What Are You Looking Forward to? Deliberate Positivity as a Promising Strategy for Conversational Agents

LIBBY FERLAND and RISAKO OWAN, Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA
ZACHARY KUNKEL and HANNAH QU, Department of Psychology, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA
MARIA GINI, Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA
WILMA KOUTSTAAL, Department of Psychology, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA

Conversational agents (CAs) are one of the most promising technologies for helping older adults maintain independence longer by augmenting their support and social networks. Voice-based technology in particular is especially powerful in this regard due to its accessibility and ease of use. There is also a growing body of evidence supporting the potential use of such technology in mitigating common issues such as loneliness and isolation, particularly for independent older adults aging in place. One of the key challenges for smart technologies deployed in this context is the development and maintenance of long-term user engagement and adoption, which is often addressed by attempting to closely mimic human social interactions. However, the more human-like the system, the more glaring fault conditions become, and the more jarring they are for users. In this study we explore the effectiveness of an alternative conversational strategy meant to encourage users to engage in positive reflection and introspection. We detail the iterative design and implementation of a prototype CA developed to engage in social conversation with older adults on selected topics of interest. We then use this system as part of a multi-method approach to investigate the effect of deliberate positivity as a conversational strategy, including its impact on user impressions and willingness to continue using the CA. Our results from different approaches, including methods such as psycholinguistic analysis, user self-report, and researcher-based coding, paint a promising picture of this conversational design. We show that the deliberate encouragement by a CA of positive conversation and reflection in users has a measurable positive impact on both user enjoyment and desire to continue engaging with a system. We further demonstrate how some user characteristics may amplify this effect, and discuss the implications of these results for the design and testing of future conversational systems for older adults.

Libby Ferland and Risako Owan contributed equally to this research.

This research was supported by an NSF EAGER grant (IIS 1927190).

Authors' Contact Information: Libby Ferland (corresponding author), Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA; e-mail: ferla006@umn.edu; Risako Owan, Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA; e-mail: owan0002@umn.edu; Zachary Kunkel, Department of Psychology, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA; e-mail: kunke166@alumni.umn.edu; Hannah Qu, Department of Psychology, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA; e-mail: qu000016@umn.edu; Maria Gini, Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA; e-mail: gini@umn.edu; Wilma Koutstaal, Department of Psychology, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA; e-mail: kouts003@umn.edu.

ACM 2160-6463/2025/7-ART14
https://doi.org/10.1145/3725738

CCS Concepts: • **Human-centered computing** → **Natural language interfaces**; **Interaction design**; *User studies*; *User models*;

Additional Key Words and Phrases: deliberate positivity, conversational assistants, conversational design, older adults, social conversational systems, user studies, PANAS, LIWC, SUS

ACM Reference format:
Libby Ferland, Risako Owan, Zachary Kunkel, Hannah Qu, Maria Gini, and Wilma Koutstaal. 2025. What Are You Looking Forward to? Deliberate Positivity as a Promising Strategy for Conversational Agents. *ACM Trans. Interact. Intell. Syst.* 15, 3, Article 14 (July 2025), 40 pages.
https://doi.org/10.1145/3725738

1 Introduction

The unprecedented aging of the global population, paired with the rapid pace of the development of increasingly sophisticated intelligent interactive systems, has prompted exploration into many important **research questions (RQs)** in the community. It is estimated that by the year 2050 there will be 2.1 billion people aged 60 or older; conversely, there will be fewer working-age adults available to help meet the care needs of older adults requiring support.[1] Out of many available technologies, conversational intelligent systems are regarded as some of the most promising avenues of technological support for older adults. Voice-activated systems are widely accessible, can be more comfortable to talk to than humans, especially when user questions are repeated, and users can experience similar psychological benefits as they would talking to a human. However, there are some recurring issues that plague the use of conversational technology as support systems, particularly since by definition these technologies would need to be adopted for long-term use and encourage regular ongoing interaction with users. Many conversational strategies have been developed to facilitate this, particularly when it comes to social support to alleviate isolation and depression faced by older community-dwelling adults, but the deliberate and planful use of positivity in conversation with an intelligent system has yet to be explicitly investigated.

Positivity in this conversational context can be defined as an interaction in which someone shares good news or positive events with others. This process is sometimes referred to as personal capitalization in the psychology and human-human communication literature [34, 53]. Capitalization can be thought of as the complement to support seeking, the social process, and coping mechanism of sharing negative events and feelings in order to alleviate stress and other negative emotions [35]. In other words positivity, or capitalization, is the process by which we "spread the love" rather than "share the burden." While capitalization as positivity is a novel question in human-**conversational agent (CA)** interactions, the phenomenon is well-studied in human-human communication, where it has been shown to be greatly beneficial psychologically. Capitalization promotes positive retrospection in sharers ("capitalizers") and increases positive thought patterns [34, 79, 87]. This positive effect can depend to some degree on the reactions of a partner or partners, but the effects in capitalizers can be long-lasting regardless [52]. Like most social processes, capitalization is an iterative process that continues and deepens over time, meaning that mimicking this social process in intelligent systems may help in promoting adoption over the long term. For older adults in particular it is also especially noteworthy that the positive effects of capitalization are even more pronounced in capitalizers experiencing depression or anxiety [4]. Capitalization has also been shown to have a positive impact on health more generally, particularly when it comes to mitigating stress [38, 54]. This makes capitalization a potentially valuable resource for older adults aging in

[1] https://www.who.int/news-room/fact-sheets/detail/ageing-and-health#:~:text=By%202050%2C%20the%20world's%20population,2050%20to%20reach%20426%20million.

place who may be experiencing increased feelings of social isolation and depression, as well as other health concerns.

The use of planful positivity incorporated into older adults' interactions with an exclusively voice-activated social conversational assistant is novel—and thus far unexplored—territory. Nonetheless, there is a variety of related work, including studies using text-based chatbots, automated Web sites, or mixed media, that provides encouraging and suggestive evidence that this is a novel direction that is well worth exploring. For example, in college-aged adults there is evidence that the process of capitalization engaged upon during social media use can yield psychological benefits just as human-human positivity can [50]. Computer-mediated social interactions have also been shown to be effective in promoting feelings of well-being and social connectedness in both working-age and older adults [9, 42]. There are likewise promising findings from research examining various smartphone text-based applications—although these have, for the most part, looked at more narrowly defined positive interventions such as encouraging gratitude, mindfulness, and the feeling of personal connectedness (see, e.g., [2, 8, 37, 46, 55] among others), rather than a broader range of positivity prompts or reminders interleaved into a more wide-ranging conversation about everyday activities. Many studies using smartphone or mobile applications also have focused specifically on narrow target groups linked by circumstances such as isolation during the COVID-19 pandemic [61], or by medical conditions such as cancer [40] or depression and anxiety [33]. More broadly, there also have been notable attempts to adopt a user-inclusive design approach with other age groups, such as children [101] and adolescents [31], and there has been sustained and growing interest in the varied ways that digital positive psychological interventions, such as automated Web sites and mobile applications, can promote human health and well-being [5, 75, 90].

Many older adults, often defined as adults 60 years of age or over [7, 94], wish to maintain independence as long as possible. This is often done by remaining in their own homes and dwelling in the community as much as they can, so-termed "aging in place." Promoting and supporting the independence of older adults is critical primarily because evidence continues to emerge that older adults remaining in the home may stay healthier longer and have better outcomes than those in assisted living (see, e.g., [14, 47, 64]), and because older adults aging in place therefore usually require less intensive support resources [65]. It is very important to note that intelligent systems cannot and should not fully replace human interaction and care; however, smart technologies are poised as potentially highly useful tools in augmenting and supplementing the effectiveness of human caregivers. Conversational technologies can be used for many things, including applications such as smart home control, encouraging medication compliance, and social interaction. Other smart technology, such as in-home sensor networks, can assist in maintaining physical health in uses such as fall detection. By their nature, however, most of these applications are designed for long-term use of technology, which presents unique challenges when considering design and adoption of technologies for older users.

The study presented here is a three-phase multi-method investigation into the effects of deliberate and encouraged positivity on user mood, interaction style, and acceptability of a prototype conversational system. *Phase I* consisted of feasibility pilot studies to establish our basic experimental approach. *Phase II* laid the groundwork for the development of an actual prototype conversational system designed from the ground up with input from older users by testing aspects of conversational design using a Wizard of Oz approach. The bulk of our work consists of *Phase III*, in which we created and tested an actual prototype conversational system implementing a conversational strategy centered around deliberate positivity. All three phases together address the following core questions about the use of positivity in conversational design for voice-based systems designed to interact with older adults.

RQ1: Can positive prompts or deliberate positivity be implemented to encourage continued interaction/use of a CA?

RQ2: How does user interaction with a CA implementing deliberate positivity influence user affect and perception of the CA across multiple sessions?

(a) Is user affect influenced by the number of positive prompts per session?

(b) Is user affect influenced by session-specific system usability, as assessed by user self-report using both general and CA-specific usability inventories?

(c) Is user affect influenced by session-specific system usability, as assessed by user behavior in response to minor system errors?

(d) Is user affect influenced by a user's general **interest level and experience (I&E)** relating to computers and information technology?

One of the core questions of interest in the design of CAs, as noted above, continues to be that of encouraging long-term user adoption and acceptability for deployment in a variety of scenarios [30]. This is a particularly salient question for our target user group of older adults. We therefore use a mixed-methods approach to characterizing participant interactions with the CA in the hopes of better understanding the challenges and opportunities for likely continued interaction with the CA. While this type of characterization is not explicitly a RQ, the development of appropriate metrics to profile interactions could be considered a "bigger picture" underpinning to this study. This mix of qualitative measures such as user self-report and quantitative measures such as psycholinguistic analysis allows us to build a holistic picture of older adults' interaction with a CA. Our unified approach also used a mix of participant feedback, such as user rating of minor system behavior, and expert *post-hoc* analysis by the research team to develop a much more nuanced characterization of these interactions in order to situate our results as potential guidelines for future system design.

In this article we first provide an overview of relevant work, including a review of deliberate positivity, or personal capitalization. We also provide background in conversational system design for older users as a framing for these experiments. We then outline our experimental design for each phase, including our methods, tools, and findings. We provide further analyses integrating results from different investigative methods, such as user self-report and conversational analysis, to provide a more holistic interpretation of our results. Finally, we discuss the implications of our results for the design of and incorporation of deliberate positivity in conversational systems for older adults and others, as well as the process of user assessment and testing involving target user groups such as older adults.

2 Related Work

2.1 Deliberate Positivity

The use of positive reflection in interpersonal interactions is well-studied. This process, also called personal capitalization, is a core relationship building and maintenance tool [79]. Capitalization is primarily asynchronous and retrospective. It is by definition dyadic, since experiencing the relationship-building benefits of capitalization relies on having a partner to respond to positive disclosure [52]. These benefits are apparent even in the case of disclosing minor positive events such as recounting a favorite TV show or relating something interesting encountered on a hike [87].

The benefits of sharing positive events or memories through capitalization extend beyond the sociorelational, however. The capitalization process can have a sizeable psychological impact on so-called capitalizers, even over the long term [4]. This is especially noteworthy considering evidence that people experiencing depression, isolation, or anxiety can actually gain stronger benefits from capitalizing [43, 56]. As such, capitalization has also been recognized as a promising clinical tool as

a part of positive psychological intervention for many different populations. This includes older adults, where the impact may be particularly far-reaching (see, e.g., [86, 92, 100], among others).

Other social processes, such as self-disclosure, have been shown to induce similar psychological benefits in the users of conversational systems as they do in human-human interactions [44], meaning that capitalization may be worth investigating as an aspect of system design. There is further evidence in the capitalization literature pointing to important elements in interaction design. Even the simple act of a potential listener taking the conversational initiative and asking basic probing questions, for example, may have a good deal of benefit as it promotes feelings of care between conversational partners [79]. Even straightforward agent-initiated dialogue such as "How was your day?" may be a powerful tool in prompting and encouraging continued positive reflection [19]. Encouraging further opportunities for positive experiences may also help increase the psychological benefits of capitalization [4, 16]. Therefore, a system designed with features such as calendar and scheduling functions that can help promote further participation in enjoyable activities may also be beneficial for combating isolation and stress in users.

Supporting mental health and further promoting mental wellness strategies is especially important for older adults, particularly adults who wish to maintain high degrees of independence for as long as possible [13, 74]. Maintaining mental resilience is an important piece of healthy and successful cognitive aging [48], and psychological intervention can help promote mental resilience. Positive interventions designed to help people manage stress and mitigate depressive or anxious symptoms are especially useful for this purpose [100]. The impact of this sort of positivity also extends beyond cognition. The effects of stress, depression, anxiety, loneliness, and related mental states on physical health are well-documented and numerous (see, e.g., [13, 17, 98], among others). A great number of disease processes benefit from the active management of mental health. This is a notable consideration for an older population, many of whom may already be managing long-term or higher-impact health concerns. Better physical health, in combination with better mental health, can make it easier for older adults to maintain independence and stay in the home longer [21, 88].

2.2 Older Users and Long-Term Adoption of Intelligent Technology

Designing for older users and designing for encouraging long-term use are difficult prospects, both individually and when taken together. Social processes such as positive capitalization are iterative and deepening, so any psychological and/or relationship building benefits may build effectively over the long term. However, maintaining long-term adoption and acceptance of conversational technologies is one of the most open and challenging questions in the field [30]. Simultaneously, older users are considered one of the hardest user groups to design for due to the heterogeneity of the user group [29, 39]. Common effects of aging, such as changes in memory, mobility, and sensation, are well understood; however, life experience, health, and even inter-group age differences can greatly impact the degree to which individuals experience any one of these changes [71]. Nonetheless, some common usage patterns and acceptability concerns for older adults in general do exist, which may provide additional insight into important design elements to help encourage long-term adoption by this user group, particularly when combined with broader potential causes of system abandonment.

Previous work has demonstrated that two primary uses of conversational systems for older adults are information seeking and listening to music, both in aging-in-place and assisted living contexts. [76, 82, 96]. Some evidence suggests that older users may also have different concerns about data security and privacy [32]. These users may engage in interactions that lead to more sensitive information being disclosed—for example, Pradhan et al. found that many information-seeking interactions between older adults and the Alexa involved questions about health [82]. Health concerns and more intense medical needs increase with age, so even social conversations

may include the disclosure of sensitive health information [6, 15]. There is also the potential that conversations and disclosed information may be overheard, particularly as a user's living circumstances may change; a user may need to have caregivers visit for health concerns, for example, or transition into assisted living, which may further limit a potential user's desire to use a device [76, 95]. The combination of the potential sensitivity of conversations and the presence of other listeners creates an increased need for fine-grained security and privacy controls, as well as a need to retain ownership over data as much as possible.

Older adults using conversational systems have also been shown to be open to social interactions with these systems, and experience feelings of increased engagement and support as a result of using them [22, 84]. Older adults also tend to be more likely to perceive systems as social presences and interact with them in a social style [22, 36, 97]. However, long-term use of CAs by older adults for social or other purposes can be limited by disappointment in exploration of what the conversational system can actually do [22, 32, 96]. In fact, the capabilities of intelligent systems (or lack thereof) have been consistently observed to be the greatest barrier to adoption of conversational technologies [20, 30, 62]. User expectations often exceed the limits of what a system can actually do—an expectation/experience gap that frequently causes frustration and eventually abandonment of the technology [30, 62]. The use of intelligent technologies has previously been modeled as a process of exploration, exploitation, and then abandonment [18, 68]. Many technologies may see high use and interest in the short-term and when users first begin using them. However, after this initial phase when the novelty has worn off, the limitations of technologies become more apparent and often more frustrating for users, at which point the technology is abandoned.

This pattern of not meeting user expectations has multiple potential root causes; however, one important contributing factor to the impression of systems overpromising and underdelivering is the interaction design for intelligent agents, particularly those that are conversation-based. Past evidence suggests that more "human-like" behaviors can lead users to believe CAs are more capable than they actually are, which in turn leads to disappointment, frustration, and potential abandonment of the technology [18, 62, 102]. However, the fact that more human-like behavior might be detrimental is complicated by the innate human tendency towards anthropomorphization, particularly when it comes to interactive and voice-based technology [51, 73]. It has long been suggested that this innate tendency makes it impossible to fully remove social or anthropomorphic cues from interactive technology [63, 73]. This is particularly true for CAs, since language-based and particularly voice-based interaction are inherently social/anthropomorphic cues, as is interactivity over the longer term [24, 63]. Fully removing these cues is therefore impossible—but cues can be reduced through conversational design to perhaps lessen the tendency of excessive social attribution for conversational technology [24].

Balancing the amount of social cues towards a more neutral social presence—that is to say, attempting to minimize the impression on users that a CA has some sort of persona or personality—may have other benefits in terms of encouraging long-term adoption. This is particularly the case for voice-activated systems, an interaction modality often noted as widely accessible and more inviting to users who may have sensorimotor issues or decreased hearing or vision [83]. These sorts of physical characteristics are often observed in older adults, so systems should be designed with an eye to physical inclusivity [29]. However, there are other dimensions to consider as well. Past evidence suggests that while older adults are receptive to adaptable systems, the desire for systems with social capabilities is quite mixed [26]. Conversely, older adults do tend to interact with intelligent systems on a more social basis [36, 97]. Therefore, the most inclusive design might also attempt to balance social cues and capabilities in an effort to not encourage false impressions and expectations from users. Using common social niceties such as thanking and greeting, while avoiding more fully anthropomorphic cues such as joking and self-disclosure, may strike a good balance

between a fully social presence and a robotic presence, thus encouraging longer-term adoption for a wider number of users—which is important for heterogeneous user groups like older adults.

3 Experimental Overview

3.1 Experiments

In this series of experiments we invited older users to interact with, and provide feedback concerning, a newly developed prototype conversational assistant, with the explicit aim of systematically examining the potential of deliberate positivity as a route to encouraging long-term adoption by this user group. We report here the results of a three-phase mixed-methods study including pilot studies (Phase I), a Wizard of Oz-based deliberate positivity experiment (Phase II), and an experiment using an actual prototype CA implementing a conversational strategy centered around deliberate positivity (Phase III). Results are presented in the same order that the studies were performed. Phase I reporting includes two pilot studies, Pilot 1a and 1b. Phase II reporting includes the Wizard of Oz-based study. Phase III reporting includes testing with older users and the prototype CA.

3.2 Participant Recruitment and Eligibility

Recruitment for all phases of the study, including Pilots 1a and 1b as well as Phases II and III, was performed via ResearchMatch. ResearchMatch is a national health volunteer registry that was created by several academic institutions and supported by the U.S. National Institutes of Health as part of the Clinical Translational Science Award program. ResearchMatch has a large population of volunteers who have consented to be contacted by researchers about health studies for which they may be eligible. The study description invited participants to take part in research to help improve a new voice-activated CA, and stated that the goal of the study was to create an accessible home-based CA to help with everyday tasks. Inclusion criteria were that participants were older adults, commonly defined by the World Health Organization as 60 years old or older [7, 94] (although, in fact, the majority of recruited participants were older—e.g., mean ages of 69 and 68 years in Phases II and III, respectively), were native speakers of English (defined as having learned English by 6 years of age), and had recently learned to use Zoom or wished to learn how to use Zoom. To facilitate scheduling, only participants who resided in the Central Standard Time zone were contacted. Review and approval for this study and all procedures was obtained from the University of Minnesota Institutional Review Board.

The research team's interactions with participants and the CA were also consistent across phases. After confirming the audio quality of the session and that it was being properly recorded, research assistants directed participants to begin the interaction. Although the research assistants had to remain on video calls by necessity, once audio quality was assured research assistants stepped away from the call after informing participants that they would stop actively listening to audio or watching the conference video. Thus, participants could interact with the system without being directly monitored by an external observer, more closely paralleling fully *in situ* conditions.

4 Phase I: Pilot Studies

4.1 Pilot 1a

In our first pilot study, we explored the appropriateness and usability of the experimental paradigm itself. We first wanted to determine whether or not older adults are comfortable with interacting with a unimodal voice-activated system. While we had originally planned to perform this study with participants in person, the COVID-19 pandemic necessitated the use of video conferencing. Therefore, Pilot 1a was a feasibility study specifically for older adults interacting with a voice-based system over Zoom. Both the original experimental design and the Zoom modification were IRB

approved, and for Pilot 1a and all subsequent experiments participants were compensated with a gift card to a retailer of their choice for an approximate pay rate of $15 an hour.

In Pilot 1a, we employed a protocol of everyday calendar-related scenarios in which older participants were asked to imagine themselves, and how they might use a CA for assisting them with the setting of reminders or planning for specific activities (e.g., an upcoming doctor's appointment or planning to host a small dinner party). Based on this Zoom-based single-session scenarios study we learned that older adults (N = 12, 60 years of age and above) are capable of, and comfortable with, taking part in CA-based interactions remotely on Zoom. We also learned that the majority of participants fully accepted that the CA is a real CA. As a precursor to development of a prototype system, the "Conversational Assistant" in Pilot 1a was deployed using a "Wizard of Oz" testing procedure in order to simulate a plausible system to evaluate different aspects of system design. This consisted of a soundboard with a set of pre-recorded prompts and positive, negative, and neutral responses that were created using the voice of "Joanna" from Amazon Web Service's "Polly" service. These results formed the basis of the actual prototype developed in later studies.

4.2 Pilot 1b

Given that Pilot 1a was promising, we wanted to gather further information with deeper probes into participants' impressions and what sorts of functionalities would be most helpful and relevant to older adults and the types of daily activities they might have. To do this, we conducted one-on-one Zoom-based follow-up personal interviews with 10 of the older adults from the first pilot study. The interviews were developed and conducted with IRB approval, and were designed to inform us of what types of events older adults might include in their daily calendars (including under the restrictions in activities related to the pandemic), so that some of the more common event types could be incorporated into the CA scripts before testing with further older adults. The interviews also provided us with knowledge of the general activities/events older participants found to be positive and/or rewarding. From this series of interviews, a clear pattern began to emerge regarding common enjoyable activities for these older adults. The majority of activities and events mentioned by interviewees were related to either leisure and hobby activities or family/social activities. These two general topics became our domains of interest in further experiments and eventually prototype development.

5 Phase II: Wizard of Oz

Prior to beginning the intensive process of prototype development, we wanted to further explore the concept of deliberate positivity used as a conversational strategy. To do this, we employed a Wizard of Oz protocol for a less constrained, more social/conversational CA than the scenario-based system from the pilot experiments. The Wizard of Oz system conversed with participants regarding the target domains of interest—family and hobbies—and further prompted users about positive experiences and memories to promote positive reflection in participants. The results of this experiment were extremely promising and guided the design of the actual prototype used in Phase III.

5.1 Experimental Design

Participants in the Wizard of Oz protocol attended two sessions a day over the course of 4 days for a total of eight sessions per participant. Figure 1 illustrates the general procedural flow for each individual session.

The first session introduced participants to the experiment and gave an overview of how the rest of their sessions would go. Immediately prior to their first interaction with the "system," participants

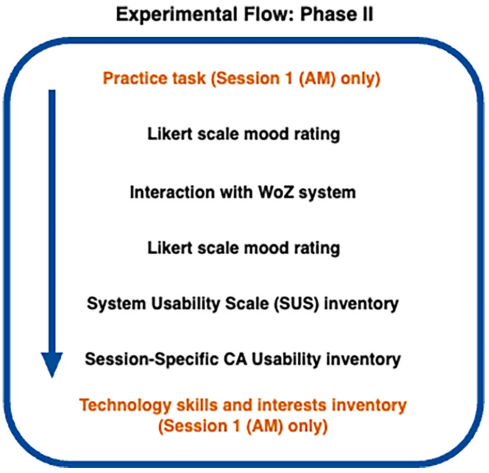

Fig. 1. Experimental flow for Phase II.

were also asked to complete a practice task to familiarize themselves with the system, which was one of the scenarios used in the first pilot study.

Most of the rest of the first session followed the same pattern as all subsequent sessions. For each session, participants were asked to report on their current mood using a 7-point Likert scale, anchored at 1 for "not very good at all" and 7 for "very good." Participants then interacted with the Wizard of Oz system, which included conversing about family- and hobby-related topics and responding to positive prompts from the system. Following each interaction, participants were asked to rate their mood again using the same 7-point Likert scale. Lastly, participants were given two brief surveys regarding the general usability of the system (the **System Usability Scale (SUS)**) and their perceptions of the CA, such as rating the pleasantness and appropriateness of the conversational interaction during that particular session (the Session-Specific CA Usability Inventory). At the end of the first session, participants were also asked to complete several brief inventories detailing their current I&E levels with technology.

Following the eighth and final interaction, the experimenter disclosed the use of deception as part of the Wizard of Oz protocol. Once they were told they had interacted with a human, participants were asked a brief series of questions regarding whether or not they had suspected at any point that they were interacting with a human under the guise of a machine rather than an artificial agent. If participants indicated the deception may have failed, they were further asked when and why that suspicion occurred. At the conclusion of the experiment participants were given compensation options to choose from. This was done with IRB approval, and all participants were provided with IRB approved consent documentation and contact details for the research team.

5.2 Methods

5.2.1 Technology Profiles. We included a number of questions in experimental surveys to better understand participants' experience with and impressions of technology. These included questions related to participant attitudes towards technology, a report of their general technological skills, and their level of experience with various uses of technology.

(a) Attitudes and Interests. Participants were asked to answer a small inventory of questions relating to their current interests and attitudes towards technology [49]. Each question was responded to on a 5-point Likert scale, anchored at 1 for "not at all" and 5 for "extremely." The questions included

a rating of their interest in technology, willingness to try new things, and frequency of technology use. Each question was asked separately for smartphones and then for computers/tablets.
(b) Technology Skills. Participants were similarly asked to rate their abilities with regards to both the physical use of technology and the use of technology for specific applications. This section of the survey included questions adopted from Boot et al. such as whether participants felt comfortable using a keyboard to type, removing a paper jam from a printer, or using technology to find information and answer questions [10]. These survey questions were also answered using a 5-point Likert scale with the same anchors as above.
(c) Experience with Information Technology. We asked participants (in Session 1) to answer 10 questions about their experience with Information Technology [81]. They were asked to rate their level of experience with different ways of "using the internet and technology." This included common online activities such as online shopping, e-mail, social networking, and streaming. Responses were on a 5-point Likert scale, anchored at 1 for "very poor" and 5 for "excellent." The full text of questions related to experience with technology can be found in Appendix A.

5.2.2 SUS. There are 10 items on the SUS written as statements that raters are asked to evaluate. Each statement is rated on a 5-point scale, anchored with 1 for "strongly disagree" and 5 for "strongly agree." These statements include ratings of system complexity and ease of use, as well as learnability and the perceived ease of adopting the system. The 10 items on the SUS are weighted, combined, and scaled to give a final usability score between 0 and 100. In previous aggregate studies, the average SUS score from about 500 studies is a 68, where anything above 68 can be considered as having above average or good usability scores [89]. The full text of the 10 SUS items can be seen in Appendix B.

5.2.3 Session-Specific Ratings of CA Usability. In order to further probe user impressions of the CA as a language-based technology, we adopted a 15-item inventory of usability statements. Similar to the SUS, these were rated on a 5-point Likert scale anchored at 1 for "completely disagree" and 5 for "completely agree." These statements covered user impressions of design elements more specific to natural language/conversational interfaces, and included items such as user ratings of system sociability, appropriateness, interaction style, and language use. These were designed to provide insight into larger interaction elements such as whether participants felt comfortable being open and sharing with the system, or whether the interaction was pleasant and enjoyable. The full text of these 15 items as presented to participants can be found in Appendix C.

5.3 Results

5.3.1 Demographics and Technological Profiles. A total of 17 participants were recruited for Phase II; of these, 14 completed all 8 experimental sessions. The mean age of participants was a little over 69 years old, with 13 out 18 participants 65 years of age and older. Participants reported a mean of approximately 17.5 years of formal education. All participants self-identified as native English speakers, and 16 participants identified as non-Hispanic White; 1 participant identified as mixed race. Of these, 13 participants identified as female, while 4 identified as male. By self-report, participants were also in very good health, indicated by a mean rating of 5.8 on a 7-point scale.

By and large, the participants for Phase II were interested in and positive towards technology and especially exploring new technology. Almost all participants reported they were daily users of computers and smartphones, and most indicated some comfort with the technologies and willingness to explore new features. Participants were most proficient at tasks like sending e-mail and information seeking, while they were least familiar with using computer and smartphone technologies for streaming purposes including movies and music. About half of the participants indicated they had previous experience with smart home assistants.

Table 1. Session-Specific Ratings of CA Usability for Phase II

Session-Specific CA Usability Question	S1	S2	S3	S4	S5	S6	S7	S8	Mean (SD)
Q1: CA pleasant	4.59	4.53	4.53	4.35	4.59	4.53	4.50	4.64	4.70 (0.47)
Q2: CA sociable	4.69	4.76	4.71	4.47	4.65	4.71	4.57	4.57	4.71 (0.45)
Q3: Felt comfortable sharing	4.41	4.29	4.24	4.00	4.35	4.00	4.36	4.14	4.44 (0.86)
Q4: Felt could be open	4.24	4.12	4.29	4.12	4.29	4.06	4.36	4.50	4.45 (0.72)
Q5: Felt involved	4.47	4.29	4.47	4.12	4.35	4.29	4.21	4.36	4.46 (0.75)
Q6: Enjoyed interaction	4.06	4.12	4.35	4.06	4.06	3.94	4.21	4.14	4.24 (1.00)
Q7: Interaction was smooth	4.35	4.00	4.24	4.00	4.18	4.24	4.14	4.07	4.22 (0.71)
Q8: Would like to interact again	4.18	3.88	4.00	3.88	3.82	3.88	3.93	3.86	4.02 (1.10)
Q9: Interaction was satisfying	4.12	3.88	3.94	3.63	3.94	3.94	3.86	4.00	4.07 (0.94)
Q10: CA said right thing	3.53	3.71	3.76	3.81	3.88	4.06	4.14	4.14	4.10 (0.88)
Q11: CA responded appropriately	4.29	4.29	3.94	3.94	4.24	4.24	4.36	4.21	4.40 (0.79)
Q12: CA communicated correctly	4.59	4.56	4.71	4.18	4.41	4.47	4.36	4.21	4.53 (0.63)
Q13: CA seemed competent	4.59	4.35	4.47	4.24	4.47	4.41	4.57	4.43	4.61 (0.56)
Q14: CA seemed natural	3.88	3.88	3.94	3.82	4.00	4.00	4.14	4.14	4.09 (0.86)
Q15: Would like a long conversation	3.24	3.35	3.35	3.12	3.29	3.29	3.50	3.36	3.46 (1.18)

N = 17, except N = 14 for Sessions 7 and 8. All items were answered on a 5-point Likert scale (1 = completely disagree, 5 = completely agree). The final column provides the average for participants who completed all eight sessions. The full text of the 15 questions can be viewed in Appendix C.

5.3.2 Mood and Affect. In each session, directly before beginning the CA interaction ("begin"), and directly after the CA interaction was completed ("end"), participants were asked to report how they were feeling on a 7-point scale. In general, participants responded that they were feeling quite good, grand mean = 6.37, 95% CI [6.10, 6.65].

Averaging across all of the sessions that participants completed, there was a modest but significant *increase* in their mood ratings from beginning to ending the interactions with the CA, $F(1, 15) = 6.42$, $p = 0.023$, partial $\eta^2 = 0.30$. Mood ratings at the beginning of the CA session were $M = 6.30$, 95% CI [6.01, 6.58]; mood ratings at the end of the CA session were $M = 6.45$, 95% CI [6.17, 7.73].

5.3.3 Session-Specific Ratings of CA Usability. Across the first six sessions, the average responses to the session-specific CA-interaction questions generally fell between 3.50 and 4.75, with only one of the questions (Q15—"I would like to have long conversations with the CA") uniformly showing an average response that was closer to the midpoint of the scale. For participants who completed all eight sessions, the mean across sessions was uniformly at or above 4.00, except for Q15. Table 1 provides session-specific values for each of the 15 questions.

5.3.4 Interim Conclusions. These data provide initial evidential support for the general approach to both the selected content domains that the prototype CA would focus on in dialogue (that is, family and hobby-related topics) and the use of deliberative positivity to encourage continued interaction with the CA. With respect to our two main RQs, these findings offer suggestive support for the potential value and feasibility of implementing positive prompts or deliberate positivity to encourage continued interaction/use of a CA (*RQ1*) and also for the possible favorable effects of deliberate positivity on user affect and perception of the CA across multiple sessions (*RQ2*). This constituted an encouraging foundation for our next phase deploying our prototype CA. These data also provided guidance for us in future experiments; there was some thought that the relatively low endorsement of engaging in longer conversations with the CA may have been related to the

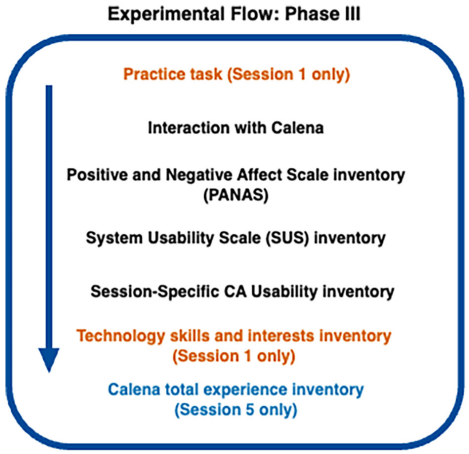

Fig. 2. Experimental flow for Phase III.

twice-daily sessions. Therefore, in Phase III participants were asked to attend one session a day over 5 days.

6 Phase III: "Calena" Prototype

6.1 Experimental Design

Participants were asked to attend one session a day over the course of a week, for a total of five sessions. This was a deliberate decrease and potential improvement over the design in Phase II, as fewer sessions and once-a-day scheduling reduced cognitive load on participants. Each session, however, had a similar workflow as in Phase II. Figure 2 demonstrates the procedural flow for each session in Phase III.

In the first session, participants were introduced to the study and gave consent to participate. Immediately prior to the first interaction with the prototype system, each participant took part in a very constrained interaction in order to both acclimate them to the system and ensure that their audio quality was sufficient to be understood by the speech-to-text service. This was an approximately 30–60 second conversation about the weather. At the completion of this task, participants then interacted with Calena in a more self-directed way.

All interactions, including in the first session, followed the same general conversational flow. Participants were asked to verbally verify both their participant ID and session number before Calena began asking about how their day had gone. After this first social nicety, participants were prompted as to whether they would like to talk about family or hobbies, the two domains of interest. The conversation was largely participant directed from there, as Calena would ask follow-up questions regarding any hobbies or family information disclosed by the participant. The follow-up questions included both prompts based on information disclosed in previous sessions and anything newly disclosed by the participant on a given day. Participants were occasionally prompted to disclose new information, and participant answers including hobby information and family structure were maintained across all sessions.

Following each session, participants were asked to complete a brief inventory related to their current mood that asked more specific questions about their positive and negative affects. This was a more comprehensive in-depth probe than was used in the prior study. Based on the promising results of the single-item affect assessment in Phase II, we improved on the design by selecting

a standardized inventory to capture a more holistic view of participants' positive and negative affect. This mood inventory was followed by a short questionnaire regarding the day's interaction, which included ratings of how well the interaction went and how usable the system was for the interaction in question. For the first session only, participants also completed self-report questionnaires about their current technology usage and experience. At the end of the fifth session and thus the conclusion of the experiment, participants were asked to complete an inventory regarding their overall experience with Calena, as well as their retrospective thoughts on the usability of the system and its potential uses in the real world. After completing these inventories participants were given the opportunity to ask any follow-up questions of the research team and selected their preference for receiving compensation for their participation in the study.

6.2 System Design

6.2.1 "Calena" Architecture. Phase III primarily focused on the development of a real prototype dialogue system that could then be deployed and tested with the target older adult user group. This prototype was given the working name of "Calena" due to its early origins in a calendar and scheduling system, and to maintain some consistency with existing (implicit) naming conventions among commercially available dialogue systems such as Alexa and Siri, as well as the AWS Polly "Joanna" voice used in previous experiments. Calena was designed from the ground up with five inter-related components meant to support key goals in a social conversational system:

(1) Robust deep learning-based support for fully customizable conversational domains.
(2) Ability to analyze sentiment in user utterances in real time in order to respond appropriately.
(3) Flexible and customizable **personal knowledge base (PKB)** to store and reason with information disclosed by users.
(4) Robust speech processing/generation with flexible timing presets to support real-time interpretation of user speech and fast generation of speech for system turns.
(5) Data protection and security, both server-side and locally, to maintain confidentiality and participant privacy due to the potentially sensitive nature of information participants might share.

6.2.2 Conversational Support—MindMeld. The MindMeld API [85] is a robust open source CA API released by Cisco in the late 2010s. The API is designed to support the creation of CAs with deep domain knowledge by leveraging gold standard large language models such as BERT. MindMeld was chosen because it is also specifically designed to enable the creation of in-depth fully customizable domain models, including user-defined entities and training data. MindMeld also allowed us to leverage both previous experimental data [28] and data generated internally by members of the research team. This enabled us to quickly and thoroughly implement additional functionality for Calena such as integrated calendar and scheduling capabilities. This in turn allowed us to make Calena more well-rounded and further reinforce any perception of the system as reasonably socially capable and flexible.

6.2.3 Sentiment Analysis—VADER. We considered sentiment analysis to be one of the most important aspects of a properly social chatbot since it would ensure we were able to respond to user utterances in an appropriate and not off-putting way. Our final tool selection for sentiment analysis was NLTK's VADER [45]. Other state-of-the-art sentiment analysis tools we tested such as FLAIR detected sentiment as a binary value—that is, the detected sentiment was either entirely positive or entirely negative [1]. VADER, on the other hand, scored sentiment on a continuous scale of intensity in addition to detecting positive or negative sentiment. This allowed us to modulate the responses Calena used to be more appropriate for varying levels of sentiment intensity. Perhaps

most importantly, VADER's scoring system also allowed us to detect neutral sentiment and respond appropriately in order to make Calena seem more natural in conversation. VADER was also a strong tool for our purposes because of its compact implementation and speed in scoring utterances, and because the use of standardized speech-to-text transcription gave us a near-guarantee of clean, correct input with a very small chance of issues such as misspellings that may impact VADER's accuracy.

6.2.4 Personalization and History. Personalization is a powerful tool to encourage users to continue engaging with CAs [80]. Previous evidence also suggests that older adults are particularly interested in personalized/adaptive systems [26]. In our case, personalization primarily involved remembering any information users disclosed about family members or hobbies in order to avoid repetitive conversation. Due to its nature as a multi-phase study, we also wished the adopted design to support future development and extensions to the project, which required flexibility. Apache Gremlin [3], a graph-based framework, was chosen for several reasons. First, it allows any number of custom designed entity types—family members and hobbies, in our case—with any desired attributes such as a name or category. Unlike rigidly structured specifications such as JSON or XML, Gremlin also supports data entries that do not have filled values for each of their attributes, which helps make user interaction and knowledge extraction more natural. Gremlin data is stored in a connected graph structure with different edges with their own attributes. This allowed us to easily capture *relationships* between entities in addition to relevant information disclosed by users—for example, an edge in the graph between two people can be labeled with the exact (literal) familial relationship. This also allowed us more flexibility and customization in traversing and reasoning with the PKB for individual users. Additionally, Gremlin is a temporary database that persists only as long as the server is active, which allowed us to build in extra data security and privacy protection; however, Gremlin data can also be backed up and saved to disk to protect against potential technical issues through the course of an individual participant's time interacting with Calena.

6.2.5 Speech Processing and Generation. We chose to use an out-of-the-box solution to handle the text-to-speech and speech-to-text requirements for the Calena dialogue system. After some comparison between commercially available products, Microsoft Azure cognitive services proved to be the best tool for the job. Azure allowed us to define customized behavior such as listening timeouts when interacting with users, which is particularly important for older users who may have longer pauses and other speech disfluencies—less interruption by the CA would hopefully lead to less user frustration. Azure was also the best solution in terms of maintaining full control of our data, and the robustness of the service also increased our processing power and allowed us to do live input sanitization without additional computing resources, which was very important in using the full power of the customized domain and entity structures we had defined in MindMeld.

6.2.6 User Privacy and Data Protection. Previous work has indicated that older users consider data protection and privacy to be particularly important [26]. Other work has also provided evidence that users will disclose a variety of personal information irrespective of system design—including with a task oriented system [27]. Older users in particular may share more sensitive data due to the prevalence of health concerns in the age group [6, 15], particularly when those health concerns are combined with older adults' tendencies to use conversational systems for information seeking [82]. All of these factors lead us to place privacy and data security at the top of the feature list for our prototype design. This turned out to guide our choices over other components of the prototype system, namely our choice of PKB representation (Gremlin) and our choice of speech processing and generation methods. We discovered that Microsoft Azure would let us to maintain ownership

of our data by keeping any cloud storage protected and private [69]. Azure also guarantees we retain full ownership of our own data, again setting it apart from comparable services [70]. The data protection and ownership features of Azure led us to choose it as the platform through which speech-to-text and text-to-speech functions were performed. Lastly Gremlin, as previously discussed, is a temporary server that only stores data until the Gremlin server is shut down or restarted [3]. Unless data is specifically backed up to disk through a separate process, Gremlin data is therefore entirely transient, which gave us an extra layer of user security and privacy protection. In addition to storing any user knowledge in the Gremlin knowledge base, the Gremlin server itself was hosted on a password-protected private server whose access was restricted to PIs and two graduate research assistants trained to run and troubleshoot Calena during experimental sessions. For a final additional layer of protection for participant data, any backups of the Gremlin database and conversational data were stored on Box, which is a HIPAA-compliant cloud storage platform. All participant data was treated as potentially containing sensitive information commensurate with **personal health information (PHI)**; the research team followed institutional guidelines for handling PHI in controlling the storage and access of participant data [77].

6.3 Conversational Design and Implementation

6.3.1 Data and Training. Much of the training data for the system was generated among project team members. In cases where large amounts of training data were needed, the workload was distributed such that each member would be asked to generate up to 50 different examples. In order to boost diversity in the dataset, data generation was not collaborative and each set of samples was created independently and blind to the samples being created by other members of the research team. Previous custom entity datasets, such as user names, were also adopted from other projects.

6.3.2 Domain Reasoning. In order to balance a reasonable amount of fluency in conversation vs. computational complexity and processing time, we attempted to streamline conversations where possible so that some parts of interactions could be rule based. Members of the research team prepared a list of over 100 hobby names and then further categorized them into hobby types. These conceptual categories were based off of common elements between different hobbies—e.g., outdoor activities, cooking, volunteer work, and so on. Independent team members then created sets of slightly generic questions applicable to each category that could be used when a user introduced a new hobby to the system. For example, a generic question might be "What do you like most about being outside?" for activities like hunting or hiking, or "What do you like about your kitchen?" for detected hobbies involving food. Categorical questions were then run against individual hobbies to ensure they were general enough to use for every activity in that category.

Hobbies also had to be handled via entity detection, and the relevant information including the proper set of categorical questions had to be retrieved from the database through some associative metric. We tested a number of methods to promote this. Our original method was to examine the semantic similarity between the category names and the activities themselves, but this had mixed results and often the semantic scoring was different than what we would expect. Next, we looked at the semantic similarity scores between the detected hobby entities and the different hobbies in the database. This proved to be the most usable method, though even after multiple attempts at training MindMeld still had some issues with the negation of hobbies (e.g., "No, I don't do X."). To circumvent this we had an additional semantic similarity check with a very high threshold so the CA ignored the detected hobby if the threshold was met.

Family detection also required some fine-tuning due to the need to keep the natural language understanding process efficient, compact, and reliable. To this end, we relied on built-in features and models to the extent possible. Family detection primarily involved entity recognition, role

classification, and entity resolution through MindMeld. A custom "person" entity with associated roles ("family," "name," etc.) was defined. A **maximum-entropy Markov model (MEMM)**-based [67] classifier was then trained using data generated as outlined in Section 6.3.1, with the addition of default personal name data from MindMeld. A logistic regression role classifier then determined whether the mentioned person belonged to the "family" role. Entity resolution was then done using a search of the Gremlin database, where people were stored as nodes in the graph with edges representing relationships. Despite the large amount of name training data used, however, the MEMM classifier was not as strong at detecting people by name as it was by relationship. To offset this, an additional pass at entity recognition was done using spaCy's RoBERTa-based [59] classifier and "en_core_web_sm" model, which had much higher performance at tagging people by name. The number of recognized entities between the two was compared, and the classifier outputs with the highest number of entities detected were used.

6.3.3 Positive Prompts. Our experimental manipulation hinged on the CA delivering positive prompts in a minimally repetitive and somewhat natural way. To facilitate this, the positive prompts kept after feedback in Phase I were further categorized into groups based on whether they were more general or participant emotion driven (e.g., "What is something you were grateful for in the last weeks or months?") or driven by specific events or activities (e.g., "What is a meal or snack you recently enjoyed?"). These two categories were then mixed so that when prompts were selected, participants received a mix of specific and more general prompts. Each prompt also had three associated responses for the CA to choose from based on the results of sentiment analysis for the participant's answer to the prompt. The response of the CA would therefore vary both by question and by whether or not the user response was initially positive, negative, or sentimentally neutral.

6.3.4 Social Impressions and Personification. As part of our investigation into conversational strategies with an eye towards encouraging long-term adoption, we considered the social capabilities and "humanness" of the prototype in the conversational design. The language- and voice-based interactive nature of the technology makes it impossible to fully remove social cues. However, we purposefully excluded highly anthropomorphizing intentional social cues to reduce the likelihood of initially inflated or unrealistic user expectations arising from overly personified CA behaviors. Using the taxonomy by Feine et al. [24], we paid particularly close attention to content cues and deliberately avoided including jokes, praise, tips and advice, and self-disclosure. We also limited small talk to "how are you today" as a conversational opener, and further limited the CA's use of references to self to be as infrequent as possible and contain no other personification other than the use of first-person pronouns. We did include a small number of cues such as thanking, greeting, and obtaining user assent to participate in dialogues, primarily to give an impression of politeness while keeping Calena as neutral as possible.

6.4 Methods

6.4.1 Deliberate Positivity. The primary manipulation in this study involved encouraging positive reflection. As such, participants were divided into two groups (Group A and Group B), each with a different schedule for the number of positive prompts per session. Each group received the same number of positive prompts overall; however, the number of positive prompts was *increased* at different times for each group. Participants in Group A had their highest number of positive prompts on days 2 and 4, while Group B had a delayed increase in the number of positive prompts and had their highest number on days 3 and 4. Both groups had the same number of positive prompts on days 1 and 5, and had the same number of increased positive prompts on day 4.

6.4.2 User Experiences with Technology. Participants were asked to provide some information about their previous experiences with and interests in technology so that we could develop a technology-specific "demographic profile" for participants. The inventories used for this purpose were adopted from Phase II and include the I&E inventories described in Section 5.2.1.

6.4.3 PANAS. The PANAS consists of 20 words describing different moods and emotions. Participants are asked to read each emotion and then rate how much they feel that way at the time of rating. Ratings are done on a 5-point Likert scale anchored at 1 for "very slightly or not at all" and 5 for "extremely." Positive and negative emotions were interleaved in the survey presented to participants. The emotions on the PANAS include: interested, distressed, excited, upset, strong, guilty, scared, hostile, enthusiastic, proud, irritable, alert, ashamed, inspired, nervous, determined, attentive, jittery, active, and afraid.

6.4.4 SUS. We retained the SUS used in Phase II as a method of assessing user satisfaction with the prototype. Section 5.2.2 contains a more detailed description of the inventory, and the full text of each inventory item can be found in Appendix B.

6.4.5 Session-Specific Ratings of CA Usability. For each experimental session, after responding to the PANAS and the SUS, participants viewed and rated 15 statements about their interactions with Calena during that particular session. This inventory was retained from Phase II, and a more detailed description of the inventory can be found in Section 5.2.3. The full text of all inventory items can be seen in Appendix C.

6.4.6 Post-Study Rating of Total CA Experience. In Session 5 only—thus after several days of interacting with the CA—participants were asked to answer a number of questions about how they might use a CA, how they perceived the personality of the CA, and so on. Participants were asked to answer these questions using a 7-point Likert scale, anchored at 1 for "not at all," 4 for "somewhat," and 7 for "very much so."

The seven questions regarding overall impressions included: (1) Could the CA assist you with reminders for things? (2) Could the CA help you with your scheduling? (3) Could the CA help you with your list making? (4) Could the CA encourage you to complete tasks? (5) Could the CA motivate you in your goals? (6) Could the CA help you stay accountable to the tasks and/or goals you need to do? (7) Could the CA help others you live with (e.g., family, friends) stay accountable to their tasks and/or goals?

6.4.7 Conversational Analysis and Linguistic Inquiry Word Count (LIWC). LIWC is a psycholinguistic analysis tool that defines different semantic and syntactic categories for a very large dictionary of words. Given a piece of text, the tool creates raw counts of the frequency of words in each category, as well as more standardized measures regarding specific semantic and sociolinguistic properties [78]. There are four summary variables, which are standardized measures based on a 100-point scale. We focused on these summary variables to characterize the language participants used when interacting with the CA [93]:

(1) Analytical Thinking. Scores for this measure indicate the level of formality and reservedness in language, where a lower score means a more flowing, natural, or less reserved response. Participants with lower scores here may feel more comfortable with a CA and more willing to interact with it on a social level.
(2) Clout. Scores for clout measure authority and confidence in text, where relative authority or power increases with the score.
(3) Authenticity. The higher the score, the more open and disclosing someone is being.

Table 2. Session Duration Statistics (Phase III)

	Mean	Median	SD	Min.	Max.	N
Session 1	19.40	15.33	10.40	10.18	54.63	26
Session 2	17.12	16.54	7.16	6.92	34.52	24
Session 3	15.11	14.67	5.75	5.17	26.93	25
Session 4	13.25	12.77	5.23	5.47	26.80	25
Session 5	15.62	13.57	8.41	3.12	37.87	22
Average per participant[a]	17.27	15.76	8.82	8.40	54.63	26

Values are in minutes.
[a]N = 26, with the average calculated based on all of the sessions a given participant took part in. The number of sessions varied across participants, with participation as follows: All five sessions, N = 21; four sessions, N = 4; one session, N = 1; thus, across all participants, data are from a total of 122 sessions.

(4) Emotional Tone. Scores here measure the positive or negative tone in conversation, where higher scores are more positive.

6.5 Results

Results for Phase III are presented in correspondence with our identified RQs. First, we give an overview of participant demographics, including statistics about the total N per session and the duration of interactions with the Calena prototype. Second, we perform a transcript-based analysis of user behavior in response to positive prompts, which was the primary manipulation for all parts of *RQ2*. This includes statistics for participant questionnaires. First and foremost we report statistics for session-wise participant behavior in response to positive prompts and participant affect (*RQ2a*, Sections 6.5.2 and 6.5.3). We further examine results pertaining to the interaction between user affect and session-wise usability assessments (*RQ2b*, Sections 6.5.4 and 6.5.6), as well as user affect and minor system errors identified through a transcript-based behavioral analysis (*RQ2c*, Section 6.5.7). Lastly, we examine the interplay between user and affect and users' self-reported I&E with technology (*RQ2d*, Sections 6.5.8 and 6.5.9). We conclude the results with overall post-study impressions of the CA.

After the presentation of results for our primary RQs we also present exploratory analyses, including further transcript-based investigation of user behavior using LIWC to provide convergent and supportive evidence of user acceptability for the prototype CA.

6.5.1 Demographics. A total of 26 individuals who met our inclusion criteria (e.g., aged 60 years or older, and native speaker of English) took part in the Calena prototype study. The mean age of participants was 68.04 years (SD = 6.09), min = 60, max = 81, with 20 of the 26 participants being 65 years of age or older. Nineteen participants self-identified as female, while seven self-identified as male. Participants were also asked to self-report on general health; the mean score was 5.23 on a 7-point scale (SD = 1.28, min = 3, max = 7). All participants reported normal or corrected-to-normal vision, while 24 out of 26 participants reported as having normal or corrected-to-normal hearing. Among participants, 3 identified as Black or African American, 1 identified as American Indian/Alaska Native, 22 were of White ethnicity: 25 not Hispanic or Latino, 1 no response.

Overall, participants attended 122 sessions. Sessions ranged in length from around 5 minutes to 15 minutes, with a median duration of nearly 16 minutes. Table 2 contains statistics regarding the duration of each session, including the number of participants included in analysis for each session.

Table 3. Average Number of Positive Participant Responses to the Average Number of Positive Prompts by Session and Group

Group	Session 1		Session 2		Session 3		Session 4		Session 5	
	CA	Part.	CA	Part.	CA	Part.	CA	Part.	CA	Part.
Group A	1.10	1.10	**4.00**	**3.30**	1.00	0.80	**3.90**	**3.30**	1.00	0.86
Group B	1.00	0.88	1.00	0.86	**3.80**	**3.40**	**3.60**	**3.33**	1.00	0.93

Sessions in which the CA (Calena) was designed to give four positive prompts are shown in bold font; sessions in which the CA was designed to give one positive prompt are in regular font. In all sessions, "Part." (for Participant) indicates the average number of times that participants responded positively to the prompt given by the CA. Group A was assigned to receive an increased number of positive prompts from the CA (Calena) in Sessions 2 and 4; Group B was assigned to receive an increased number of positive prompts in Sessions 3 and 4. For Group A, N = 10, 10, 10, 10, and 7 for Sessions 1 through 5, respectively; for Group B, N = 16, 14, 15, 15, and 15 for Sessions 1 through 5, respectively.

6.5.2 User Behavior in Response to Positive Prompts. For all included participants, the transcribed and checked transcriptions were read and coded line by line for seven variables, including six variables that pertained to the CA's dialogue or actions and one that pertained to the participants' response to the CA. In terms of behavior directly related to the experimental manipulation, we noted any turn in which the CA selected a positive prompt. We also noted the overall sentiment of participant responses to each positive prompt in order to measure how well the prompts were received. Table 3 shows statistics regarding the number of positive prompts the CA gave per session and the number of positive responses to those prompts from participants.

As can be seen from Table 3, participants' responses to the positive prompts from the CA (Calena) were predominantly positive, and clearly increased in frequency when the CA provided more positive prompts (for Group A, in Sessions 2 and 4; for Group B, in Sessions 3 and 4). Contrasting the average number of positive participant responses for the two sessions with more positive prompts compared with the average number of positive participant responses for the three sessions where the CA gave only one positive prompt, revealed a highly significant difference both for Group A (means = 3.57 vs. 1.00 respectively, $F(1, 6) = 937.22$, $p < 0.001$), and Group B (means = 3.36 vs. 0.88 respectively, $F(1, 13) = 207.39$, $p < 0.001$). The comparatively infrequent occasions when the CA provided a positive prompt but the participant did not respond positively to that prompt primarily involved times when the participant did not respond at all (e.g., they did not hear the question or were interrupted), or the participant explicitly stated that they could not think of anything/did not have a response.

6.5.3 PANAS. The PANAS was the second inventory administered during each experimental session to develop a better day-to-day picture of participants' affective state. Figure 3 shows the trends in participant positive and negative affect across all five sessions.

A 2 (Valence Type: positive or negative) × 5 (Session: Sessions 1–5) × 2 (Group: A or B) mixed factor ANOVA on the PANAS scores showed a highly significant effect of Valence Type, with Positive Affect (M = 30.96) markedly higher than Negative Affect ($M = 10.62$), $F(1, 17) = 189.78$, $p < 0.001$, partial $\eta^2 = 0.92$. There was no effect of Session, $F < 1$, no Session × Valence Type interaction, $F = 1.24$, no Session × Group interaction, $F < 1$, and no Session × Group × Valence Type interaction, $F < 1$. There was however, a main effect of Group, $F(1, 17) = 5.62$, $p = 0.03$, and a Valence Type × Group interaction, $F(1, 17) = 5.49$, $p = 0.032$, reflecting generally higher positive affect in Group A than in Group B, and a greater difference between positive and negative affect for Group A than for Group B.

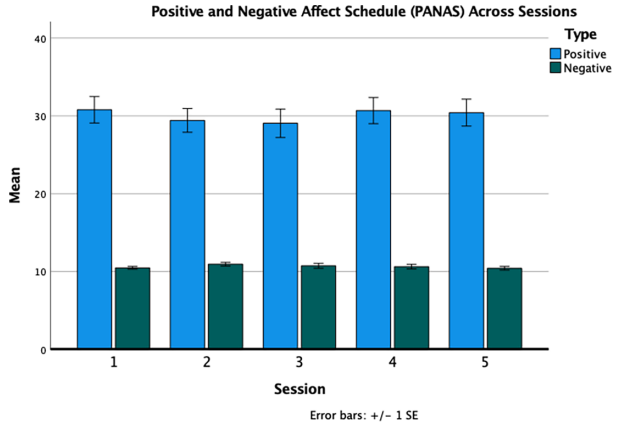

Fig. 3. Participant positive and negative affect scores across five sessions.

These results show that there was no differential effect of the frequency of positive prompts intervention. However, a more focused probe of the effects of the intervention on PANAS scores might compare the mean PANAS scores (especially for positive valence) for the average of the two sessions with more positive prompts vs. the average of the two sessions with fewer positive prompts, setting aside the first (baseline session) that was held constant for both groups.

A 2 (Group A vs. Group B) × 2 (Session Type: One vs. Four CA Positive Prompts) mixed-factor ANOVA on the averaged Positively Valenced PANAS responses revealed no effect of Session Type, $F < 1$, and no Group × Session Type interaction, $F < 1$, but a main effect of Group, $F(1, 17) = 5.22$, $p = 0.036$, partial $\eta^2 = 0.24$. This main effect of Group reflected the observation that Group A generally reported higher positive valence ($M = 34.25$) than did Group B ($M = 27.35$), but this generally higher positive valence for Group A was apparent both for sessions with one vs. four CA positive prompts ($M = 34.64$ and $M = 33.86$, respectively) whereas Group B had consistently slightly lower positive affect ($M = 27.54$ and $M = 27.17$, respectively).

Nonetheless, although the intervention did not *differentially* boost positive affect (*RQ2a*), across the five interaction sessions with Calena participants' positive affect remained high and negative affect remained low (*RQ1*). This across-group pattern of high and sustained positive affect across all of the sessions is shown in Figure 3.

6.5.4 SUS and PANAS. Participants completed a brief system usability inventory (SUS) after every session. We examined the SUS scores across the five sessions separately for Group A and Group B. Group A generally had SUS scores at or slightly above the mean ($M = 70.09$); Group B had a lower mean score ($M = 59.95$). Notably, in both groups, after an initial decline in Session 2, SUS scores for both groups did not decline further, and actually increased in the later sessions, particularly for Group A. Across all five sessions and all participants the average SUS score was 62.83 (SD = 11.47).

Across all sessions and participants (N = 26), positive affect on the PANAS was significantly positively correlated with SUS scores, Pearson r = 0.51 [0.15, 0.75], p = 0.008. The uniformly low levels of negative affect on the PANAS were not correlated with SUS, Pearson r = −0.19 [−0.54, 0.22], p = 0.36. These correlations addressing *RQ2b* (how is user affect influenced by *general* system usability?) are demonstrated in Figures 4 and 5, respectively.

6.5.5 Dimensions of CA Impressions from Session-Specific Usability Ratings. To develop a better picture of user responses to the CA, we ran principal components analysis using the post-session

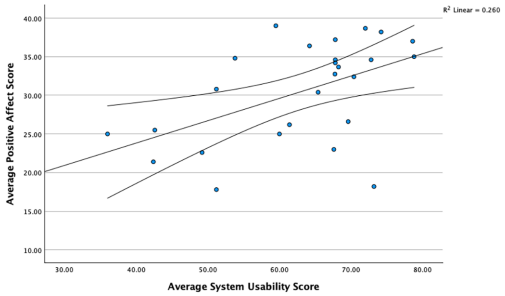

Fig. 4. Positive affect as measured by PANAS had a significant positive correlation with usability scores.

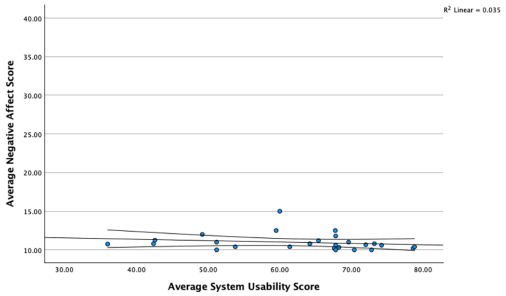

Fig. 5. Negative affect as measured by PANAS was not correlated with usability scores.

Table 4. Component Analysis of Session-Specific CA Usability Questions

Question	Component 1	Component 2	Component 3
Q1: CA pleasant	0.444	**0.723**	0.098
Q2: CA sociable	0.486	**0.643**	0.166
Q3: Felt comfortable sharing	0.155	0.113	**0.893**
Q4: Felt could be open	0.206	0.326	**0.884**
Q5: Felt involved	0.362	**0.506**	0.374
Q6: Enjoyed interaction	0.164	**0.923**	0.130
Q7: Interaction was smooth	**0.813**	0.381	−0.040
Q8: Would like to interact again	0.029	**0.885**	0.118
Q9: Interaction was satisfying	0.440	**0.711**	0.230
Q10: CA said right thing	**0.775**	0.172	0.199
Q11: CA responded appropriately	**0.836**	0.152	0.103
Q12: CA communicated correctly	**0.826**	0.230	0.302
Q13: CA seemed competent	**0.811**	0.238	0.253
Q14: CA seemed natural	**0.830**	0.287	0.118
Q15: Would like a long conversation	0.408	**0.669**	0.318

Extraction method: PCA. Rotation method: Varimax with Kaiser normalization. Rotation converged in five iterations. Each column (components 1, 2, and 3) corresponds to an identified dimension of participants' CA impressions. A bolded element in these columns indicates that the CA usability question belongs to the dimension, for example, Q8 (would like to interact again) belongs to Component 2, Pleasant-Enjoyable. The full text of the 15 post-session questions can be found in Appendix C.

survey questions pertaining to participant impressions of the CA. These analyses are reported in Table 4.

From this component analysis, we find three primary dimensions of participants' CA impressions. We call the first Open-Share (Component 3 in Table 4), which encompasses impressions related to how genuine participants felt they were able to be with the CA and how authentic the interactions were. The second, Pleasant-Enjoyable (Component 2 in Table 4), relates more to the social aspects of the interaction such as whether it was enjoyable and whether the participant would like to repeat the experience. The third, Appropriate-Natural (Component 1 in Table 4), encompasses impressions such as whether or not the CA gave appropriate responses to a variety of input and seemed generally competent as an intelligent conversational partner.

Table 5. Correlations of Session-Specific CA Usability Components with PANAS Scores

Component and Session	Positive Affect	Negative Affect
S1 Open_Share	r = 0.39, p = 0.057^{m}	r = 0.23, p = 0.29
S2 Open_Share	r = 0.31, p = 0.17	r = 0.12, p = 0.059
S3 Open_Share	r = 0.26, p = 0.21	r = −0.18, p = 0.41
S4 Open_Share	r = 0.42, p = 0.036*	r = 0.27, p = 0.19
S5 Open_Share	r = 0.37, p = 0.087^{m}	r = −0.20, p = 0.37
S1 Pleasant_Enjoyable	r = 0.44, p = 0.29*	r = 0.07, p = 0.74
S2 Pleasant_Enjoyable	r = 0.54, p = 0.009**	r = 0.17, p = 0.46
S3 Pleasant_Enjoyable	r = 0.74, p < 0.001***	r = −0.29, p = 0.17
S4 Pleasant_Enjoyable	r = 0.58, p = 0.002**	r = −0.37, p = 0.065^{m}
S5 Pleasant_Enjoyable	r = 0.72, p < 0.001***	r = −0.26, p = 0.25
S1 CA_Appropriate _Natural	r = 0.53, p = 0.006**	r = −0.10, p = 0.63
S2 CA_Appropriate _Natural	r = 0.25, p = 0.30	r = 0.02, p = 0.92
S3 CA_Appropriate _Natural	r = 0.60, p = 0.002**	r = −0.19, p = 0.38
S4 CA_Appropriate _Natural	r = 0.52, p = 0.007**	r = −0.047, p = 0.018*
S5 CA_Appropriate _Natural	r = 0.73, p < 0.001***	r = −0.41, p = 0.056^{m}

*p < 0.05, **p < 0.01, ***p < 0.001, ^{m} 0.05 < p < 0.10.

6.5.6 Session-Specific CA Usability and PANAS. We further examined the dimensions of participant impressions as relates to participant scores on the PANAS inventory. In general, all three CA Impression Components positively correlated with PANAS Positive Affect. Across the five sessions, the average (Fisher *Zr* transformed) correlation of Positive Affect with Open_Share was r = 0.37; the corresponding average for Positive Affect with Pleasant_Enjoyable was r = 0.72, and for Positive Affect with CA_Appropriate_Natural was r = 0.61. These averaged correlations address *RQ2b* (how is user affect influenced by *CA-specific* usability?); Table 5 also presents the session-wise correlations of these three CA impression components with positive affect and with negative affect.

6.5.7 User Behavior in Response to Minor System Errors. For our transcript-based analyses, we classified different system behaviors related to error states. We identified five distinct error types. These included:

(1) Loops, which we defined as instances where the CA repeated the same question it asked immediately previously.
(2) Interruptions. This included instances where the CA interrupted the participant while the participant was speaking. Instances of the participant interrupting the CA were ignored.
(3) Memory Failures. This included instances where participants indicated through dialogue that the CA failed to remember previously disclosed information—e.g., the participant said "I already told you that."
(4) Inappropriate Responses. We defined this error as instances where the CA responded to user utterances with the incorrect sentiment. We also included instances where the CA misidentified a person or hobby when it was disfluent enough with the rest of the interaction that the participant said something about it.
(5) Other Errors. This primarily consisted of situations where the CA entirely crashed and had to be restarted, or instances of the CA backtracking ("Let me go back.") for no apparent reason.

In general, across all sessions, there were comparatively few instances classified as *CA loops* (mean = 0.33); the frequency of loops was largely similar for Group A (mean = 0.29) and Group B (mean = 0.37), and across sessions (for Sessions 1–5, respectively, means = 0.39, 0.29, 0.32, 0.36, and 0.29).

Instances classified as *CA interruptions* of the participant occurred about 1.52 times per session. Interruptions were slightly and non-significantly more common for Group A (mean = 1.94) than for Group B (mean = 1.10); interruptions were comparatively constant across sessions (for Sessions 1–5, respectively, means = 1.54, 1.61, 2.11, 1.21, and 1.14).

Overall, *CA memory failures* were noted about 0.35 times per session. Memory failures were equally commonly noted in Group A (mean = 0.37) and Group B (mean = 0.32), but appeared to be elevated in particularly Session 2 compared with the first session and later sessions (means for Sessions 1–5, respectively = 0.15, 0.81, 0.23, 0.25, and 0.30), effect of session, $F(4, 72) = 2.33$, $p = 0.064$, but with a significant cubic effect, $F(1,18) = 5.43$, $p = 0.032$, partial $\eta^2 = 0.23$.

On average, about 1.53 of the CA responses were evaluated as *inappropriate*. The rated number of inappropriate responses was largely similar for the two groups (mean for Group A = 1.17, mean for Group B = 1.89), and was generally similar across sessions (means for Sessions 1 through 5, respectively = 1.39, 1.75, 1.43, 1.64, and 1.43), $F < 1$ for the effect of session, $F < 1$ for session × group, and $F = 1.40$ for the effect of group.

Other CA errors were relatively infrequent (mean = 0.57); they occurred approximately equally often in Group A (mean = 0.46) and Group B (mean = 0.69), and approximately equally across the five sessions (means for Sessions 1 through 5, respectively = 0.43, 0.82, 0.68, 0.43, and 0.50). $F < 1$ for the effect of session, $F < 1$ for session × group, and $F = 1.33$ for the effect of group.

Given that, across all sessions, there were comparatively few minor system errors, we did not separately correlate minor system errors with user affect (*RQ2c*). Nonetheless, considering specifically the CA Impression Component CA_Appropriate_Natural, from the results presented in Table 5 it can be seen that there was generally a modest to strong positive correlation between participants' positive affect and their assessment of the extent to which the CA was natural and appropriate; given the generally low levels of negative affect, the correlations of this component with negative affect were generally near zero.

6.5.8 Participant Profiles with Technology. Participants completed surveys at the start of the experiment to describe their recent use and perception of technology. These self-report measures allowed us to develop a more complete picture of participants' interests in and experiences with recently evolving technology in their day-to-day lives, which we present as a profile here.

(a) Interests. Participants were generally curious about existing technologies and interested in exploring new ones, and also indicated some attachment to the devices they currently have. Statistics for each item on the interests inventory are included in Table D1 (Appendix D).
(b) Current Technology Skills. Participants were generally fairly skilled with existing technologies and self-reported as feeling confident in using technology to support a variety of tasks, from engaging with community events to streaming media. Participants also reported as feeling confident in performing tasks such as troubleshooting and printer maintenance, indicating that overall participants had high competency with both hardware and software. Table D2 (Appendix D) provides a profile of participant technology skills.
(c) Current Experience Levels. In general, participants rated themselves as having more experience with the first five activities in Table D2 and less experience with the last five activities (shown by the means for the items). There was also greater variability in levels of experience with the last five activities in Table D2, as shown by the SD and range, with all five levels of experience (from very poor to excellent) reported across the different participants for the last five activities.

Fig. 6. Means for participant interest-experience groupings.

6.5.9 Effects of Experience with Technology. To develop a better picture of user behavior, we examined whether the pattern of results observed in the full sample session-by-session data differed for individuals with comparatively higher vs. lower levels of interest, skills, and experience with computers and information technology.

(a) Interest-Experience Groups. To examine this question we first found participants' average responses to each of the three subsets of questions relating to Technology Interests/Attitudes (8 items), Technology Skills (12 items), and Experience with Information Technology (10 items). The averages for these subscales were significantly positively correlated with one another (Interests-Skills, $r = 0.55$, $p = 0.003$, Interests-Experience, $r = 0.55$, $p = 0.003$, Experience-Skills, $r = 0.83$, $p < 0.001$), so we then averaged these subset averages to find a grand average. Participants with values above the median on this grand average measure were placed in the High Interest, Skills, and Experience Group ($N = 13$, average = 4.65, $SD = 0.14$) and participants below the median were placed in the Low Interest, Skills, and Experience Group ($N = 13$, average = 3.72, $SD = 0.50$). Figure 6 shows the means for the two Interest-Experience groups for the three subscales.

We then compared participants in these two Interest-Experience groups who had complete data across the five sessions for: (i) PANAS scores; (ii) the three components regarding their session-specific ratings of the CA, that is, Open-Share, Pleasant-Enjoyable, and Appropriate-Natural; and (iii) SUS scores.

(b) Affect and Interest-Experience Level. Figure 7 shows the means ($N = 11$ for High I&E, $N = 8$ for Low I&E) for Positive affect scores across the five sessions. From this we can see that, across the five sessions, Positive affect scores for the high Interest/Experience Group uniformly exceeded Positive affect scores for the low Interest/Experience Group. Nonetheless, the level of positive affect in the low Interest/Experience group remained quite consistent across the five sessions, and positive affect in both groups slightly (numerically) increased in the later sessions. Paralleling the findings for the full sample, Figure 8 demonstrates that negative affect scores were uniformly low, and there was no main effect of Interest-Experience group and no interaction, potentially due to a floor effect. These results address *RQ2d*: Despite between-group differences in overall positive affect, positive affect remained quite high in both the group with lower and the group with stronger I&E with technology; furthermore, regardless of participants' level of technology-related experience, positive affect did not decline (and negative affect did not increase) in the later of the five interaction sessions with the CA.

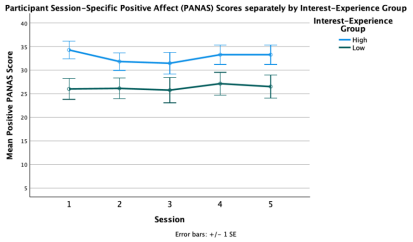

Fig. 7. Session-specific positive affect scores by I-E group.

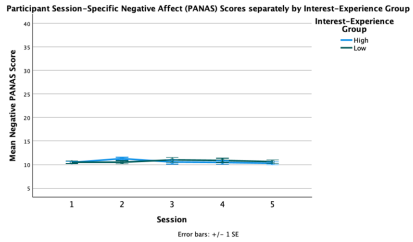

Fig. 8. Session-specific negative affect scores by I-E group.

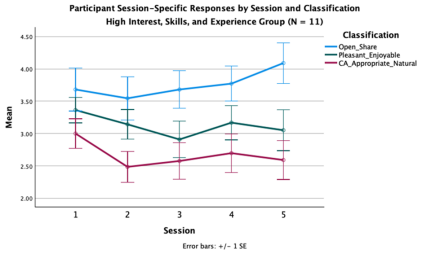

Fig. 9. Session-specific CA usability dimensions among the High Interest-Experience Group.

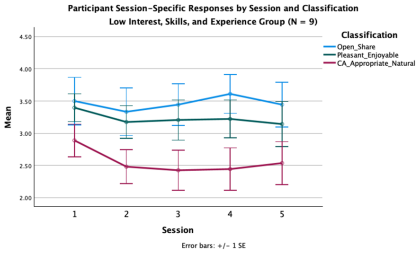

Fig. 10. Session-specific CA usability dimensions among the Low Interest-Experience Group.

(c) Session-Specific CA Usability Dimensions. Figures 9 and 10 present the three component scores (Open-Share, Pleasant-Enjoyable, and CA-Appropriate-Natural) across the five sessions separately for the High Interest-Experience Group (N = 11) and the Low Interest-Experience Group (N = 9).

Notably, in both High and Low Interest-Experience groups, the Open-Share component had the highest endorsement rate, and this stayed comparatively high across the five sessions. Although the strength of endorsement varied between groups, they showed generally similar patterns that remained consistent across groups. The scores across sessions for Appropriate-Natural were generally the lowest; however, both groups indicated they found the conversations fairly enjoyable and felt comfortable sharing information with the CA. The Low Interest-Experience group was generally less comfortable sharing and this endorsement stayed fairly consistent, whereas the High Interest-Experience group increased their scores for this component over time. This indicates that further investigation into Low Interest-Experience users may be helpful in determining any other factors influencing these scores, as well as making the use of the CA more enjoyable regardless of a user's previous experience with technology.

6.5.10 Post-Study Rating of Total CA Experience. Participant responses to the seven specific uses questions again showed large variability, such that across participants both the minimum and maximum possible score were given for each item. The highest mean endorsements were given in response to the questions of whether the CA could "assist with reminders" and "help with scheduling"; the lowest mean endorsement (even below the midpoint of the scale) was given in response to the question of whether the CA could "help others you live with (e.g., family, friends) stay accountable to their tasks and/or goals."

6.6 Exploratory Analyses

6.6.1 Conversational Analysis with LIWC. In addition to inventories at the start of the experiment and user self-report metrics, we also performed psycholinguistic analysis of the contents

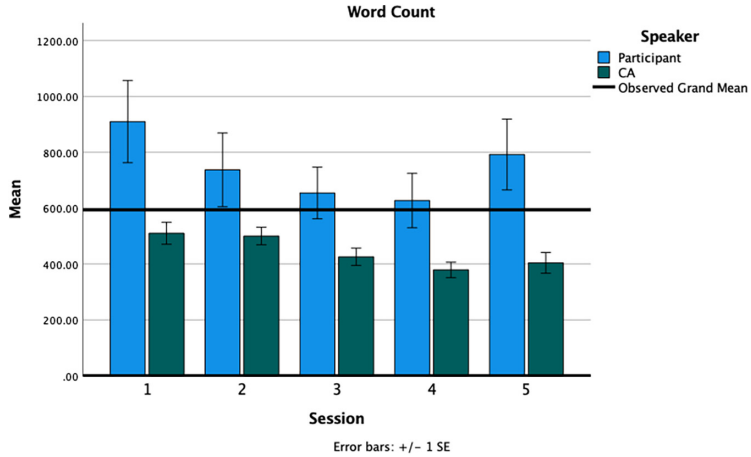

Fig. 11. Raw word count for Calena and participants across five sessions.

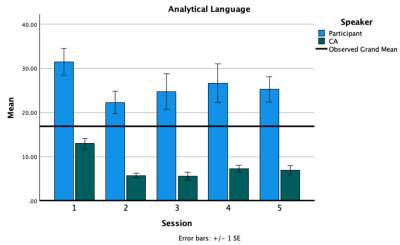

Fig. 12. Participant and CA analytical language across five sessions.

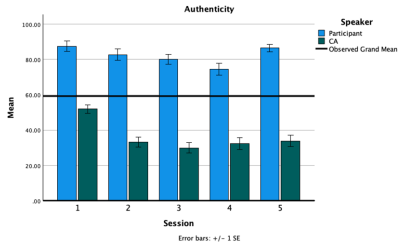

Fig. 13. Participant and CA authenticity across five sessions.

of the interactions with the prototype CA. In order to provide the most sensitivity to changes in psycholinguistic characteristics across time, the results are presented by session for the full sample. We focused our examination on the four summary variables discussed above, plus the raw word count. These analyses were performed with participants who attended all five experimental sessions (N = 21).

Raw word count provides a general picture of user involvement in an interaction separate from the actual contents of the interaction. A 2 (speaker) × 2 (group) × 5 (session) mixed-factor ANOVA on Word Count revealed only a significant effect of speaker, with participants overall speaking substantially more (M word count = 793.02) than the CA (M word count = 449.76), $F(1, 19) = 23.09$, $p < 0.001$, partial $\eta^2 = 0.55$. Figure 11 demonstrates that over five sessions, participants remained fairly consistent in their involvement and the length of conversations.

Examination of the four LIWC summary variables was also revelatory. First, we considered Analytical Language and Authenticity across time for both participants and the CA. These trends are shown in Figures 12 and 13, respectively.

Consistent with these trends, a 2 (speaker) × 2 (group) × 5 (session) mixed-factor ANOVA on Analytical Language revealed a significant effect of speaker, with participants overall using more analytical language ($M = 25.95$) than the CA ($M = 7.80$), $F(1, 19) = 42.06$, $p < 0.001$, partial $\eta^2 = 0.69$. There was also a significant effect of session, $F(4, 76) = 5.55$, $p < 0.001$, partial $\eta^2 = 0.23$, reflecting both (within the main effect of session) a linear decrease in analytical language across sessions, $F(1, 19) = 6.42$, $p = 0.02$, partial $\eta^2 = 0.25$, and a quadratic effect, with a larger decrease from the

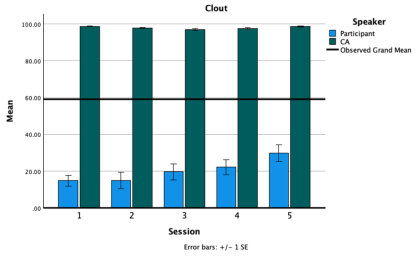

Fig. 14. Participant and CA clout across five sessions.

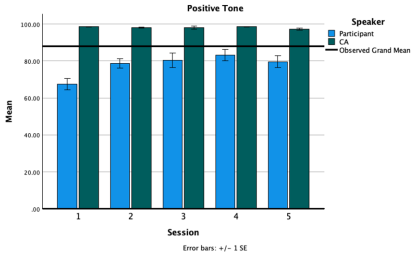

Fig. 15. Participant and CA positive tone across five sessions.

first to the second session than for subsequent sessions, $F(1, 19) = 8.72$, $p = 0.008$, partial $\eta^2 = 0.32$. The marked decrease in analytical language, as seen in Figure 12, indicated participants became more comfortable and open with the CA over time.

Authenticity is another good measure of participant comfort and openness. A 2 (speaker) × 2 (group) × 5 (session) mixed-factor ANOVA on Authenticity revealed a significant effect of speaker, $F(1, 19) = 288.85$, $p < 0.001$, partial $\eta^2 = 0.94$, with participants overall using more authentic language ($M = 81.51$) than the CA ($M = 35.62$). There was also a main effect of session, $F(4, 76) = 9.52$, $p < 0.001$, and a speaker × session interaction, $F(4, 76) = 5.48$, $p < 0.001$, mostly reflecting a greater decrease in authenticity across sessions for the CA than for participants, particularly from Session 1 to Session 2. The overall consistent level of participant authenticity over five sessions indicated not only a comfort with the system but pointed to ongoing and sustained engagement, vital for long-term deployment. This trend can be seen in Figure 13.

We next considered the LIWC summary variables Clout and (Positive) Tone. In LIWC analyses, Clout is a measure of authority and confidence in a conversation, while Tone is a measure of the emotional valence of a conversation, where a higher score indicates a more positive valence. The trends for these two variables over five sessions are shown in Figures 14 and 15, respectively.

A 2 (speaker) × 2 (group) × 5 (session) mixed-factor ANOVA on Clout revealed a significant effect of speaker, with the CA overall demonstrating substantially more and ceiling levels of clout ($M = 97.86$) than the participants ($M = 21.70$), $F(1, 19) = 923.90$, $p < 0.001$, partial $\eta^2 = 0.98$. There was also a significant effect of session, $F(4, 76) = 3.19$, $p = 0.018$, partial $\eta_2 = 0.14$, and a significant speaker × session interaction, $F(4, 76) = 2.97$, $p = 0.025$, partial $\eta^2 = 0.14$ which reflected increasing clout across the sessions for the participant, $F(1, 19) = 14.78$, $p = 0.001$, partial $\eta^2 = 0.44$ for the linear effect within the speaker × session interaction. Participant clout increased steadily throughout all five sessions, indicating that participants became more acclimated to and confident with using the CA. This can be seen in Figure 14.

A 2 (speaker) × 2 (group) × 5 (session) mixed-factor ANOVA on (Positive) Tone revealed a significant effect of speaker, $F(1, 19) = 140.31$, $p < 0.001$, partial $\eta^2 = 0.88$, which reflected a near-ceiling level of positive tone for the CA ($M = 98.00$) but also quite positive tone overall for participants ($M = 77.71$). Notably, there was also a main effect of session, $F(4, 76) = 3.07$, $p = 0.021$, and a speaker × session interaction, $F(4, 76) = 3.65$, $p = 0.009$, reflecting the finding that positive tone generally increased across the sessions for participants, $F(1, 19) = 10.52$, $p = 0.004$, partial $\eta^2 = 0.36$ for the linear effect in the speaker × session interaction, including an increase from Session 1 to Session 2, $F(1, 19) = 4.97$, $p = 0.038$, partial $\eta^2 = 0.21$ for the quadratic effect in the speaker × session interaction. The slight but general increase in positive tone is another promising indicator that users remained engaged and interested in the interactions with the CA over time.

7 Discussion

7.1 Key Findings

This study was designed to address four primary RQs. *RQ1* pertained to whether deliberate positivity can be effectively implemented as a conversational strategy to encourage continued user engagement with a conversational system. We find multidimensional results indicating that the answer to this is firmly *yes*. Behavioral measures such as psycholinguistic analysis indicate that user engagement over a week of interactions remained consistent and relatively high. Further, participant mood did not decline either in Phase II with a single prompt or in Phase III using a more standardized inventory measuring affect. The findings regarding *RQ2a* (the impact of number of positive prompts on user affect) and *RQ2b* (the impact of session-specific usability metrics on user affect) are mixed under this paradigm. There was no evidence that user affect was *differentially* influenced by the number of positive prompts per session (RQ2a); however, user positive affect remained consistently high across sessions. When examining *RQ2b* and *RQ2c*, it is unclear whether session-specific usability influences user affect. There was a floor effect in that the number of minor system behavior incidents (e.g., loops or memory failures) was low and very consistent across sessions, which resulted in low variance in user self-report of usability. Moreover, positive affect remained high and very consistent across sessions. This potentially indicates that any effect probed by RQ2b and 2c may be better captured with a finer-grained or strongly temporally tied measure, such as probing for usability much closer to any incidence of a minor behavioral issue in the system. RQ2a is also still somewhat of an open question as it is unclear whether our findings are fully generalizable or are more a result of constraining the number of positive prompts per session to one or four prompts. Future work extending this paradigm to contain even more positive prompts or to have participants interact with a CA over a longer time frame may prove more revelatory in addressing RQ2a, RQ2b, and RQ2c.

With regard to *RQ2d* and the question of the influence of user technological experience and interest levels on affect, our results provide a promising avenue for future study. While participant I&E levels with technology did not seem to influence the overall impact of deliberate positivity, participants in the High I&E group demonstrated a markedly higher positive affect throughout the study than did participants in the Low I&E group. However, patterns in affect observed in both groups mirrored one another closely. This indicates that any effect of the deliberate positivity intervention may be consistent regardless of participant experience level, even if the overall emotional valence is lower in the case of participants with lower I&E levels.

7.2 Implications for Future Design

The answers to the RQs for this study point to several key implications for the future design of conversational systems. These primary implications involve both the use of deliberate positivity in conversational design and the structuring of user-driven design methods, particularly with older adults, but potentially also with other (often underserved) user groups.

7.2.1 Positivity as a Tool. Our findings suggest that, in convergence with the literature on the benefits of positive capitalization in human-human conversation, deliberate positivity as a conversational strategy can have a measurable positive impact on user affect, and may help drive and maintain user engagement with a system over the longer term. We find evidence through multiple methodologies that users stayed consistently positive throughout the interactions, and moreover are consistently receptive to and engaged with a system implementing the deliberate positivity conversational strategy. This is further strengthened by the length of the study; while the experiments in Phase III were limited to a working week (5 days), evidence of user engagement

and positive affect did not drop appreciably over the course of the experiment despite indications of the novelty effect subsiding.

It is also noteworthy as a strength of this study that the evidence that our implementation of deliberate positivity is acceptable is diverse and is reflected both in user behavior and in user report. The psycholinguistic characteristics of user interactions with the system are consistent with profiles of sustained user engagement and interest. The raw word count of participant interactions with Calena, for example, can be used as a base metric of engagement; despite minor variations in later sessions, participants consistently remained fairly talkative, and more importantly were more consistently verbose than the Calena prototype. Authenticity, an LIWC psycholinguistic measure relating to the amount of openness, honesty, or "sharing" in a given linguistic sample, also remained high through all five sessions. LIWC Clout, or authoritativeness/confidence, is also a particularly interesting indicator of engagement; while participant clout was consistently lower than that of Calena, it also noticeably increased throughout the course of the experiment, potentially indicating a consistent increase in participant comfort and familiarity with a system—a stage of the adoption process in which interest generally wanes—while other measures of engagement such as word count and authenticity stayed high.

Outside of linguistic evidence, the standardized measure of positive and negative affect offered by PANAS for participants also revealed the general acceptability of deliberate positivity as a strategy for users. Positive affect across all sessions remained consistently high, as seen in Figure 3, and negative affect scores also remained uniformly low. Equally importantly, similar patterns in positive affect were reflected across participants at both ends of the interest-experience spectrum, as seen in Figure 7, and there was no real effect on participant negative affect in both groups as seen in Figure 8. This is an indicator that deliberate positivity in the *worst case* had a neutral impact on user affect, which is a very promising result in terms of the potential long-term use of a deliberately positive conversational strategy. Even in cases where no discernible increase in positive affect was seen, positive affect remained consistent and, equally importantly, negative affect was not elevated or exacerbated.

7.2.2 User-Driven Design. This work highlights implications for the *process* of designing intelligent systems for older users just as much as the design of the systems themselves. Chief among the design practices foregrounded by this study is the necessity of ensuring that users involved in the design process are a representative sample of the target user base, particularly when the target users are older adults. The iterative process we used in designing the Calena prototype included older adults in all phases of the design, and allowed us to fine-tune our implementation of positivity via direct feedback from users. The reporting done by participants in the pilot projects of Phase I and in Phase II (Wizard of Oz) gave us insight as to what sorts of questions users would be receptive to when interacting with the system, which further enabled us to design a system that allowed us to actually investigate our RQs more directly. The fact that large portions of the prototype design (domains of interest, positive prompts from the CA) were directly driven by feedback from our target user population is a major strength of this study, and fully supports the value of including a user group in design from the ground up.

Ensuring users involved in the design process are a representative sample of the target user base is also a design practice we underline the importance of—not only for older adults, but users of any age group or life circumstance. Many of these user groups are noted in previous work to have drastically different characteristics or requirements; for example, the design of smart technologies for children may need to consider developmental impacts or involve parental oversight, as suggested in [91], or be well-facilitated by play and physical prototyping in the design process as in [31, 101]. In contrast, design for adults might emphasize data privacy or transparency [58, 96] in addition to

trustworthiness [15], while design for adolescents might focus on social aspects like support for navigating peer and school stressors or simply being an empathetic listening presence [41, 60, 66]. We highlight that target users—whether grouped by age, medical condition, lifestyle, or other characteristics—should always be a fundamental part of the design and testing process precisely because acceptability and usability varies so widely. In addition, defining and measuring usability and acceptability is still not well understood or explored for a number of user groups and related technologies (e.g., [23]). In the case of adults, and particularly older adults, addressing the issues of long-term adoption that have long plagued intelligent systems (see, e.g., [30]) requires the inclusion of both users with little experience with technology and those with high interest in and experience with technology in order to gain the best insight into the acceptability of any given piece of technology. A major strength of the current study is exactly this portion of the methodology, as discussed in Sections 6.5.8 and 6.5.9. The emergence of two distinct participant groups with regards to attitudes towards and experience with technology clearly demonstrates that our results have some generalizability across a wide range of users. It is also noteworthy in considering future design and testing practices—despite the necessity of conducting the study via video conference on Zoom, we were still able to reach a reasonably wide variety of participants, indicating that remote testing may be feasible for future work focusing on different user demographics.

7.2.3 Challenges in Real-World Use. The design, implementation, and testing of the systems used in this study also reinforce some existing challenges in deploying a CA such as Calena in the real world—namely, security and long-term adoption. First, data security and privacy remains a major concern for both storage of any gathered user data and access to the CA itself. While an informal examination of experimental transcripts showed that little very highly sensitive information such as health data was disclosed *for this study specifically*, evidence suggests that social conversations with older users can still lead to the disclosure of things such as health information [6, 15]. Multiple possibilities exist to help keep user data as secure as possible. First, as in the PKB in this study, data should all be stored locally; in real-world deployment this would mean devices with higher storage capacity that should be password protected and ideally involve additional security restrictions (e.g., additional independent authentication). Second, we designed the Calena CA such that the contents of the knowledge base cannot be queried directly by users. Third, the challenges specific to unimodal voice interaction must be considered.

Voice interaction presents a number of challenges by its nature, both in general and in ways more specific to applications such as we propose in this study. We considered two primary sources of vulnerability in our prototype and experimental design: data security and the presence of others in the home. The CA pipeline involves both speech-to-text and text-to-speech steps, which ideally should be done locally on the device itself to provide the best security. In this study, we primarily mitigated potential risks associated with these aspects by ensuring that conversational data was processed through a service that guaranteed data privacy and ownership, as well as storing all data securely. For older users in particular the potential presence of others in the room or house may also be a concern for voice systems. We mitigated potential risks related to overhearing interactions with the CA by restricting the number of people attending an experimental session to one RA interacting with the participant and one trained RA on standby to troubleshoot if necessary; the participants were also asked to be in an empty room. In wider deployment, this is more difficult to guard against—particularly when a primary design consideration for older users is accessibility. Any authentication measures should be as unobtrusive as possible and, ideally, not depend on a user remembering a password, interacting with a device such as a smartphone, or wearing a separate sensor. A number of approaches have been suggested that may facilitate this. For example, a voiceprint biometric identifier—particularly one linked to a wake word or phrase—would minimize

cognitive load. An embedded sensor in a frequently used item such as glasses or jewelry may also provide a natural way to authenticate users [25]. Checks like wider scale liveness detection may also be useful in safeguarding against any malicious actors [99]. Some combination of these sorts of security features may give users the highest level of both ease of use and security.

The second major challenge in deploying a system like Calena for wider use lies in promoting and maintaining long-term adoption. We note that for our design there were two primary concerns: avoiding repetitive interactions by remembering prior user responses, and encouraging further user engagement by adding depth to the dialogue tree. Both of these can be addressed by learning about users, which foregrounds the need for a well-structured user knowledge base that can be effectively reasoned with in order to enrich interactions. In this study, we addressed this by restricting the domains of interest and incorporating rule-based dialogue trees as described in Section 6.3.2. In real-world usage, this may be largely infeasible. However, the success of domain-restricted design like Calena suggests that one potential way to better engage with users is to focus on domain expertise in agent design rather than generalization, and limit supported features and tasks based on user characteristics and contexts. This is an especially impactful consideration when taking into account existing evidence about use patterns among older adults for conversational systems demonstrating that certain sets of tasks are more widely adopted than others by these users [76, 82, 96]. Further evidence also suggests that usability on a task- and domain-based level, such as frustration with lack of customization and support for deeper dialogue trees in older adults seeking health information [15], may have a heavy impact on long-term acceptance [72]. Both of these factors, combined with our findings, reinforce that designs for CAs such as Calena should pay close attention to the selection of supported conversational domains, particularly in mental health and social contexts where a higher degree of expertise and longer dialogues may be desirable.

7.3 Limitations and Future Directions

The current experimental paradigm has helped pinpoint further questions to be investigated and modifications that can be made to future experiments. First, the promising results regarding the effects of deliberate positivity in conversational design on user affect and perceived usability of a system indicate that positivity as an interaction strategy is worth investigating further. However, this study was limited to one week of interactions with users. Much of the evidence in human-human interaction surrounding the effectiveness of deliberate positivity as a strategy in improving health and well-being, particularly mental health, implies that the full impact of encouraging positivity may not be observed in the short-term—that is, the overall effect on well-being is not immediately temporally tied, and seeing the entire picture requires longer follow-up. Therefore, the full impact of positivity as a conversational strategy requires a more longitudinal approach; extending the experimental time frame, perhaps with interactions occurring over a series of weeks or several months, rather than a single week, may give us deeper insight into the true effectiveness of the strategy. In the current experiment, we explored positivity as a conversational tool by varying the number of deliberate positive injections into a conversation with a user. With a longer-term design, it would be possible to probe deeper into these effects, such as increasing the variation in the number of positive prompts or studying the effects of different patterns of positive manipulations such as clustering.

Our probes into positivity as a conversational strategy were also constrained to purely prompting users to reflect on positive emotions and events. Previous literature, as well as our results, suggests this is an effective method of promoting positive capitalization. However, there may be other effective socioconversational strategies to promote positivity that are worth investigating. A logical extension of this work would be to vary the directness or indirectness of the conversational strategy in prompting positive reflection. The encouragement towards positivity in our conversational

design was also exclusively agent driven, which may have altered some of the dynamics of the interactions with users. The means of engaging in deliberate positivity in more mixed-initiative interactions would also be an interesting question to explore in future work.

Part of exploring the effectiveness of this conversational strategy also involved a session-wise examination of the usability and acceptability of the prototype system itself. For this work, we opted to use the SUS as a general usability assessment, supplemented by usability questions tailored to language-based interactive systems specifically. We chose the SUS both because it is one of the most widely used tools for assessing usability and further because it probes user impressions of adoptability, which is key for systems designed for long-term use. We also included an inventory of ratings specific to the design and implementation of the conversational system, including both language and social impressions. This new tool was beneficial in revealing and separating dimensions of CA usability such as how comfortable a user felt with being open and sharing with the system, how pleasant or enjoyable the interaction was, and how appropriate or natural the interaction felt. More recent work has introduced in-depth usability inventories specifically for conversational systems, which have since been validated [11, 12]. Future work extending the current study might add the use of more recent fine-tuned inventories to probe impressions of language-based interactions specifically, particularly as future work may explore user interactions and acceptability over longer time periods than the 1 week limitation of this study. This would give us a more holistic view of evolving user attitudes over time, building on the period of initial adoption covered in this study.

The current study was also somewhat limited environmentally. Interactions were only partially *in situ*, primarily by necessity. All phases of the study were moved to video conferencing due to the COVID-19 pandemic, which allowed participants to interact with all versions of the system (Wizard of Oz or actual prototype) from their own homes. This gave us an opportunity to examine user behavior with the prototype in their own home environments, which is important when examining questions related to design for independent older adults. However, this also meant that an experimenter was always present on the video call (albeit not observing or listening to the interaction), and the actual prototype agent was physically hosted elsewhere. Evidence suggests that spatial and social contexts do impact user behavior with dialogue systems, particularly when it comes to sharing behaviors and more personal disclosure [57]. An immediate next extension to this work might then be to create prototype hardware that can be deployed in participants' homes so that participants can interact with the system in the most natural context possible.

8 Conclusions

In this three-phase mixed-methods study we have shown both the feasibility and acceptability of a social conversational technology for older adults. We have shown that a conversational strategy employing deliberate positivity may be effective in promoting continued use and interest in a conversational system. We have provided a profile of user interactions with a CA designed to discuss select topics of interest with older adults, namely hobbies and family. We further perform an analysis of a combination of multiple forms of user self-report and psycholinguistic content, as well as an *a posteriori* examination of user and system behavior through transcript analysis. Through this, we show that a deliberately positive conversational strategy may be effective in maintaining user interest over a longer series of interactions.

Acknowledgments

The authors would like to thank Dr. Yihan Wu, Naome Etori, and Kayla Chan for their discussion and help with data collection and processing. The authors also express gratitude to Drs. Serguei

Pakhomov, Martin Michalowski, and Michael Kotlyar for their contributions to discussion and experimental design, as well as Raymond Finzel for his troubleshooting assistance.

References

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) (NAACL '19)*. Association for Computational Linguistics, 54–59. DOI: https://doi.org/10.18653/v1/N19-4010

[2] Felwah Alqahtani, Alaa Alslaity, and Rita Orji. 2022. Usability testing of a gratitude application for promoting mental well-being. In *International Conference on Human-Computer Interaction*. Springer International Publishing, Cham, 296–312. DOI: https://doi.org/10.1007/978-3-031-05412-9_21

[3] Apache. 2023. *Apache TinkerPop Gremlin Query Language*. The Apache Software Foundation. Retrieved December 20, 2023 from https://tinkerpop.apache.org/gremlin.html

[4] Anne Arewasikporn, John A. Sturgeon, and Alex J. Zautra. 2019. Sharing positive experiences boosts resilient thinking: Everyday benefits of social connection and positive emotion in a community sample. *American Journal of Community Psychology* 63, 1–2 (2019), 110–121. DOI: https://doi.org/10.1002/ajcp.12279

[5] Raluca Balan, Anca Dobrean, and Costina R. Poetar. 2024. Use of automated conversational agents in improving young population mental health: A scoping review. *NPJ Digital Medicine* 7, 1 (2024), 75. DOI: https://doi.org/10.1038/s41746-024-01072-1

[6] Scott Beach, Richard Schulz, Julie Downs, Judith Matthews, Bruce Barron, and Katherine Seelman. 2009. Disability, age, and informational privacy attitudes in quality of life technology applications: Results from a national web survey. *ACM Transactions on Accessible Computing* 2, 1 (2009), 1–21. DOI: https://doi.org/10.1145/1525840.1525846

[7] John R. Beard, Alana Officer, Islene Araujo De Carvalho, Ritu Sadana, Anne Margriet Pot, Jean-Pierre Michel, Peter Lloyd-Sherlock, JoAnne E. Epping-Jordan, G. M. E. E. Geeske Peeters, Wahyu Retno Mahanani, et al. 2016. The world report on ageing and health: A policy framework for healthy ageing. *The Lancet* 387, 10033 (2016), 2145–2154. DOI: https://doi.org/10.1016/S0140-6736(15)00516-4

[8] Eileen Bendig, Benjamin Erb, Echo Meißner, Natalie Bauereiß, and Harald Baumeister. 2021. Feasibility of a software agent providing a brief intervention for self-help to uplift psychological wellbeing ("SISU"). A single-group pretest-posttest trial investigating the potential of SISU to act as therapeutic agent. *Internet Interventions* 24 (2021), 100377. DOI: https://doi.org/10.1016/j.invent.2021.100377

[9] Idrissa Beogo, Drissa Sia, Eric Tchouaket Nguemeleu, Junqiang Zhao, Marie-Pierre Gagnon, and Josephine Etowa. 2022. Strengthening social capital to address isolation and loneliness in long-term care facilities during the COVID-19 pandemic: Protocol for a systematic review of research on information and communication technologies. *JMIR Research Protocols* 11, 3 (2022), e36269. DOI: https://doi.org/10.2196/46753

[10] Walter R. Boot, Neil Charness, Sara J. Czaja, Joseph Sharit, Wendy A. Rogers, Arthur D. Fisk, Tracy Mitzner, Chin Chin Lee, and Sankaran Nair. 2015. Computer proficiency questionnaire: Assessing low and high computer proficient seniors. *The Gerontologist* 55, 3 (2015), 404–411. DOI: https://doi.org/10.1093/geront/gnt117

[11] Simone Borsci, Alessio Malizia, Martin Schmettow, Frank Van Der Velde, Gunay Tariverdiyeva, Divyaa Balaji, and Alan Chamberlain. 2022. The Chatbot Usability Scale: The design and pilot of a usability scale for interaction with AI-based conversational agents. *Personal and Ubiquitous Computing* 26 (2022), 95–119. DOI: https://doi.org/10.1007/s00779-021-01582-9

[12] Simone Borsci, Martin Schmettow, Alessio Malizia, Alan Chamberlain, and Frank Van Der Velde. 2023. A confirmatory factorial analysis of the Chatbot Usability Scale: A multilanguage validation. *Personal and Ubiquitous Computing* 27, 2 (2023), 317–330. DOI: https://doi.org/10.1007/s00779-022-01690-0

[13] Anda Botoseneanu, Miriam R. Elman, Heather G. Allore, David A. Dorr, Jason T. Newsom, Corey L. Nagel, and Ana R. Quiñones. 2023. Depressive multimorbidity and trajectories of functional status among older Americans: Differences by racial/ethnic group. *Journal of the American Medical Directors Association* 24, 2 (2023), 250–257. DOI: https://doi.org/10.1016/j.jamda.2022.11.015

[14] Denis Boucaud-Maitre, Luc Letenneur, Moustapha Dramé, Nadine Taubé-Teguo, Jean-François Dartigues, Hélène Amieva, and Maturin Tabué-Teguo. 2023. Comparison of mortality and hospitalizations of older adults living in residential care facilities versus nursing homes or the community. A systematic review. *PLoS One* 18, 5 (2023), e0286527. DOI: https://doi.org/10.1371/journal.pone.0286527

[15] Robin Brewer, Casey Pierce, Pooja Upadhyay, and Leeseul Park. 2022. An empirical study of older adult's voice assistant use for health information seeking. *ACM Transactions on Interactive Intelligent Systems* 12, 2 (2022), 1–32. DOI: https://doi.org/10.1145/3484507

[16] Farid Chakhssi, Jannis T. Kraiss, Marion Sommers-Spijkerman, and Ernst T. Bohlmeijer. 2018. The effect of positive psychology interventions on well-being and distress in clinical samples with psychiatric or somatic disorders: A systematic review and meta-analysis. *BMC Psychiatry* 18 (2018), 1–17. DOI: https://doi.org/10.1186/s12888-018-1739-2

[17] Dennis S. Charney. 2004. Psychobiological mechanisms of resilience and vulnerability: Implications for successful adaptation to extreme stress. *American Journal of Psychiatry* 161, 2 (2004), 195–216. DOI: https://doi.org/10.1176/appi.ajp.161.2.195

[18] Minji Cho, Sang-su Lee, and Kun-Pyo Lee. 2019. Once a kind friend is now a thing: Understanding how conversational agents at home are forgotten. In *2019 ACM Designing Interactive Systems Conference (DIS '19)*. ACM, New York, NY, 1557–1569. DOI: https://doi.org/10.1145/3322276.3322332

[19] Kassandra Cortes and Joanne V. Wood. 2019. How was your day? Conveying care, but under the radar, for people lower in trust. *Journal of Experimental Social Psychology* 83 (2019), 11–22. DOI: https://doi.org/10.1016/j.jesp.2019.03.003

[20] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. What can I help you with? Infrequent users' experiences of intelligent personal assistants. In *19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, New York, NY, 1–12. DOI: https://doi.org/10.1145/3098279.3098539

[21] Nancy J. Donovan and Dan Blazer. 2020. Social isolation and loneliness in older adults: Review and commentary of a National Academies report. *The American Journal of Geriatric Psychiatry* 28, 12 (2020), 1233–1244. DOI: https://doi.org/10.1016/j.jagp.2020.08.005

[22] Melisa Duque, Sarah Pink, Yolande Strengers, Rex Martin, and Larissa Nicholls. 2021. Automation, wellbeing and digital voice assistants: Older people and Google devices. *Convergence* 27, 5, SI (2021), 1189–1206. DOI: https://doi.org/10.1177/13548565211038537

[23] Stefano Federici, Maria Laura de Filippis, Maria Laura Mele, Simone Borsci, Marco Bracalenti, Giancarlo Gaudino, Antonello Cocco, Massimo Amendola, and Emilio Simonetti. 2020. Inside Pandora's box: A systematic review of the assessment of the perceived quality of chatbots for people with disabilities or special needs. *Disability and Rehabilitation: Assistive Technology* 15, 7 (2020), 832–837. DOI: https://doi.org/10.1080/17483107.2020.1775313

[24] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies* 132 (2019), 138–161. DOI: https://doi.org/10.1016/j.ijhcs.2019.07.009

[25] Huan Feng, Kassem Fawaz, and Kang G. Shin. 2017. Continuous authentication for voice assistants. In *23rd Annual International Conference on Mobile Computing and Networking*. ACM, New York, NY, 343–355. DOI: https://doi.org/10.1145/3117811.311782

[26] Libby Ferland, Thomas Huffstutler, Jacob Rice, Joan Zheng, Shi Ni, and Maria Gini. 2019. Evaluating older users' experiences with commercial dialogue systems: Implications for future design and development. In *the 2nd AAAI Workshop on Reasoning and Learning for Human-Machine Dialogues (DEEP-DIAL)*. AAAI Press, Washington, DC. DOI: https://doi.org/10.48550/arXiv.1902.04393

[27] Libby Ferland and Wilma Koutstaal. 2020. How's your day look? The (un)expected sociolinguistic effects of user modeling in a conversational agent. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. ACM, New York, NY, 1–8. DOI: https://doi.org/10.1145/3334480.3375227

[28] Libby Ferland, Jude Sauve, Michael Lucke, Runpeng Nie, Malik Khadar, Serguei Pakhomov, and Maria Gini. 2021. Tell me about your day: Designing a conversational agent for time and stress management. In *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability*. David L. Buckeridge, Arash Shaban-Nejad, and Martin Michalowski (Eds.), Springer, Cham, Palo Alto, California, 297–303. DOI: https://doi.org/10.1007/978-3-030-53352-6_28

[29] Björn Fischer, Alexander Peine, and Britt Östlund. 2020. The importance of user involvement: A systematic review of involving older users in technology design. *The Gerontologist* 60, 7 (2020), e513–e523. DOI: https://doi.org/10.1093/geront/gnz163

[30] Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, et al. 2021. Future directions for chatbot research: An interdisciplinary research agenda. *Computing* 103, 12 (2021), 2915–2942. DOI: https://doi.org/10.1007/s00607-021-01016-7

[31] Kate Freire, Rod Pope, Kate Jeffrey, Kristen Andrews, Melissa Nott, and Tricia Bowman. 2022. Engaging with children and adolescents: A systematic review of participatory methods and approaches in research informing the development of health resources and interventions. *Adolescent Research Review* 7, 3 (2022), 335–354. DOI: https://doi.org/10.1007/s40894-022-00181-w

[32] Alisa Frik, Leysan Nurgalieva, Julia Bernd, Joyce Lee, Florian Schaub, and Serge Egelman. 2019. Privacy and security threat models and mitigation strategies of older adults. In *15th Symposium on Usable Privacy and Security (SOUPS '19)*. USENIX Association, Santa Clara, CA, 21–40. DOI: https://doi.org/10.5555/3361476.3361479

[33] Russell Fulmer, Angela Joerin, Breanna Gentile, Lysanne Lakerink, and Michiel Rauws. 2018. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. *JMIR Mental Health* 5, 4 (2018), e9782. DOI: https://doi.org/10.2196/mental.9782

[34] Shelly L. Gable and Harry T. Reis. 2010. Good news! Capitalizing on positive events in an interpersonal context. In *Advances in Experimental Social Psychology*. Mark P. Zanna (Ed.), Vol. 42, 195–257. DOI: https://doi.org/10.1016/S0065-2601(10)42004-3

[35] Shelly L. Gable, Harry T. Reis, Emily A. Impett, and Evan R. Asher. 2018. What do you do when things go right? The intrapersonal and interpersonal benefits of sharing positive events. In *Relationships, Well-Being and Behaviour: Selected Works of Harry Reis* (1st ed.), Harry T. Reis (Ed.), Routledge, 144–182. DOI: https://doi.org/10.4324/9780203732496-6

[36] Kallirroi Georgila, Maria Wolters, Vasilis Karaiskos, Melissa Kronenthal, Robert Logie, Neil Mayo, Johanna Moore, and Matt Watson. 2008. A fully anotated corpus for studying the effect of cognitive ageing on users' interactions with spoken dialogue systems. In *6th International Conference on Language Resources and Evaluation (LREC '08)*. Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias (Eds.), European Language Resources Association (ELRA), 7 pages.

[37] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2019. EMMA: An emotion-aware wellbeing chatbot. *In 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE*, 1–7. DOI: https://doi.org/10.1109/ACII.2019.8925455

[38] Jean-Philippe Gouin, Carsten Wrosch, Jennifer J. McGrath, and Linda Booij. 2020. Interpersonal capitalization moderates the associations of chronic caregiving stress and depression with inflammation. *Psychoneuroendocrinology* 112 (2020), 104509. DOI: https://doi.org/10.1016/j.psyneuen.2019.104509

[39] Miriam G. Grates, Ann-Christin Heming, Marina Vukoman, Peter Schabsky, and Jonas Sorgalla. 2019. New perspectives on user participation in technology design processes: An interdisciplinary approach. *The Gerontologist* 59, 1 (2019), 45–57. DOI: https://doi.org/10.1093/geront/gny112

[40] Stephanie Greer, Danielle Ramo, Yin-Juei Chang, Michael Fu, Judith Moskowitz, and Jana Haritatos. 2019. Use of the chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: Randomized controlled feasibility trial. *JMIR MHealth and UHealth* 7, 10 (2019), e15018. DOI: https://doi.org/10.2196/15018

[41] Camilla Gudmundsen Høiland, Asbjørn Følstad, and Amela Karahasanovic. 2020. Hi, can I help? Exploring how to design a mental health chatbot for youths. *Human Technology* 16, 2 (2020). DOI: https://doi.org/10.17011/ht/urn.202008245640

[42] Jeffrey Hall, Natalie Pennington, and Amanda Holmstrom. 2021. Connecting through technology during COVID-19. *Human Communication & Technology* 2, 1 (2021), 1–18. DOI: https://doi.org/10.17161/hct.v3i1.15026

[43] Rachel Hershenberg, Joanne Davila, and Shirley H. Leong. 2014. Depressive symptoms in women and the preference and emotional benefits of discussing positive life events. *Journal of Social and Clinical Psychology* 33, 9 (2014), 767–788. DOI: https://doi.org/10.1521/jscp.2014.33.9.767

[44] Annabell Ho, Jeff Hancock, and Adam S. Miner. 2018. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication* 68, 4 (2018), 712–733. DOI: https://doi.org/10.1093/joc/jqy026

[45] C. J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *In 8th International Conference on Weblogs and Social Media (ICWSM)*. AAAI, Palo Alto, California. DOI: https://doi.org/10.1609/icwsm.v8i1.14550

[46] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR MHealth and UHealth* 6, 11 (2018), e12106. DOI: https://doi.org/10.2196/12106

[47] Tomoko Ito, Mikiya Sato, Hideto Takahashi, Chihiro Omori, Yuta Taniguchi, Xueying Jin, Taeko Watanabe, Haruko Noguchi, and Nanako Tamiya. 2022. Mortality differences in disabled older adults by place of care in Japan: Nationwide 10-year results. *Journal of Public Health Policy* 43, 4 (2022), 542–559. DOI: https://doi.org/10.1057/s41271-022-00369-3

[48] Dilip V. Jeste, Colin A. Depp, and Ipsit V. Vahia. 2010. Successful cognitive and emotional aging. *World Psychiatry* 9, 2 (2010), 78. DOI: https://doi.org/10.1002/j.2051-5545.2010.tb00277.x

[49] Astrid Karnoe, Dorthe Furstrand, Karl Bang Christensen, Ole Norgaard, and Lars Kayser. 2018. Assessing competencies needed to engage with digital health services: Development of the eHealth literacy assessment toolkit. *Journal of Medical Internet Research* 20, 5 (2018), e178. DOI: https://doi.org/10.2196/jmir.8347

[50] Bumsoo Kim and Yonghwan Kim. 2017. College students' social media use and communication network heterogeneity: Implications for social capital and subjective well-being. *Computers in Human Behavior* 73 (2017), 620–628. DOI: https://doi.org/10.1016/j.chb.2017.03.033

[51] Youjeong Kim and S. Shyam Sundar. 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior* 28, 1 (2012), 241–250. DOI: https://doi.org/10.1016/j.chb.2011.09.006

[52] Nathaniel M. Lambert, A. Marlea Gwinn, Roy F. Baumeister, Amy Strachman, Isaac J. Washburn, Shelly L. Gable, and Frank D. Fincham. 2013. A boost of positive affect: The perks of sharing positive experiences. *Journal of Social and Personal Relationships* 30, 1 (2013), 24–43. DOI: https://doi.org/10.1177/0265407512449400

[53] Christopher A. Langston. 1994. Capitalizing on and coping with daily-life events: Expressive responses to positive events. *Journal of Personality and Social Psychology* 67, 6 (1994), 1112. DOI: https://doi.org/10.1037/0022-3514.67.6.1112

[54] Brian Leavy, Brenda H. O'Connell, and Deirdre O'Shea. 2023. Gratitude, affect balance, and stress buffering: A growth curve examination of cardiovascular responses to a laboratory stress task. *International Journal of Psychophysiology* 183 (2023), 103–116. DOI: https://doi.org/10.1016/j.ijpsycho.2022.11.013

[55] Minha Lee, Jessica Contreras Alejandro, and Wijnand IJsselsteijn. 2024. Cultivating gratitude with a chatbot. *International Journal of Human–Computer Interaction* 40, 18 (2024), 4957–4972. DOI: https://doi.org/10.1080/10447318.2023.2231277

[56] Y. Irina Li, Lisa R. Starr, and Rachel Hershenberg. 2017. Responses to positive affect in daily life: Positive rumination and dampening moderate the association between daily events and depressive symptoms. *Journal of Psychopathology and Behavioral Assessment* 39 (2017), 412–425. DOI: https://doi.org/10.1007/s10862-017-9593-y

[57] Ziyang Li, Pei-Luen Patrick Rau, and Dinglong Huang. 2019. Self-disclosure to an IoT conversational agent: Effects of space and user context on users' willingness to self-disclose personal information. *Applied Sciences* 9, 9, Article 1887 (2019). DOI: https://doi.org/10.3390/app9091887

[58] Yuting Liao, Jessica Vitak, Priya Kumar, Michael Zimmer, and Katherine Kritikos. 2019. Understanding the role of privacy and trust in intelligent personal assistant adoption. In *14th International Conference on Information in Contemporary Society (iConference '19)*. Springer, 102–113. DOI: https://doi.org/10.1007/978-3-030-15742-5_9

[59] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv:1907.11692. DOI: https://doi.org/10.48550/arXiv.1907.11692

[60] Irene Lopatovska, Olivia Turpin, Jessika Davis, Ellen Connell, Chris Denney, Hilda Fournier, Archana Ravi, Ji Hee Yoon, and Eesha Parasnis. 2022. Capturing teens' voice in designing supportive agents. In *4th Conference on Conversational User Interfaces. ACM*, New York, NY, 1–12. DOI: https://doi.org/10.1145/3543829.3543838

[61] Nicola Ludin, Chester Holt-Quick, Sarah Hopkins, Karolina Stasiak, Sarah Hetrick, Jim Warren, and Tania Cargo. 2022. A chatbot to support young people during the COVID-19 pandemic in New Zealand: Evaluation of the real-world rollout of an open trial. *Journal of Medical Internet Research* 24, 11 (2022), e38743. DOI: https://doi.org/10.2196/38743

[62] Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: The gulf between user expectation and experience of conversational agents. In *2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 5286–5297. DOI: https://doi.org/10.1145/2858036.2858288

[63] George M. Marakas, Richard D. Johnson, and Jonathan W. Palmer. 2000. A theoretical model of differential social attributions toward computing technology: When the metaphor becomes the model. *International Journal of Human-Computer Studies* 52, 4 (2000), 719–750. DOI: https://doi.org/10.1006/ijhc.1999.0348

[64] Karen Dorman Marek, Lori Popejoy, Greg Petroski, David Mehr, Marilyn Rantz, and Wen-Chieh Lin. 2005. Clinical outcomes of aging in place. *Nursing Research* 54, 3 (2005), 202–211. DOI: https://doi.org/10.1097/00006199-200505000-00008

[65] Karen Dorman Marek, Frank Stetzer, Scott J. Adams, Lori L. Popejoy, and Marilyn Rantz. 2012. Aging in place versus nursing home care: Comparison of costs to Medicare and Medicaid. *Research in Gerontological Nursing* 5, 2 (2012), 123–129. DOI: https://doi.org/10.3928/19404921-20110802-01

[66] Audrey Mariamo, Caroline Elizabeth Temcheff, Pierre-Majorique Léger, Sylvain Senecal, and Marianne Alexandra Lau. 2021. Emotional reactions and likelihood of response to questions designed for a mental health chatbot among adolescents: Experimental study. *JMIR Human Factors* 8, 1 (2021), e24343. DOI: https://doi.org/10.2196/24343

[67] Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. 2000. Maximum entropy Markov models for information extraction and segmentation. In *17th International Conference on Machine Learning (ICML)*. ACM, New York, NY, 591–598. DOI: https://doi.org/10.5555/645529.658277

[68] Graeme McLean and Kofi Osei-Frimpong. 2019. Hey Alexa… Examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior* 99 (2019), 28–37. DOI: https://doi.org/10.1016/j.chb.2019.05.009

[69] Microsoft. 2023. *Azure Customer Data Protection*. Microsoft. Retrieved December 19, 2023 from https://learn.microsoft.com/en-us/azure/security/fundamentals/protection-customer-data

[70] Microsoft. 2023. *Data Management at Microsoft*. Microsoft. Retrieved December 19, 2023 from https://www.microsoft.com/en-us/trust-center/privacy/data-management

[71] Harry R. Moody and Jennifer R. Sasser. 2020. *Aging: Concepts and Controversies*. Sage Publications, Thousand Oaks, CA.

[72] Isabela Motta and Manuela Quaresma. 2021. Understanding task differences to leverage the usability and adoption of voice assistants (VAs). In *International Conference on Human-Computer Interaction*. Springer, 483–502. DOI: https://doi.org/10.1007/978-3-030-78227-6_35

[73] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 56, 1 (2000), 81–103. DOI: https://doi.org/10.1111/0022-4537.00153

[74] Nataliya Nerobkova, Yu Shin Park, Eun-Cheol Park, and Jaeyong Shin. 2023. Frailty transition and depression among community-dwelling older adults: The Korean Longitudinal Study of Aging (2006–2020). *BMC Geriatrics* 23, 1 (2023), 148. DOI: https://doi.org/10.1186/s12877-022-03570-x

[75] Andreia Nunes, São Luís Castro, and Teresa Limpo. 2020. A review of mindfulness-based apps for children. *Mindfulness* 11 (2020), 2089–2101. DOI: https://doi.org/10.1007/s12671-020-01410-w

[76] Bruna Oewel, Tawfiq Ammari, and Robin N. Brewer. 2023. Voice assistant use in long-term care. In *5th International Conference on Conversational User Interfaces*. ACM, New York, NY, 1–10. DOI: https://doi.org/10.1145/3571884.3597135

[77] University of Minnesota. 2025. *Rules for Storing Protected Health Information in Box*. University of Minnesota. Retrieved January 15, 2025 from https://box.umn.edu/static/phi_rules

[78] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015*. Technical Report. University of Texas at Austin.

[79] Brett J. Peters, Harry T. Reis, and Shelly L. Gable. 2018. Making the good even better: A review and theoretical model of interpersonal capitalization. *Social and Personality Psychology Compass* 12, 7 (2018), e12407. DOI: https://doi.org/10.1111/spc3.12407

[80] Noemi da Paixao Pinto, Juliana Baptista dos Santos Franca, Henrique Prado de Sa Sousa, Adriana Santarosa Vivacqua, and Ana Cristina Bicharra Garcia. 2021. Conversational agents for elderly interaction. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, New York City, 1–6. DOI: https://doi.org/10.1109/CSCWD49262.2021.9437883

[81] Arianna Poli, Susanne Kelfve, and Andreas Motel-Klingebiel. 2019. A research tool for measuring non-participation of older people in research on digital health. *BMC Public Health* 19, 1 (2019), 1–12. DOI: https://doi.org/10.1186/s12889-019-7830-x

[82] 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* 27, 4 (2020), 1–27. DOI: https://doi.org/10.1145/3373759

[83] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility came by accident": Use of voice-controlled intelligent personal assistants by people with disabilities. In *2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–13. DOI: https://doi.org/10.1145/3173574.3174033

[84] Sandeep Purao, Haijing Hao, and Chenhang Meng. 2021. The use of smart home speakers by the elderly: Exploratory analyses and potential for big data. *Big Data Research* 25 (2021), 100224. DOI: https://doi.org/10.1016/j.bdr.2021.100224

[85] Arushi Raghuvanshi, Lucien Carroll, and Karthik Raghunathan. 2018. Developing production-level conversational interfaces with shallow semantic parsing. In *2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Brussels, Belgium, 157–162. DOI: https://doi.org/10.18653/v1/D18-2027

[86] Encarnación Ramírez, Ana Raquel Ortega, Alberto Chamorro, and José María Colmenero. 2014. A program of positive intervention in the elderly: Memories, gratitude and forgiveness. *Aging & Mental Health* 18, 4 (2014), 463–470. DOI: https://doi.org/10.1080/13607863.2013.856858

[87] Harry T. Reis, Shannon M. Smith, Cheryl L. Carmichael, Peter A. Caprariello, Fen-Fang Tsai, Amy Rodrigues, and Michael R. Maniaci. 2010. Are you happy for me? How sharing positive events with others provides personal and interpersonal benefits. *Journal of Personality and Social Psychology* 99, 2 (2010), 311. DOI: https://doi.org/10.1037/a0018344

[88] Noralou P. Roos and Betty Havens. 1991. Predictors of successful aging: A twelve-year study of Manitoba elderly. *American Journal of Public Health* 81, 1 (1991), 63–68. DOI: https://doi.org/10.2105/AJPH.81.1.63

[89] Jeff Sauro. 2010. *A Practical Guide to Measuring Usability*, Vol. 12. Measuring Usability LLC, Denver.

[90] Stephen M. Schueller and Acacia C. Parks. 2012. Disseminating self-help: Positive psychology exercises in an online trial. *Journal of Medical Internet Research* 14, 3 (2012), e1850. DOI: https://doi.org/10.2196/jmir.1850

[91] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, what's up?" A mixed-methods studies of in-home conversational agent usage. In *2018 Designing Interactive Systems Conference*. ACM, New York, NY, 857–868. DOI: https://doi.org/10.1145/3196709.3196772

[92] Jennifer L. Smith and Linda Hollinger-Smith. 2015. Savoring, resilience, and psychological well-being in older adults. *Aging & Mental Health* 19, 3 (2015), 192–200. DOI: https://doi.org/10.1080/13607863.2014.986647

[93] Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54.

[94] Barbra Teater and Jill M. Chonody. 2020. How do older adults define successful aging? A scoping review. *The International Journal of Aging and Human Development* 91, 4 (2020), 599–625. DOI: https://doi.org/10.1177/0091415019871207
[95] Biju Thankachan, Markku Turunen, and Kristiina Jokinen. 2023. Challenges with voice assistants for the elderly in semi-public spaces. In *16th International Conference on Pervasive Technologies Related to Assistive Environments*. ACM, New York, NY, 658–661. DOI: https://doi.org/10.1145/3594806.3596576
[96] Milka Trajkova and Aqueasha Martin-Hammond. 2020. "Alexa is a toy": Exploring older adults' reasons for using, limiting, and abandoning echo. In *2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–13. DOI: https://doi.org/10.1145/3313831.3376760
[97] Ravichander Vipperla, Maria Wolters, Kallirroi Georgila, and Steve Renals. 2009. Speech input from older users in smart environments: Challenges and perspectives. In *5th International Conference on Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments (UAHCI)*. Springer, Berlin, 117–126. DOI: https://doi.org/10.1007/978-3-642-02710-9_14
[98] Robert S. Wilson, Kristin R. Krueger, Steven E. Arnold, Julie A. Schneider, Jeremiah F. Kelly, Lisa L. Barnes, Yuxiao Tang, and David A. Bennett. 2007. Loneliness and risk of Alzheimer disease. *Archives of General Psychiatry* 64, 2 (2007), 234–240. DOI: https://doi.org/10.1001/archpsyc.64.2.234
[99] Qiang Yang, Kaiyan Cui, and Yuanqing Zheng. 2024. Room-scale voice liveness detection for smart devices. *IEEE Transactions on Dependable and Secure Computing*21, 5 (2024), 4982–4996. DOI: https://doi.org/10.1109/TDSC.2024.3367269
[100] Xiuyu Yao, Yidan Wang, Ying Zhou, and Zheng Li. 2023. A nurse-led positive psychological intervention among elderly community-dwelling adults with mild cognitive impairment and depression: A non-randomized controlled trial. *International Journal of Geriatric Psychiatry* 38, 6 (2023), e5951. DOI: https://doi.org/10.1002/gps.5951
[101] Svetlana Yarosh and Stephen Matthew Schueller. 2017. "Happiness inventors": Informing positive computing technologies through participatory design with children. *Journal of Medical Internet Research* 19, 1 (2017), e14. DOI: https://doi.org/10.2196/jmir.6822
[102] Yue You, Chun-Hua Tsai, Yao Li, Fenglong Ma, Christopher Heron, and Xinning Gui. 2023. Beyond self-diagnosis: How a chatbot-based symptom checker should respond. *ACM Transactions on Computer-Human Interaction* 30, 4 (2023), 1–44. DOI: https://doi.org/10.1145/3589959

Appendices

A Experience with Technology Items (Phases II and III)

Participants were asked to rate their experiences with various uses of Internet-enabled technology. Responses were on a 5-point Likert scale, anchored at 1 for "very poor" and 5 for "excellent." The specific items included:

(1) Sending/receiving e-mails
(2) Buying goods or services over the Internet
(3) Reading or downloading online news, newspaper, or magazines
(4) Internet banking
(5) Accessing institutions
(6) Playing or downloading games, images, films, or music
(7) Listening to Web radio or watching Web television
(8) Telephoning or making video calls over the Internet
(9) Social networking, for example Facebook or X (formerly known as Twitter), and
(10) Posting messages to chat sites, blogs or forums, or instant messaging.

B SUS Items (Phases II and III)

There are 10 statements on the SUS, each rated on a 5-point Likert scale. The scale is anchored at 1 for "strongly disagree" and 5 for "strongly agree." These statements include:

(1) I think that I would like to use this system frequently.
(2) I found the system unnecessarily complex.
(3) I thought the system was easy to use.
(4) I think that I would need the support of a technical person to be able to use this system.

(5) I found the various functions in this system were well integrated.
(6) I thought there was too much inconsistency in this system.
(7) I would imagine that most people would learn to use this system very quickly.
(8) I found the system very cumbersome to use.
(9) I felt very confident using the system.
(10) I needed to learn a lot of things before I could get going with this system.

C Post-Session Ratings of CA Usability Items (Phases II and III)

At the end of each experimental session, participants were asked 15 CA-specific questions, listed here. All questions were rated on a 5-point Likert scale, anchored at 1 for "completely disagree" and 5 for "completely agree."

(1) The CA was pleasant to be with.
(2) The CA was sociable with me.
(3) I felt comfortable sharing personal information during the interaction.
(4) I felt like I could be open during the interaction.
(5) I felt involved in the interaction.
(6) I enjoyed the interaction with the CA.
(7) I considered the interaction with the CA to be smooth.
(8) I would like to interact with the CA again.
(9) I feel the interaction with the CA was satisfying.
(10) The CA said the right thing to make me feel better.
(11) The CA responded appropriately to my feelings and emotions.
(12) The CA communicated correctly.
(13) The CA came across as competent.
(14) The CA came across as natural.
(15) Overall, I would like to have a long conversation with the CA.

D Participant I&E with Technology Statistics (Phase III)

D.1 Interest

Table D1 contains question-wise statistics for the interest in technology questionnaire administered at the beginning of the experiment for Phase III.

D.2 Experience

Table D2 contains detailed statistics for participants' self-reported skill level with various technologies in Phase III.

Table D1. Participant Interests in Technology for Phase III

Item	Mean	SD	Min.	Max.
Interested in computers or tablets	4.31	1.12	1	5
Fond of computers/tablets	4.08	1.26	1	5
Not afraid to try new functions on computer/tablet	3.96	1.43	1	5
Frequency of use of computers/tablets	4.73	0.78	2	5
Interested in using smartphone	4.50	1.11	1	5
Fond of smartphone	4.31	1.16	1	5
Not afraid to try new functions on smartphone	3.96	1.34	1	5
Frequency of use of smartphone	4.58	1.21	1	5

Statements were evaluated using a 5-point Likert scale, anchored at 1 for "not at all" and 5 for "extremely."

Table D2. Participant Technology Skills for Phase III

Item	Mean	SD	Min.	Max.
Able to type on keyboard	4.77	0.43	4	5
Able to use mouse	4.96	0.20	4	5
Able to load ink into printer	4.77	0.43	4	5
Able to fix paper jams	4.46	0.71	3	5
Able to open e-mail	4.88	0.43	3	5
Able to send e-mail	5.00	0.00	5	5
Able to find information about local community resources	4.77	0.51	3	5
Able to find information about hobbies/interests	4.85	0.37	4	5
Able to use computer to enter events into calendar	4.04	1.46	1	5
Able to check the date and time of upcoming or prior appointments	4.31	1.29	1	5
Able to use computer to watch movies/videos	4.31	1.19	1	5
Able to use computer to listen to music	4.12	1.31	1	5

Statements were evaluated using a 5-point Likert scale, anchored at 1 for "not at all" and 5 for "extremely."

Received 30 December 2023; revised 18 January 2025; accepted 10 March 2025