowns for participants.

6.3 Different modalities complemented each other, and helped conversations from going awry

On several occasions, we observed participants rely on the alternative modality to the one that they were using to recover from an interaction error. For example, P2, P3, and P10 used the touch input modality to select an answer when what they were saying was not appropriately interpreted by the system, whether it was due to a transcription error or because they said their answer while the system was speaking. For example, when asked about her marital status, P2 (English not her first language) answered *"widowed"* and the system transcribed it to *"video"*, so it repeated the question. She repeated her answer with a stronger enunciation, yet the same error occurred. By the third attempt, she used the buttons on the screen for the first time (after approximately 15 minutes of having been interacting only using the speech modality). This was a particularly important alternative path for P2, because this was an emotionally painful question for her, as she expressed during the follow-up interview. In instances requiring the system to re-ask a question (a total of 40 vertical black lines in the interaction sequence shown in Figures 7 and 8), alternative modalities offered participants options in error repair, especially through interaction style switching (a total of nine switches portrayed by changes in square colors green to pink or blue, or vice versa for the same question). Similarly, for the speech modality, there were ten instances of rewording responses, as depicted by the changes from pink to blue or vice versa for the same question.

Speech served as an alternative when button presses did not register. P3, P7, and P10 said their answers aloud when their tapping gestures were not registered by the iPad. Finally in cases where the provided answer choices were not sufficient, the voice modality offered means to give a different response (e.g., P4, P9, and P10). For example, when asked *"how is your vision?"*, P4 responded *"fair with glasses"*. Relatedly, when asked, *"do you wear glasses?"*, P9, who had mostly been using buttons for the previous set of questions, hesitated as her fingers hovered over the screen, and instead of tapping either of the provided buttons "yes" or "no," she responded *"sometimes"*. P10 also said *"sometimes"* when asked *"do you use any assistive devices to get around, such as a cane, walker, crutches, or wheelchair?"*

6.4 Different modalities increased accessibility, but also made the system more cumbersome to some

P1 was blind, and had to fully rely on the audio modality of our system to complete the GA independently. During our first session, we noticed a severe heuristic violation for him: the system provided a visual cue of its status processing the information, but no audio cue. As a result, he thought the system had not heard his response and repeated himself one or more times while the display showed a spinning icon saying "one moment please." The system was also having another issue that made us stop the session, and, with the participant's permission, restart it the next day[8]. While we fixed the issue (something was wrong with the iPad we were using), we also added an audio cue of typing sounds to accompany visual cues. The next day, he did not experience those usability issues and was able to successfully complete the full assessment. In our post-completion interview, he remarked that for blind people like him the support our system provides to fill out applications "*is just a prayer answered.*"

While the interaction with P1 showed our system's ability to fully work with only the audio modality, P4's and P10's interactions demonstrated the strength of the multiple modalities. P4, who had poor hearing and used hearing aids, and had fair vision with glasses, explained how the many modalities helped her, "*it was wonderful that the person was speaking and I could also read it because I was going back and forth. I mean, I was hearing so that validated one input to my mind, but then I could read it and then if I didn't, if say my mind wasn't catching it, I validated it.*" P4 completed the assessment using our system without any issues despite her audio and visual impairments, and as she states, the multiple modalities enhanced her experience. P10 (Hungarian accent, arthritis) expressed similar feedback, and when asked by the interviewer to describe her experience switching between voice and touch input, she stated: *"it's providing comfort because—this is good actually—sometimes I would give the answer, what the machine would not evaluate how I think or would not understand what I am thinking. And also the physical comfort."* Throughout the interaction P10 would often rub her hands in discomfort, and as she later expressed, *"depending also on the condition of my hands [because] my finger hurt, then I'm giving voice answer."* This resulted from her arthritis condition, which as she explained, limits her use of technology: *"I cannot work too much on the computer either, because my fingers are hurting."* She also motioned to her frozen shoulder during the interaction to the interviewer, indicating that it was limiting her mobility to some degree. Like P4, multiple modalities enhanced P10's experience by being able to cater to multiple limitations, both on her part and on our system's (see discussion on conversational repair in Section 6.3).

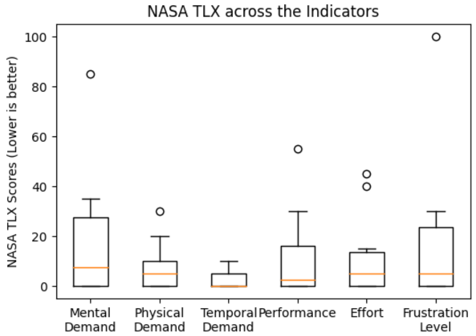


Fig. 9. Box plot of NASA TLX results. Excluding outliers, all means are under ten on a 100 scale. Mental demand and frustration levels have the largest box sizes, and physical and temporal demands the smallest ones. Note that a lower score is better.

[8] This first attempt is included in the qualitative analysis and depicted in Figure 7, but omitted for the quantitative analysis for consistency across participants.

6.5 Challenges with personalizing pace, addressing transcription errors, and asking difficult questions.

In this section we describe three major challenges we observed, including adapting to various pace needs and preferences, transcription errors, and issues with the question design.

6.5.1 Adapting the system's pace. Though participants expressed a positive experience stemming from an inclusion-centered system design, participants who did not experience limitations found the system irritating (i.e. P6, and P8). Under the impression that the questionnaire must be completed only by voice, P6 often answered before the system finished dictating the answer choices (a useful auditory cue for blind or visually impaired individuals) resulting in repeated answers and longer waiting times. Consequently, though she was one of the youngest participants (64 years old) she took the second longest amount of time to complete the questionnaire (approximately 18 minutes). P8, under the same impression, completed the questionnaire by voice without any difficulties but later commented to *"kill the voice"* as it *"slows the whole process down"*. Additionally, he did not appreciate the commentary in the machine's responses. P8 remarked during his interview, *"I was frustrated by it because it was so slow. And I don't care for the thing repeating what I said."* Note, repeating the answer establishes conversational grounding and allows for error repair. Both P6 and P8 indicated that given the option to do it again, they would only complete it by touch, P6 saying *"if you can read it, and you can just do it immediately... Why wait for all of that? All the different choices?"*

Other participants, in particular those with higher degrees of frailty, did not seem to mind the prolonged wait times (e.g., P1, P4, P5, and P7). For example, P5 (83 years old, web developer), quickly grasped how the system worked and was able to easily use the audio, visual, and touch modalities. Even though his body language suggested some frustration with the processing times (e.g., sticking his tongue out), he was tolerant of the delays. For example, by reviewing his interaction video we were able to count that he waited 27 seconds before performing self-repair after our system had not captured his response. Specifically, he said his answer before the device was done talking. This meant that his answer was not recorded. After 27 seconds he repeated his response, which the device then successfully captured. In conversational time, this is an unusually long silence; however, this was not a remarkable moment for P5 as he did not bring it up in any of the conversations after the interaction.

Finally, in contrast to the previous examples, P4 preferred for the voice to be *"just a tad bit slower"*, given that, as she explained, *"with dyslexia, you don't grasp things (as I understand it) as a norm. But you're speeding everything up in your mind to grasp, and so you're actually wearing yourself out to keep even with what is coming at you."* For P4 and P7 (the participant who took the longest to complete the questionnaire), a slower pace could be highly beneficial, and a system like ours can serve them by having abundant patience. These examples demonstrate that though the pace of our system generally worked well for many participants, a predetermined pace does not work well for everyone—there were participants at both ends of the spectrum who preferred a faster or slower system.

6.5.2 Recovering from transcriptions errors. Our system often transcribed speech incorrectly. Sometimes the LLM successfully handled these errors, in a way that made them unnoticeable from a user perspective. For example when P2 (English is not her first language) was asked *Are you taking any "supplements including herbals and vitamins*", she responded "*Yes, I take vitamin [sic]*". Though our system registered "*Yes, I take right tommy*", it successfully handled her answer and responded, "*Great, it's good to keep track of what you're taking.*" In other occasions, it gracefully performed self-repair by establishing conversational grounding and asking participants for the piece of information that it was missing. When P5 responded "*not limited*" to the question, "*How much has your bladder and bowel control been limited by your health condition*", our system registered "*but*". Consequentially unable to categorize his response, our system responded "*I understand that this might be a sensitive topic*" and asked the question again. However, on some occasions, the repairs harmed the interactions. For example, P2 was emotionally distraught when answering *"widowed"* and *"video"* registered, and our system kept asking for her marital status.

6.5.3 Some questions affected participants negatively. We observed that our participants' reactions to questions ranged from neutral, to surprising, to (on the most extreme end) emotionally triggering. The majority of questions did not affect participants, as observed through their body language. Rather, most emotional responses we saw were a reaction to the voice and the content of extraneous commentary, or the need to perform conversational repair (which participants like P6 found frustrating). For example, P4 responded *"Poor"* when our system asked her about her hearing, to which our system responded *"I'm sorry, I didn't quite catch that. . . "* and proceeded to restate the question. This caused P4 to laugh, perhaps finding it humorous that our system also had poor hearing, or potentially releasing tension. There were some questions that surprised participants. For example, for the question, *"How much has your bladder and bowel control been limited by your health condition"*, many participants were shocked (e.g., P7 opened her mouth and was momentarily speechless before she laughed). P1 commented that the question seemed *"intrusive,"* stating *"I don't see why it should be questioned."* Along the same lines, P2's interaction with the marital status question prompted her to later ask the interviewer *"Why do they want to know if my husband died?"* This indicates that the lack of communication on the qualifications of the questions (why a certain question might be included in the questionnaire) and a lack of contextual awareness of the participant's reaction are limitations of our system. The lack of contextual awareness of the participant's reaction was also a problem that occurred with P9. When our system began to ask her questions in the Social Section of the questionnaire (refer to Table 6), she mentioned that she found these questions to be very *"personal towards the reality that these questions might be something that [she needs] to answer in the future."* She mentioned that the question, *"How often do you have someone to love and make you feel wanted?"* was particularly triggering since that created a lot of sadness with the loss of her sister. After her interaction with our system, she expressed that she was *"too sad to continue,"* and rescheduled her interview for a couple of days later. Clarifying why these items are clinically important, or offering participants the option to skip some questions might mitigate the feelings of intrusiveness.

6.6 Participants had diverse ideas for other applications for this system

Finally, when asked about other settings in which our system could be used, we received a wide range of feedback stemming from the diverse backgrounds that our participants had. P3 and P5, both war veterans, offered application ideas that would improve their experience in the VA's health program. For example, P3 indicated that our system could help support people with mental health needs. Participants saw other potential applications of a system with our capabilities, including: enhancing the accessibility of online dating sites for older adults (P5), ordering food at restaurants for blind people (P1), and supporting learners with ADHD through multiple modalities (P1).

6.7 Quantitative results

Here we quantitatively evaluate our system's performance based on its accuracy and participant questionnaire responses.

6.7.1 System performance. A nurse practitioner examined the videos of participant interactions and reported the responses she would have registered in a clinical setting. When comparing her answers with those registered in our system's database by treating the nurse practitioner's responses as ground truth, 92.8% matched exactly, including correctly skipped questions. The remaining 7.2% (42 errors) fall within four major categories: 1) the system incorrectly skipped questions, likely because it deemed them irrelevant based on prior user answers (12 errors); 2) the system recorded an answer as "null" in the database (12 errors); 3) the content displayed on the

Measure [range]	Our study (N = 10), M [SD]	White low error, M [SD]	Black low error, M [SD]
NASA TLX [1-100]	11.9 [10.0]	n/a	n/a
SUS [1-100]	71.8 [15.7]	n/a	n/a
Self-Consciousness [1-7]	4.8 [2.4]	4.9 [0.8]	4.3 [0.8]
Transportation [1-7]	5.7 [1.6]	4.3 [0.7]	4.3 [0.3]
Technology Evaluation [1-7]	5.3 [1.0]	4.8 [0.9]	5.3 [0.5]

Table 5. Summary of quantitative results (N=10) without excluding outliers. The two columns on the right show the means and averages for the same measures for Black and white participants in the low error interaction condition for a study with voice assistants ran by Wenzel et al. [96].

screen did not match the spoken question, likely because the prompt to generate a response did not respect instructions to remain on a question until it reached a classification (6 errors); or 4) the system classified the user response differently than the nurse practitioner (12 errors). Note, the nurse practitioner could not perfectly see where the participants tapped in some questions, which may have increased the number of mismatches. Moreover, in a single instance, the system hallucinated an additional question, asking the user whether they had anyone to confide in, which was not a provided question.

When treating the nurse practitioner and the system as independent raters, we obtain almost perfect or perfect agreement in the majority of the questions based on the guidelines from Landis and Koch [58], with a few exceptions showing substantial agreement, and one showing fair agreement[9]. See Table 6 in Appendix A for each question's Cohen's Kappa.

6.7.2 Self-reported measures. The questionnaire results are summarized in Table 5. The NASA TLX measure reveals low cognitive load scores (mean: 11.9, standard deviation: 10.0 out of 100.0)[10], and an above average SUS score of 71.8 out of 100.0[11].

Moreover, our participants' self-consciousness, transportation, and technology evaluation measures were all better than the combined measures for white and Black participants in the low error rate condition of an experimental study conducted by Wenzel et al. [96] on interactions with voice assistants. Even though we do not analyze race as an independent factor, especially given only one participant in our sample was not white, we provide these measures as a benchmark. These comparisons provide an indication of the performance of our system compared to experiences with similar devices. A deeper analysis of the NASA TLX data, shown in the boxplot in Figure 9, reveals outliers in our data, providing a richer view of the scores. Note, P9, who is accountable for half of the outliers, was the only participant who did not complete the NASA TLX immediately after her interaction with our system, because she requested a break. Thus, she completed the second part of the study on a different day.

[9] Two of three mismatches here were for questions where participants used buttons, and the nurse made a guess about which button was pressed.

[10] A meta-analysis of NASA-TLX global workload scores found that for daily activities, such as completing telephone inquiries, a score of 11.9 is on the bottom quartile [35].

[11] https://measuringu.com/sus/

7 DISCUSSION

Despite the wide range of literature about multimodal interfaces to improve accessibility or support aging in place [23, 71, 98], very little is known about how to create holistic multimodal interfaces that support inclusion. To the contrary, much is known about how often voice-based interfaces fail to live up to their promises [8, 18, 73, 90]. Simultaneously, prior literature in the medical field calls for innovative solutions to help frail patients complete the GA without assistance [1, 25]. In this work, we respond to these needs by carefully building a system that overcomes many of the issues surfaced in prior work by following a holistic framework, and innovating with the needs of a group that has been historically marginalized in the development of technology at the forefront of our design process. As a result, we set a foundation for using LLMs for creating inclusive and scalable digital forms for all.

In this section, we further discuss how our findings address our research questions. We first describe the opportunities this work opens up, especially how well it was received by older adults in need of assistance (RQ1). Second, we discuss the challenges entailing the design of a holistic multimodal systems, and lay out three main areas for future research (RQ1). Finally, we generate design guidelines based on our findings related to RQ2. We note that a key difference between our system and other multimodal systems is its focus on serving people who have been left behind with the existing trend towards digitization.

7.1 Opportunities for holistic multimodal systems for digital forms

Seven participants who currently require paper-based methods or human assistance stand to greatly benefit from the independence our system grants. Our work is serving a gap that deeply impacts users whose needs have been overlooked despite extraordinary technological advances that allow us to address those needs. Our system may not only grant more or complete independence to those who currently need human assistance, but may also increase the amount of agency they have over their personal health data. For those that currently use paper-based forms, our system provides a new option for when those forms may no longer be available. The three participants that are already served by existing digital form filling mechanisms would be relatively unaffected by our system, especially if optimizations to adapt the system to operate at a quicker pace were made.

Our system performed better for participants that typically require assistance, than it did for those already served by existing form filling mechanisms. This is specifically because it had slow processing times, and talked too much for some participants' preferences. Addressing these issues is relatively simple compared to building a system that meets the needs of those currently excluded. Our system performance evaluation shows high accuracy for an early stage system with clear signals for how to further improve it through fine tuning. Moreover, the quantitative data from the post-completion questionnaires, despite being preliminary in terms of not offering a scientifically valid comparison, shows low cognitive load levels, and above-average usability scores, which is especially remarkable given that our system is a research artifact, not a commercial product. Moreover, when comparing technology evaluation scores to similar interactions with voice assistants in an intentionally low-error condition [96], which is artificially better than reality by underestimating the amount of errors that tend to happen in naturalistic interactions [8, 18, 73, 90], our system's technology evaluation scores are higher. Furthermore, as our participants' expressed, a system like ours could support participating in essential everyday life activities such as dating, ordering food, and learning enjoyed by the wider society. These comments indicate that the implications could extend into many sectors beyond healthcare, including education, transportation, the food and restaurant industry, accounting, and shopping—imagine being able to ask a tax document for clarification on ambiguous questions based a person's specific situation.

7.2 Challenges and future work

Despite the virtues and promises of our system, many challenges remain, opening up interesting areas for future work. Even though we strived to follow Chan et al. [17]'s HMID framework, building a truly holistic system will require more effort. There are not yet well established guidelines about how to handle multiple, simultaneous inputs, and turn-taking between humans and a conversational system with the technology constraints presently at hand. Moreover, these challenges will increase as much needed modalities (e.g., facial expressions, gestures, and audio-prosodic features [22]) are integrated into holistic multimodal systems. Our study has highlighted the importance of three major areas for future work: pace and content personalization, physical form factors, and speech-based questionnaire best practices; which we describe below.

7.2.1 Pace and content personalization. Chan et al. [17] added a fourth principle to their framework after conducting a storyboard-based user study: personalization. We saw the need for this principle in our findings as well, specifically for pace and content. While some participants needed the system to be slower and more patient, others wanted it to move quicker. Future work could investigate how a holistic multimodal system might dynamically adapt to diverse pace and content needs. For example, a user might tell it "I want to be done with this as quickly as possible," or, alternatively, "please go slowly and provide more examples." Similarly, a user may express content preferences. P8, who did not like it when the system commented on his responses, may request it not to do so. For people who may be experiencing emotional sensitivity to certain topics (e.g., the passing of a spouse), the system may be instructed to avoid questions that may exacerbate their emotional state. Pace and content personalization will likely make systems like ours more usable for more people, and also for the same person in different use cases (e.g., filling a form that is not very important on the go versus one that is critical and needs full focus).

7.2.2 Physical form factors. Additionally, while tablets are multimodal, they are not designed to be able to exclusively function with voice-only input. This created barriers for our system's alternative individual modes to work fully independently of one another. Specifically, we were unable to permanently enable the microphone on Safari, which periodically required touch input. Future work should look into how to overcome these barriers. Similarly, another interesting area for future work is to investigate various hardware options for different use cases of our system that extend beyond the GA, similar to how Bartle et al. [6] looked at voice assistant form factors with design signifiers for device ownership. For example, our system may operate on a wide range of devices, from personal earbuds and wearable displays to public kiosks, similar to ATMs.

7.2.3 Speech-based questionnaire best practices. Finally, as we transition to fully multimodal systems for form filling, we will need more design guidelines. For example, there is a need to understand best practices for converting questions that are better completed visually, such as scales with rich descriptions for each scale level. Similarly, we must figure out best design practices for auto-fill in the speech format. Moreover, it is unclear whether the speech modality made some questions, that are standard questions, seem more intrusive than they might seem in written format. If so, more work is needed to determine how an automated system may ask potentially intrusive, or worse, triggering, questions in a thoughtful, compassionate way. Finally, it is crucial to ensure that questions from already validated questionnaires preserve their validity when adapted to the voice format.

7.3 Design guidelines for independent alternative modalities

We now present design guidelines for creating holistic multimodal systems for form-filling using independent alternative modalities.

7.3.1 Modality redundancy provides needed paths for error recovery. We observed how alternative individual input modes helped participants interact based on their interaction preference at any given moment. In addition, they were able to recover if the current mode failed, in line with prior work [86]. This was a strength of holistic design, in which both modalities worked interchangeably towards the same objective. For example, the speech modality provided a means to give a response when tapping was not being recognized (e.g., one participant pressed "yes" three times and the iPad failed to register the button press, so she said the response aloud as a recovery mechanism), and the touch modality allowed participants to move on to the next question when the speech transcription was failing for one particular word (e.g., “widow” versus “video”). The redundancy helped the shared objective of registering an answer choice, whether voice failed or tapping a button failed. This may be specifically the case for multiple-choice, form-filling use cases in which the answers can be similarly inferred through speech or button presses transmitting comparable answer choices.

7.3.2 Multimodal foundation models may help us fuse more modalities to further decrease errors for the two main alternative modalities. A visual input modality facilitated through a computer vision capable foundation model could take into account a participant's expressions to strengthen conversational grounding. If the user changes from having a happy or neutral expression to one that appears extremely sad, the system may be able to infer that something is emotionally impacting the user and try to perform repair (e.g., by offering to skip a question). Similarly, an automated analysis of the non-verbal qualities of speech could further support these interactions, such as by being able to infer whether a participant is pausing or done speaking. These features may help with turn-taking, such as by supporting verbal interruptions while the system is speaking without having the system listen to itself, a major challenge we faced when determining system states.

7.3.3 Different modalities can adopt separate roles, as either primary or supporting information sources. The input modes described in the guideline above can serve as *supporting information sources* for the primary alternate input modes: speech and touch. That is, facial expressions and pauses in speech, by themselves, may not help a user answer a question, but they can inform the interaction or provide additional information for the system to determine how to categorize a response. Similarly, and in line with prior work describing “unification” [19], the primary alternate input modes could be fused in such a way that one serves as a primary input mode and one as a secondary one. For example, when P3 tapped the “advanced degree” button as he verbally expressed, *“I'm smart as hell”*, the system could determine that the button press is the primary input mode, and the speech can serve as supporting information to validate the selection. That is, if P3 had said, “I had to quit college”, the contradiction may trigger the system to confirm that “advanced degree” was the participant's intended choice.

7.3.4 The speech modality will elicit unstructured responses that LLMs can handle, and can also be used to refine the questionnaire. The speech modality provided a means to give an unstructured response that was not offered (e.g., “sometimes,” instead of only “yes” or “no”). If the system asked, “do you wear glasses?” and a patient wears glasses only sometimes, how might the system help them provide the response that will be most useful to the patient's care team? Unstructured responses may leave design traces that could be used to refine the instrument itself. That is, the question might be revised to state, “do you wear glasses? Say yes, if you wear glasses even occasionally.” Alternatively, the answer choices provided could include, “yes, always,” “just for specific activities,” and “no.”

7.3.5 If a particular modality is interfering with user preferences, the user must be able to opt out of that modality. Given our accessibility-oriented approach to inclusion, we erred on the side of ensuring our system was usable for those who have been historically excluded in the design of digital interfaces, such as older adults with high degrees of medical frailty. However, this in turn created a more burdensome, or annoying experience, for participants who are currently well-served by the existing systems. An accessibility-oriented approach is a critical step towards more inclusive interfaces later on because we will carry these findings forward as we refine the experience for other users. To have a truly inclusive interface, we must ensure that the benefits of our holistic multimodal interface are also helpful to people without special accessibility needs. This can be achieved by giving users more options and creating a system that adapts to individual preferences. For a user who prefers standard digital forms, our system could go silent and morph into a standard form. While a specific modality might always be available, it could be turned off if it is causing a disturbance, in the same way one might watch a movie using only captions and no audio, or using only audio and no captions.

7.4 Limitations

Our study was conducted with a sample of only ten older adults, in one geographic region of the world, and for one use case, the GA. Despite the diversity of interactions we observed in our small participant sample, our results must be interpreted taking into account our study's relatively small scale and homogeneity (e.g., nine of ten participants were white). Relatedly, our system was developed with one narrow, yet highly impactful, use case in mind, to increase the independence of older adult patients with cancer, for medical form filling, and with a vision to supply recovery support later on. However, our system is versatile and can support independence, thus increasing inclusion in a wide range of domains. We did not study use cases that our system may be able to support, such as the ones brought up by our participants, including: supporting blind people placing restaurant orders, making educational experiences more accessible to people with ADHD, or facilitating self-reflection and providing mental health support to veterans. An exciting area for future work will be to explore our system's applicability in different domains with more participants, participant diversity, and geographic locations.

Moreover, activating our system requires touch input to enable the microphone on Safari. This policy is likely meant to protect users' privacy, but in our case, it does not allow us to develop a fully inclusive system. Future work might explore how to build this into voice-activated, LLM-based devices, similar to the Amazon Echo Show, that may allow voice-only activation in a privacy-preserving manner towards more inclusivity.

Finally, our system only provides the core functionality of a fully multimodal system for filling out forms. There are many features and areas for potential improvement that we considered during our iterative design process but were unable to implement due to the extensive need for developer support they would require. With more developer time, we could significantly improve our system's usability, such as by shortening processing times, facilitating speech speed and volume adjustments, enabling users to review and change responses, and allowing system navigation such as by skipping ahead (or backwards) using buttons, reacting to non-verbal interaction cues, listening while speaking, and understanding multiple languages and dialects. These potential improvements were unnecessary to address our main research questions but would be exciting areas for future work to better understand how to further improve equitable usability and inclusion.

8 CONCLUSION

In this work, we built a system that can increase the accessibility, and thus inclusivity, of digital forms for many individuals who have been overlooked in the process of digitizing information systems. Our system utilizes speech and touch inputs, and speech and visual outputs, powered by LLMs, to create a multimodal experience for filling digital forms. We evaluated our system with a diverse group of older adults (N=10), validating its ability to increase independence for some people who would otherwise require assistance. Notably, all of our participants were able to complete a 47-item GA independently, despite challenges including significant vision and hearing impairments and advanced age. From our findings, we generate a set of design guidelines for creating holistic multimodal interfaces for health data entry.

ACKNOWLEDGMENTS

We would like to thank our previous collaborators, especially Professor Deborah Estrin and Dr. Armin Shahrokni, for their contributions establishing and developing the ideas that this work is built on. We would additionally like to thank Professor Monica Lam and Shicheng (George) Liu from the Stanford Open Virtual Assistant Lab for their support with systematically prompting GPT-4. Thanks to Kimi Wenzel and Professor Geoff Kaufman for sharing their questionnaires with us, and for taking time to talk to us about our work. Thanks to Professor Shiela Strauss for her support computing IRR. Thanks to our colleagues in the Interaction Design and Hybrid Digital Physical Spaces groups at Stanford for their feedback throughout the development of our system, in particular to Jackie Yang, Matt Jörke, and Alan Cheng. Finally, thanks to our participants! This work was supported in part by Stanford HAI and a Memorial Sloan Kettering Cancer Center Support Grant/Core Grant (Grant No. P30 CA008748) through funding from the National Cancer Institute, and through the Gordon and Betty Moore Foundation, Grant GBMF9048 to Dr. Fessele. Dr. Andrea Cuadra was additionally supported by a Stanford PRISM Baker Postdoctoral Fellowship.

REFERENCES

[1] Sarah Abdi, Luc de Witte, Mark Hawley, et al. 2020. Emerging technologies with potential care and support applications for older people: review of gray literature. *JMIR aging* 3, 2 (2020), e17286.

[2] Icek Ajzen and Martin Fishbein. 1977. Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological bulletin* 84, 5 (1977), 888.

[3] Abdel Rahman Feras AlSamhori, Jehad Feras AlSamhori, and Ahmad Feras AlSamhori. 2023. ChatGPT Role in a Medical Survey. *High Yield Medical Reviews* 1, 2 (2023).

[4] Ibraheem Altamimi, Abdullah Altamimi, Abdullah S Alhumimidi, Abdulaziz Altamimi, and Mohamad-Hani Temsah. 2023. Artificial Intelligence (AI) Chatbots in Medicine: A Supplement, Not a Substitute. *Cureus* 15, 6 (2023).

[5] Anneliese Arnold, Stephanie Kolody, Aidan Comeau, and Antonio Miguel Cruz. [n. d.]. What does the literature say about the use of personal voice assistants in older adults? A scoping review. 0, 0 ([n. d.]), 1–12. https://doi.org/10.1080/17483107.2022.2065369 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17483107.2022.2065369.

[6] Vince Bartle, Liam Albright, and Nicola Dell. 2023. " This machine is for the aides": Tailoring Voice Assistant Design to Home Health Care Work. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[7] Michael Barz, Mohammad Mehdi Moniri, Markus Weber, and Daniel Sonntag. 2016. Multimodal multisensor activity annotation tool. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 17–20.

[8] Erin Beneteau, Olivia K Richards, Mingrui Zhang, Julie A Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication breakdowns between families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[9] Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. *Journal of medical Internet research* 20, 9 (2018), e11510.

[10] Richard A Bolt. 1980. "Put-that-there" Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. 262–270.

[11] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[12] Eliane M Boucher, Nicole R Harake, Haley E Ward, Sarah Elizabeth Stoeckl, Junielly Vargas, Jared Minkel, Acacia C Parks, and Ran Zilca. 2021. Artificially intelligent chatbots in digital mental health interventions: a review. *Expert Review of Medical Devices* 18, sup1 (2021), 37–49.

[13] Virginia Braun and Victoria Clarke. 2021. *Thematic analysis: A practical guide*. Sage.

[14] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189, 3 (1996), 189–194.

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[16] James C Byers, AC Bittner, and Susan G Hill. 1989. Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary. *Advances in industrial ergonomics and safety* 1 (1989), 481–485.

[17] Eric Chan, Gerry Chan, Assem Kroma, and Ali Arya. 2022. Holistic Multimodal Interaction and Design. In *International Conference on Human-Computer Interaction*. Springer, 18–33.

[18] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[19] Philip R Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow. 1997. Quickset: Multimodal interaction for distributed applications. In *Proceedings of the fifth ACM international conference on Multimedia*. 31–40.

[20] Diane J Cook, Juan C Augusto, and Vikramaditya R Jakkula. 2009. Ambient intelligence: Technologies, applications, and opportunities. *Pervasive and mobile computing* 5, 4 (2009), 277–298.

[21] Valerie Crooks, Susan Waller, Tom Smith, and Theodore J Hahn. 1991. The use of the Karnofsky Performance Scale in determining outcomes and risk in geriatric outpatients. *Journal of gerontology* 46, 4 (1991), M139–M144.

[22] Andrea Cuadra, Hyein Baek, Deborah Estrin, Malte Jung, and Nicola Dell. 2022. On Inclusion: Video Analysis of Older Adult Interactions with a Multi-Modal Voice Assistant in a Public Setting. In *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development*. 1–17.

[23] Andrea Cuadra, Jessica Bethune, Rony Krell, Alexa Lempel, Katrin Hänsel, Armin Shahrokni, Deborah Estrin, and Nicola Dell. 2023. Designing Voice-First Ambient Interfaces to Support Aging in Place. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2189–2205.

[24] Andrea Cuadra, Yen-Hao Chen, Kae-Jer Cho, Deborah Estrin, and Armin Shahrokni. [n. d.]. Introducing the v-RFA, a voice assistant-based geriatric assessment. 13, 8 ([n. d.]), 1253–1255. https://doi.org/10.1016/j.jgo.2022.05.001 Publisher: Elsevier.

[25] Andrea Cuadra, Amy L Tin, Gordon Taylor Moffat, Koshy Alexander, Robert J Downey, Beatriz Korc-Grodzicki, Andrew J Vickers, and Armin Shahrokni. 2023. The association between perioperative frailty and ability to complete a web-based geriatric assessment among older adults with cancer. *European Journal of Surgical Oncology* 49, 3 (2023), 662–666.

[26] Fred D Davis. 1993. User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. *International journal of man-machine studies* 38, 3 (1993), 475–487.

[27] Hannes Devos, Kathleen Gustafson, Pedram Ahmadnezhad, Ke Liao, Jonathan D Mahnken, William M Brooks, and Jeffrey M Burns. 2020. Psychometric properties of NASA-TLX and index of cognitive activity as measures of cognitive workload in older adults. *Brain sciences* 10, 12 (2020), 994.

[28] Rishika Dwaraghanath, Rahul Majethia, and Sanjana Gautam. 2023. ECHO: An Automated Contextual Inquiry Framework for Anonymous Qualitative Studies using Conversational Assistants. *arXiv preprint arXiv:2312.07576* (2023).

[29] Bassem Elsawy and Kim E Higgins. 2011. The geriatric assessment. *American family physician* 83, 1 (2011), 48–56.

[30] Allan Fenigstein, Michael F Scheier, and Arnold H Buss. 1975. Public and private self-consciousness: Assessment and theory. *Journal of consulting and clinical psychology* 43, 4 (1975), 522.

[31] Olga T Filippova, Dennis S Chi, Kara Long Roche, Yukio Sonoda, Oliver Zivanovic, Ginger J Gardner, William P Tew, Roisin O'Cearbhaill, Saman Sarraf, Sung Wu Sun, et al. 2019. Geriatric co-management leads to safely performed cytoreductive surgery in older women with advanced stage ovarian cancer treated at a tertiary care cancer center. *Gynecologic oncology* 154, 1 (2019), 77–82.

[32] William W Gaver. 1997. Auditory interfaces. In *Handbook of human-computer interaction*. Elsevier, 1003–1041.

[33] Melanie C Green and Timothy C Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology* 79, 5 (2000), 701.

[34] Melanie C Green and Kaitlin Fitzgerald. 2017. Transportation theory applied to health and risk messaging. In *Oxford research encyclopedia of communication*.

[35] Rebecca A Grier. 2015. How high is high? A meta-analysis of NASA-TLX global workload scores. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 59. SAGE Publications Sage CA: Los Angeles, CA, 1727–1731.

[36] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field methods* 18, 1 (2006), 59–82.

[37] Sadrieh Hajesmaeel-Gohari, Firoozeh Khordastan, Farhad Fatehi, Hamidreza Samzadeh, and Kambiz Bahaadinbeigy. 2022. The most used questionnaires for evaluating satisfaction, usability, acceptance, and quality outcomes of mobile health. *BMC Medical Informatics and Decision Making* 22, 1 (2022), 22.

[38] Christina N. Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. [n. d.]. "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking. In *CHI Conference on Human Factors in Computing Systems* (New Orleans LA USA, 2022-04-29). ACM, 1–15. https://doi.org/10.1145/3491102.3501995

[39] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[40] Peter Hoonakker, Pascale Carayon, Ayse P Gurses, Roger Brown, Adjhaporn Khunlertkit, Kerry McGuire, and James M Walker. 2011. Measuring workload of ICU nurses with a questionnaire survey: the NASA Task Load Index (TLX). *IIE transactions on healthcare systems engineering* 1, 2 (2011), 131–143.

[41] Mohammed Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. 2013. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. 697–706.

[42] Arti Hurria, Chie Akiba, Jerome Kim, Dale Mitani, Matthew Loscalzo, Vani Katheria, Marianna Koczywas, Sumanta Pal, Vincent Chung, Stephen Forman, et al. 2016. Reliability, validity, and feasibility of a computer-based geriatric assessment for older adults with cancer. *Journal of oncology practice* 12, 12 (2016), e1025–e1034.

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

[44] Brigitte Jordan and Austin Henderson. 1995. Interaction analysis: Foundations and practice. *The journal of the learning sciences* 4, 1 (1995), 39–103.

[45] Sidney Katz. 1983. Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living. *Journal of the American Geriatrics Society* 31, 12 (1983), 721–727.

[46] Shahedul Huq Khandkar. 2009. Open coding. *University of Calgary* 23 (2009), 2009.

[47] Sunyoung Kim and Abhishek Choudhury. [n. d.]. Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study. 124 ([n. d.]), 106914. https://doi.org/10.1016/j.chb.2021.106914

[48] Tae Soo Kim, DaEun Choi, Yoonseo Choi, and Juho Kim. 2022. Stylette: Styling the web with natural language. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[49] Young-Ho Kim, Diana Chou, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E Conroy, Hernisa Kacorri, and Eun Kyoung Choe. 2022. Mymove: Facilitating older adults to collect in-situ activity labels on a smartwatch with speech. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–21.

[50] Young-Ho Kim, Sungdong Kim, Minsuk Chang, and Sang-Woo Lee. 2022. Leveraging Pre-Trained Language Models to Streamline Natural Language Interaction for Self-Tracking. *arXiv preprint arXiv:2205.15503* (2022).

[51] Young-Ho Kim, Bongshin Lee, Arjun Srinivasan, and Eun Kyoung Choe. 2021. Data@ hand: Fostering visual exploration of personal data on smartphones leveraging speech and touch interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[52] Oscar Kjell, Katarina Kjell, and H Andrew Schwartz. 2023. AI-based large language models are ready to transform psychological health assessment. (2023).

[53] Oscar NE Kjell, Sverker Sikström, Katarina Kjell, and H Andrew Schwartz. 2022. Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific reports* 12, 1 (2022), 3918.

[54] Rafal Kocielnik, Raina Langevin, James S George, Shota Akenaga, Amelia Wang, Darwin P Jones, Alexander Argyle, Callan Fockele, Layla Anderson, Dennis T Hsieh, et al. 2021. Can I Talk to You about Your Social Needs? Understanding Preference for Conversational User Interface in Health. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. 1–10.

[55] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[56] Małgorzata Kowalska, Aleksandra Gładyś, Barbara Kalańska-Łukasik, Monika Gruz-Kwapisz, Wojciech Wojakowski, and Tomasz Jadczyk. [n. d.]. Readiness for Voice Technology in Patients With Cardiovascular Diseases: Cross-Sectional Study. 22, 12 ([n. d.]), e20456. https://doi.org/10.2196/20456

[57] Saewon Kye, Junhyung Moon, Juneil Lee, Inho Choi, Dongmi Cheon, and Kyoungwoo Lee. 2017. Multimodal data collection framework for mental stress monitoring. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 822–829.

[58] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[59] M Powell Lawton and Elaine M Brody. 1969. Assessment of older people: self-maintaining and instrumental activities of daily living. *The gerontologist* 9, 3_Part_1 (1969), 179–186.

[60] James R Lewis. 2018. The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction* 34, 7 (2018), 577–590.

[61] Fabio Masina, Valeria Orso, Patrik Pluchino, Giulia Dainese, Stefania Volpato, Cristian Nelini, Daniela Mapelli, Anna Spagnolli, and Luciano Gamberini. [n. d.]. Investigating the Accessibility of Voice Assistants With Impaired Users: Mixed Methods Study. 22, 9 ([n. d.]), e18431. https://doi.org/10.2196/18431 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[62] Fabio Masina, Valeria Orso, Patrik Pluchino, Giulia Dainese, Stefania Volpato, Cristian Nelini, Daniela Mapelli, Anna Spagnolli, and Luciano Gamberini. 2020. Investigating the Accessibility of Voice Assistants With Impaired Users: Mixed Methods Study. *Journal of medical Internet research* 22, 9 (2020), e18431.

[63] Jesús Mateos-Nozal, Nuria Pérez-Panizo, Carlota Manuela Zárate-Sáez, María Nieves Vaquero-Pinto, Cristina Roldán-Plaza, Manuel Vicente Mejía Ramírez-Arellano, Elisabet Sánchez García, Alejandro Javier Garza-Martínez, and Alfonso José Cruz-Jentoft. 2022. Proactive geriatric comanagement of nursing home patients by a new hospital-based liaison geriatric unit: a new model for the future. *Journal of the American Medical Directors Association* 23, 2 (2022), 308–310.

[64] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[65] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[66] Supriya G Mohile, William Dale, Mark R Somerfield, Mara A Schonberg, Cynthia M Boyd, Peggy S Burhenn, Beverly Canin, Harvey Jay Cohen, Holly M Holmes, Judith O Hopkins, et al. 2018. Practical assessment and management of vulnerabilities in older patients receiving chemotherapy: ASCO guideline for geriatric oncology. *Journal of Clinical Oncology* 36, 22 (2018), 2326.

[67] Ashwin Nayak, Sharif Vakili, Kristen Nayak, Margaret Nikolov, Michelle Chiu, Philip Sosseinheimer, Sarah Talamantes, Stefano Testa, Srikanth Palanisamy, Vinay Giri, et al. 2023. Use of Voice-Based Conversational Artificial Intelligence for Basal Insulin Prescription Management Among Patients With Type 2 Diabetes: A Randomized Clinical Trial. *JAMA Network Open* 6, 12 (2023), e2340232–e2340232.

[68] Laurence Nigay and Joëlle Coutaz. 1993. A design space for multimodal systems: concurrent processing and data fusion. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 172–178.

[69] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Commun. ACM* 42, 11 (1999), 74–81.

[70] Stuart G Parker, P McCue, K Phelps, A McCleod, S Arora, K Nockels, S Kennedy, H Roberts, and S Conroy. 2018. What is comprehensive geriatric assessment (CGA)? An umbrella review. *Age and ageing* 47, 1 (2018), 149–155.

[71] Anne Marie Piper, Nadir Weibel, and James D Hollan. 2010. Introducing multimodal paper-digital interfaces for speech-language therapy. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*. 203–210.

[72] Alisha Pradhan, Amanda Lazar, and Leah Findlater. 2020. Use of intelligent voice assistants by older adults with low technology use. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 4 (2020), 1–27.

[73] Alisha Pradhan, Kanika Mehta, and Leah Findlater. [n. d.]. "Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018-04-21) *(CHI '18)*. Association for Computing Machinery, 1–13. https://doi.org/10.1145/3173574.3174033

[74] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[75] Kathryn E Ringland, Rodrigo Zalapa, Megan Neal, Lizbeth Escobedo, Monica Tentori, and Gillian R Hayes. 2014. SensoryPaint: a multimodal sensory intervention for children with neurodevelopmental disorders. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 873–884.

[76] Marcia Y Shade, Kyle Rector, Rasila Soumana, and Kevin Kupzyk. 2020. Voice assistant reminders for pain self-management tasks in aging adults. *Journal of gerontological nursing* 46, 10 (2020), 27–33.

[77] Armin Shahrokni, Amy L Tin, Saman Sarraf, Koshy Alexander, Steve Sun, Soo Jung Kim, Sincere McMillan, Heidi Yulico, Farnia Amirnia, Robert J Downey, et al. 2020. Association of geriatric comanagement and 90-day postoperative mortality among patients aged 75 years and older with cancer. *JAMA Network Open* 3, 8 (2020), e209265–e209265.

[78] Armin Shahrokni, Bella Marie Vishnevsky, Brian Jang, Saman Sarraf, Koshy Alexander, Soo Jung Kim, Robert Downey, Anoushka Afonso, and Beatriz Korc-Grodzicki. 2019. Geriatric assessment, not ASA physical status, is associated with 6-month postoperative survival in patients with cancer aged 75 years. *Journal of the National Comprehensive Cancer Network* 17, 6 (2019), 687–694.

[79] Sverker Sikström, Alfred Pålsson Höök, and Oscar Kjell. 2023. Precise language responses versus easy rating scales—Comparing respondents' views with clinicians' belief of the respondent's views. *Plos one* 18, 2 (2023), e0267995.

[80] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617* (2023).

[81] Arjun Srinivasan, Mira Dontcheva, Eytan Adar, and Seth Walker. 2019. Discovering natural language commands in multimodal interfaces. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 661–672.

[82] Arjun Srinivasan, Bongshin Lee, Nathalie Henry Riche, Steven M Drucker, and Ken Hinckley. 2020. InChorus: Designing consistent multimodal interactions for data visualization on tablet devices. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[83] Arjun Srinivasan, Bongshin Lee, and John Stasko. 2020. Interweaving multimodal interaction with flexible unit visualizations for data exploration. *IEEE Transactions on Visualization and Computer Graphics* 27, 8 (2020), 3519–3533.

[84] Lucy Suchman and Lucy A Suchman. 2007. *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.

[85] L Suchman and R Trigg. 1991. Understanding Practice: Video as a Medium for Reflection and Design. Design at Work: Cooperative Design of Computer Systems. M. Kyng.

[86] Bernhard Suhm, Brad Myers, and Alex Waibel. 2001. Multimodal error correction for speech user interfaces. *ACM transactions on computer-human interaction (TOCHI)* 8, 1 (2001), 60–98.

[87] De Wet Swanepoel, Vinaya Manchaiah, and Jan-Willem A Wasmann. 2023. The Rise of AI Chatbots in Hearing Health Care. *The Hearing Journal* 76, 04 (2023), 26–30.

[88] John C Tang. 1991. Findings from observational studies of collaborative work. *International Journal of Man-machine studies* 34, 2 (1991), 143–160.

[89] Janani Thillainadesan, Sarah J Aitken, Sue R Monaro, John S Cullen, Richard Kerdic, Sarah N Hilmer, and Vasi Naganathan. 2022. Geriatric comanagement of older vascular surgery inpatients reduces hospital-acquired geriatric syndromes. *Journal of the American Medical Directors Association* 23, 4 (2022), 589–595.

[90] Milka Trajkova and Aqueasha Martin-Hammond. 2020. " Alexa is a Toy": exploring older adults' reasons for using, limiting, and abandoning echo. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[91] Michael Tsang, George W Fitzmaurice, Gordon Kurtenbach, Azam Khan, and Bill Buxton. 2002. Boom chameleon: simultaneous capture of 3D viewpoint, voice and gesture annotations on a spatially-aware display. In *Proceedings of the 15th annual ACM symposium on User interface software and technology*. 111–120.

[92] Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247* (2021).

[93] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2023. Leveraging large language models to power chatbots for collecting user self-reported data. *arXiv preprint arXiv:2301.05843* (2023).

[94] Zhuxiaona Wei and James A Landay. 2018. Evaluating speech-based smart devices using new usability heuristics. *IEEE Pervasive Computing* 17, 2 (2018), 84–96.

[95] Laurie Weingart, Philip Smith, and Mara Olekalns. 2004. Quantitative coding of negotiation behavior. *International negotiation* 9, 3 (2004), 441–456.

[96] Kimi Wenzel, Nitya Devireddy, Cam Davison, and Geoff Kaufman. 2023. Can Voice Assistants Be Microaggressors? Cross-Race Psychological Responses to Failures of Automatic Speech Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[97] Jacob Wobbrock and Brad Myers. 2006. Trackball Text Entry for People with Motor Impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 479–488. https://doi.org/10.1145/1124772.1124845

[98] Mingrui Ray Zhang, Ruolin Wang, Xuhai Xu, Qisheng Li, Ather Sharif, and Jacob O. Wobbrock. 2021. Voicemoji: Emoji Entry Using Voice for Visually Impaired People. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 37, 18 pages. https://doi.org/10.1145/3411764.3445338

[99] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493* (2022).

[100] Bin Zheng, Xianta Jiang, Geoffrey Tien, Adam Meneghetti, O Neely M Panton, and M Stella Atkins. 2012. Workload assessment of surgeons: correlation between NASA TLX and blinks. *Surgical endoscopy* 26 (2012), 2746–2750.

[101] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625* (2022).

[102] Tamara Zubatiy, Kayci L Vickers, Niharika Mathur, and Elizabeth D Mynatt. 2021. Empowering dyads of older adults with mild cognitive impairment and their care partners using conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

A IRR FOR NURSE PRACTITIONER AND SYSTEM

Table 6. Computations of Cohen’s Kappa between the nurse practitioner and the system by question.

Section	Item	Paired observations	Cohen’s Kappa
ADL	Bathing	10	1.000
	Dressing	10	1.000
	Grooming	10	1.000
	Feeding	10	0.737
	Walk inside	10	1.000
	Walk outside	9	1.000
	Bowel and Bladder	10	0.815
iADL	Telephone	10	0.615
	Laundry	8	0.789
	Shopping	10	0.831
	Meal preparation	8	1.000
	Housework	9	1.000
	Medications	8	1.000
Falls	Fallen	9	1.000
	When fall	9 (5)*	1.000 (1.000)
	Where fall	1 (6)*	1.000 (1.000)
Assistive Devices	Vision	10	1.000
	Glasses	10	1.000
	Hearing	9	1.000
	Hearing aid	10	1.000
	Which device	10	0.857
Social	Help with daily chores	10	1.000
	Emotional support	10	1.000
	Enjoyable	10	1.000
	Love wanted	9	1.000
	Interfere	8	0.830
Social Activities	Social activities six months	6	0.750
	Social activities comparison	6	0.379
Weight	Weight change	8 (7)*	1.000 (1.000)
	Weight loss	-**	-
	Weight gain	-**	-
Emotional Status	Satisfaction	10	1.000
	Dropped activity	10	1.000
	Happy	9	1.000
	Prefer staying home	10	1.000
Medications & Supplements	Has medications	10	1.000
	Number of medications	10 (6)*	1.000 (1.000)
	Has supplements	9	1.000
	Number of supplements	10 (8)*	1.000 (1.000)
Marital Status	Marital status	10	1.000
Lifestyle	Living situation	10	1.000
Educational background	Education	10	0.851
Smoking & Alcohol	Smoking	10	1.000
	Consider quitting	-**	-
	Alcohol	10	1.000
Home Aide Help	Help	10	1.000
Questionnaire Assistance	Assessment taker	10	0.615

* The values not in parentheses include correctly skipped questions due to branching, and the values not in parentheses do not.

** Too few values not skipped to compute Cohen’s Kappa.

B CARE QUESTIONNAIRE ITEMS

Table 7. Questions and answer options.

Section	Question	Answer Choices
ADL	How much has your bathing been limited by your health condition?	Not limited, Limited a little, Limited a lot
	How much has your dressing been limited by your health condition?	
	How much has your grooming been limited by your health condition?	
	How much has your feeding been limited by your health condition?	
	How much has your walking inside the home been limited by your health condition?	
	How much has your walking outside the home been limited by your health condition?	
	How much has your bladder and bowel control been limited by your health condition?	
iADL	How independently can you use the telephone?	Without help, With some help, Unable to do so
	How independently can you do your laundry?	
	How independently can you go shopping?	
	How independently can you prepare meals?	
	How independently can you do housework?	
	How independently can you handle your own medication?	
Falls	Have you ever fallen (lost your balance, tripped, or dropped to the ground without control)?	Yes, No
	When did you last fall?	In the last week, In the last month, In the last 6 months, In the last year, More than 1 year ago
	Where did you last fall?	Home, Outside the home
Assistive Devices	How is your vision?	Excellent, Good, Fair, Poor
	Do you wear glasses?	Yes, No
	How is your hearing?	Excellent, Good, Fair, Poor
	Do you use a hearing aid?	Yes, No
	Do you use any assistive devices to get around such as a cane, walker, crutches, or wheelchair?	Cane, Walker, Crutches, Wheelchair, Prosthesis
Social	How often do you have someone to help you with daily chores if you were sick?	All the time, Most of the time, Some of the time, A little of the time, Not at all
	How often do you have someone to turn to for suggestions about how to deal with a personal problem?	
	How often do you have someone to do something enjoyable with?	
	How often do you have someone to love and make you feel wanted?	
	During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting friends or relatives)?	
Social Activities	Compared to your usual level of social activity, has your social activity during the past 6 months changed because of a change in your physical or emotional condition?	Much less, Somewhat less, Same, Somewhat more, Much more
	Compared to others your age, are your social activities more or less limited because of your physical health or emotional problems?	

Section	Question	Answer Choices
Weight	In the past 6 months, has your weight changed? How much weight have you lost? How much weight have you gained?	Lost weight, No change, Gained weight, I’m not sure Less than 5 pounds, Between 5 and 10 pounds, Between 11 and 20 pounds, More than 20 pounds
Emotional Status	Are you basically satisfied with your life? Have you dropped many of your activities and interests? Do you feel happy most of the time? Do you prefer to stay at home rather than going out and doing new things?	Yes, No
Medications & Supplements	Are you taking any prescribed medications? Approximately how many different prescribed medications do you take each day? Are you taking any supplements including herbals and vitamins? Approximately how many herbs, vitamins, and supplements do you take each day?	Yes, No One to four, Five to ten, More than ten
Marital Status	What is your current marital status?	Married, Divorced, Domestic Partnership, Separated, Single, Widowed
Lifestyle	What is your current living situation?	Living alone, Living with family or partner, Living with someone else, Assisted living facility, Nursing home, Other
Educational Background	What is your highest level of education?	Less than high school diploma, High school diploma, Some college, College graduate, Advanced degree
Smoking & Alcohol	In the past 30 days, have you smoked cigarettes (even a single puff) or used other tobacco products? Are you willing to consider quitting? How many times in the past year have you had more than 4 drinks in a day if you are a woman or 5 in a day if you are a man?	Yes, No Yes, Maybe, No Zero, One or more
Home Aide Help	Do you have home care services now (such as a visiting nurse or home health aide)?	Yes, No
Questionnaire Assistance	Who completed this assessment?	I did it myself, I did it with some help, Someone else did it for me