

Promoting Cognitive Health in Elder Care with Large Language Model-Powered Socially Assistive Robots

Maria R. Lima
Mechanical Engineering
Imperial College London
London, United Kingdom
Care Research and Technology Centre
UK Dementia Research Institute
London, United Kingdom
maria.lima18@imperial.ac.uk

Amy O'Connell
Computer Science
University of Southern California
Los Angeles, California, USA
ao71627@usc.edu

Feiyang Zhou
Computer Science
University of Southern California
Los Angeles, California, USA
zhoufeiyang@alumni.usc.edu

Alethea Nagahara
Computer Science
University of Southern California
Los Angeles, California, USA
nagahara@usc.edu

Avni Hulyalkar
Computer Science
University of Southern California
Los Angeles, California, USA
hulyalka@usc.edu

Anura Deshpande
Computer Science
University of Southern California
Los Angeles, California, USA
anurades@usc.edu

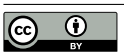
Jesse Thomason
Computer Science
University of Southern California
Los Angeles, California, USA
jessetho@usc.edu

Ravi Vaidyanathan
Mechanical Engineering
Imperial College London
London, United Kingdom
Care Research and Technology Centre
UK Dementia Research Institute
London, United Kingdom
r.vaidyanathan@imperial.ac.uk

Maja Matarić
Computer Science
University of Southern California
Los Angeles, California, USA
mataric@usc.edu

Abstract

As the global population ages, there is increasing need for accessible technologies that promote cognitive health and detect early signs of cognitive decline. This research demonstrates the potential for in-residence monitoring and assessment of cognitive health using large language model (LLM)-powered socially assistive robots (SARs). We conducted a 5-week within-subjects study involving 22 older adults in retirement homes to investigate the feasibility of LLM-powered SARs for promoting and assessing cognitive health. We designed tasks that involved verbal dialogue based on clinically validated cognitive tools. Our findings reveal improved task performance after three robot-administered sessions, with significantly more detailed picture descriptions, fewer word repetitions in semantic fluency, and reduced need for hints. We found that older adults were more socially engaged in robot-administered tasks compared to those administered by a human, and they accepted and were willing to engage with SARs in this context, which had not been tested before.



This work is licensed under a Creative Commons Attribution 4.0 International License.

CHI '25, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713582

CCS Concepts

• **Human-centered computing** → **User studies**; *Usability testing*; *Interaction design*; *Natural language interfaces.*

Keywords

socially assistive robotics, large language models, cognitive health, elder care

ACM Reference Format:
Maria R. Lima, Amy O'Connell, Feiyang Zhou, Alethea Nagahara, Avni Hulyalkar, Anura Deshpande, Jesse Thomason, Ravi Vaidyanathan, and Maja Matarić. 2025. Promoting Cognitive Health in Elder Care with Large Language Model-Powered Socially Assistive Robots. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3706598.3713582

1 Introduction

The world's aging population is set to double by 2050 [60]. As a result, the number of people living with dementia (PLWD) is rapidly increasing, with 10 million new cases identified each year [88]. Dementia is a public health priority as the leading cause of death and dependency among older adults worldwide [88]. As the demand for physical and cognitive support for older adults increases, the growing shortage of long-term care workers is becoming apparent [54]. This shortage places an enormous psychological and financial burden on families and caregivers who support PLWD in daily activities [9]. The dramatic demographic shift toward an aging

population underscores a pressing need for accessible technologies that can promote cognitive health, support independence at home, and detect early signs of cognitive deterioration, to facilitate timely and comprehensive evaluation by medical professionals.

There is a clear need for accessible home-based tools to promote and monitor cognitive health in elder care. Neuropsychological tests (NPT) conducted through in-person interviews, such as the Mini-Mental State Examination (MMSE) [21], remain the primary method for assessing cognitive functions, including attention, memory, language, and visuospatial abilities. However, NPT are limited by clinician availability, are often qualitative in nature, and are susceptible to errors and high inter-rater variability. Furthermore, traditional NPT, typically administered in paper-pencil format, can be effortful, repetitive, and disengaging.

SARs and conversational artificial intelligence (AI) are recognized for their potential to support aging and dementia care [25, 47, 53]. SARs interact through social and emotional cues, combining physical embodiment, gestures, facial expressions, and speech to improve user engagement and rapport compared to disembodied agents [17, 20]. Extensive past research has demonstrated SARs' utility in providing home assistance, including medication reminders [8], cognitive stimulation [83], remote therapy [48], companionship [2], and daily living support [74], while reducing the caregiver burden [12]. Voice interaction, in particular, offers an intuitive, hands-free, and natural user interface, promoting usability among older adults [38] and providing safety in scenarios where voice may be the only form of communication [45]. When integrated into home environments, conversational AI can offer proactive verbal support by responding to safety risks, such as falls or agitation, querying users to verify health symptoms, and initiating alerts in emergencies [46, 68].

Furthermore, home-based SARs could provide an accessible user interface for cognitive assessment as a side effect of regular human-robot interaction (HRI). Recent studies have suggested that collecting speech patterns via mobile applications and voice assistants to detect Alzheimer's disease (AD) and mild cognitive impairment (MCI) are feasible, often using memory-probing questions that mimic outpatient consultations [22, 58, 89]. However, there is a significant gap in research evaluating SARs in real-world settings, particularly as in-residence tools for cognitive support in elder care.

Moreover, in-home deployment of SARs that record speech raises ethical privacy concerns. These concerns are prevalent among older populations and represent a barrier to technology adoption [38]. If not carefully designed with end-user involvement, SARs may be perceived as intrusive. Therefore, it is essential to conduct user studies with stakeholders from the earliest HRI design stages to understand their perceptions and concerns toward facilitating long-term acceptance, trust, and perceived utility of SARs.

As conversational AI through the use of LLMs continues to change the landscape of contextual dialogue, integration with SARs opens new opportunities for human-robot engagement [91]. However, the ethical implications of deploying these technologies, especially among older people with cognitive care needs, must be carefully considered to ensure they contribute positively to their well-being and independence.

To investigate the feasibility of using LLM-powered SARs to promote and assess cognitive health, we conducted a 5-week within-subjects study in two retirement homes with 22 older adults, encompassing individuals with normal aging and those with cognitive impairment. We designed tasks that involved verbal dialogue based on clinically validated cognitive tools [11, 26, 36], including picture descriptions and semantic fluency. To minimize task memorization by older adult participants, we alternated two variations of each task between five weekly study sessions and established an individual baseline through an initial human-administered session. Participants then engaged in three weekly robot-administered sessions to mitigate the novelty effect, which leads to initial engagement but quickly wears off, resulting in decreased interest and changing attitudes toward robots [44].

Our results show that: 1) older adults accept and are willing to engage with SARs in the context of promoting cognitive health, which had not been tested before; 2) older adults' task performance significantly improved after the weekly robot-administered sessions, with more detailed picture descriptions, fewer word repetitions in semantic fluency, and a reduced need for hints by the administrator to complete the tasks; 3) older adults were more socially engaged in verbal cognitive tasks when administered by the robot compared to those administered by a human; and therefore 4) there is potential for monitoring and assessing cognitive health using LLM-powered SARs at home or in other residences.

This work contributes novel insights into the efficacy of LLM-powered SARs in promoting cognitive health in elder care, as well as the potential for SARs to provide accessible in-residence cognitive support. With consistent use, SARs could monitor cognitive health over time and identify individuals requiring more comprehensive diagnostic evaluation.

2 Related Work

In this section, we introduce prior work on SARs to support well-being and independence in elder care. Next, we review previous studies that have begun to explore SARs in cognitive assessment scenarios. We then explore recent work on LLMs integrated with SARs for more natural and engaging interactions. Finally, we identify the research gaps this work aims to address.

2.1 SARs for Well-being of Older Adults

Previous studies have reviewed the effectiveness of SARs in supporting the well-being and independence of older adults [1, 25, 47, 67, 86]. Particularly, SARs have been validated for use in elder care contexts with potential to offer social interaction and companionship [2], mental health support [48] assistance with daily living [72], health monitoring [57], and cognitive support [83]. Furthermore, prior work reports on their ability to reduce stress and anxiety among PLWD [63].

SARs offer multimodal interactive features beyond speech, such as facial expressions, embodiment, and gestures. Their embodiment and social presence enhances engagement and enjoyment in HRI [17, 37, 87]. For instance, Heerink et al. [27] reported social presence as a key feature in the design of SARs for older people. Nishio et al. [55] showed older people were more engaged in dialogues with a physically embodied robot than a virtual one.

Interest in and acceptance of SARs among older adults is increasing [40]. Prior research has highlighted psychosocial support as a key support function to promote acceptance of SARs among older populations, both as assistive tools and companions [5]. However, there is a lack of longitudinal studies due to limited user-friendly, accessible, and reliable SAR platforms capable of operating autonomously and engaging users in long-term in-home deployments. In this study, we aim to address such gap by evaluating a SAR for in-residence cognitive support of older adults.

2.2 SARs for Cognitive Assessment

Few studies have investigated the use of SARs in cognitive health assessment scenarios, particularly among older populations and individuals with cognitive impairments. Chang et al. [14] investigated the feasibility of detecting MCI using SARs in a between-subjects study with HRI-based memory tasks and standardized NPT [14]. Similarly, Rossi et al. [73] assessed personality traits via SAR-administered psychometric questionnaires and Pou-Prom et al. [64] evaluated a SAR platform for cognitive assessment with 19 seniors diagnosed with AD in picture description tasks, comparing interactions with a human administrator. That study focused on miscommunication and linguistic analysis from manual annotations, finding that richer HRI occurred with individuals with milder cognitive impairment. Additionally, participants had longer interactions and used a broader vocabulary with the human administrator compared to the robot. The authors also identified significant variations in word and sequence length across different levels of cognitive impairment, which were consistent in interactions with both human and robot administrators.

While these studies suggest the potential of SARs for cognitive support, further user studies are needed in real-world settings with larger and more diverse cohorts. Moreover, challenges related to speech recognition and HRI dialogue flexibility have not been fully addressed, particularly within elder care. Our study aims to address these gaps by evaluating the feasibility of an LLM-powered SAR to promote cognitive health through a 5-week, in-residence study. We recruited a diverse and multilingual cohort of older adults who completed cognitive tasks based on clinically validated tools through human-robot verbal dialogue in English or Spanish.

2.3 LLMs for Human-Robot Engagement

LLMs are transformer-based models, such as BERT [18] and OpenAI's GPT [70], trained on large datasets to perform natural language processing (NLP) tasks, including question answering, translation, summarization, and content generation [84]. Recent developments in LLMs have unlocked opportunities for open dialogue and HRI [91]. LLM-powered robots have elicited increased trust in collaborative tasks with humans due to their ability to facilitate more intuitive and natural language interactions [91]. However, the use of LLMs for engaging home-based cognitive assessment and monitoring with conversational agents has not yet been explored.

In elder care, LLM-powered SARs have the potential to provide emotional support and promote cognitive health through more effective human-robot engagement [31], although only a few studies have utilized LLMs for this purpose. For instance, Yang et al. [90] developed a phone-based LLM-powered chatbot, Talk2Care, to collect health information from older adults and generate a summary for their healthcare provider. They found that the voice agent collected more detailed information than a voice recording with less effort from the provider. This study suggested how LLMs can be useful for gathering health information from older adults in an engaging way, but the system was designed to replace provider voicemail systems and was not evaluated for health monitoring. Furthermore, Jo et al. [35] developed CareCall, an LLM-powered chatbot designed to support socially isolated individuals through weekly 2-3 minute phone conversations about the user's health. They deployed the system among adults aged 50-60 and found that the calls helped mitigate loneliness. While the system proved beneficial for emotional support, the study highlighted significant challenges in maintaining long-term engagement and supporting individual health needs due to limited personalization. The lack of multilingual LLMs is another factor that limits personalization of interactions, as most models are English-only. In this work, we evaluated an LLM-powered SAR to promote cognitive health in elder care within residential settings in a longitudinal study with both English and Spanish speakers.

LLMs have also shown potential in building rapport with users and sustain therapeutic conversations [43, 49, 51]. Early evidence suggests that LLM-based interactions may reduce symptoms of depression, anxiety, and stress [59]. For instance, a chatbot trained on domain-specific health data to engage in mental health assessments has been shown to ask more detailed questions about symptoms such as sleep disturbances than human psychiatrists [13]. Despite their increased promise, the integration of LLMs into clinical and care settings remains in its early stages.

Using LLMs in sensitive environments, such as in elder care, involves significant risks, including the potential for disinformation. Research has shown that while LLMs are effective communicators, they can generate inaccurate or biased information, often without transparency [30]. To ensure that LLM-powered SARs enhance the independence and well-being of older adults, it is essential to verify their output for accuracy and safety.

3 Methods

We aimed to answer the following research questions:

- **RQ1**: Are LLM-powered SARs viable tools for engaging older adults in interactive tasks that promote cognitive health?
- **RQ2**: How does task performance change in a cognitive task after three weekly robot-administered sessions?
- **RQ3**: Do older adults accept a SAR that administers cognitive tasks?
- **RQ4**: What are the perceptions and experiences of older adults who interact with a SAR in this context?

We conducted a 5-week within-subjects study that aimed to explore the feasibility of SARs in promoting and assessing cognitive health in elder care. In this section, we present details on the conversational SAR platform used, the interaction design, older adult participants, data collection and analysis.

3.1 Conversational SAR

We used the updated version of the inexpensive Blossom robot, an open-source, 3D-printed platform [78] with a gray crocheted

exterior, to create a simple and engaging appearance. We chose a female voice from Unreal Speech text-to-speech (TTS) API [80] that allowed us to select appropriate speech speed and pitch parameters to enhance engagement and robot speech clarity for older adults. Our system implementation consisted of speech-to-text (STT), LLM-generated responses, Blossom behavior, and TTS, as illustrated in Figure 1.

3.1.1 Study Setup. The study setup, shown in Figure 2, consisted of a Blossom robot, a Raspberry Pi 4 computer that controlled the robot's movement, an external speaker and microphone for verbal interaction, a monitor to display the picture used in one of the cognitive tasks, a webcam to video record the sessions, a laptop to run the main program that integrates the verbal tasks with robot behavior, and an eye gaze tracker.

3.1.2 Robot Behavior. We designed a set of idle motions to enhance user engagement with the robot. We implemented all idle motions as actuations for each of Blossom's four motors to manually selected goal positions, as described in previous work [56]. These include different sequences of sighing, posture sways, and gaze shifts that were designed to avoid distracting behaviors. We synchronized the length of the robot's motion sequences with its speech duration to promote natural and engaging interactions. We used an Amazon Web Services (AWS) server to securely activate robot behavior for the selected motion sequence.

3.2 Interaction Design

Here, we describe the two cognitive tasks, the LLM prompt designed for verbal dialogue in robot-administered tasks, and the moderator used as a pilot safety mechanism to check the quality of LLM-generated responses.

3.2.1 Interactive Cognitive Tasks. We implemented two interactive tasks based on clinically validated cognitive tools [11, 26, 36]:

- *Picture description.* The robot explains the activity and gives participants time to describe a picture displayed on the monitor (referred to as the first attempt). If the participant pauses for a long time (> 7 seconds) after this first attempt, the robot provides hints, guiding attention to areas of the picture that have not been mentioned, based on predefined areas of interest (detailed in Section 3.2.2).
- *Semantic fluency*. The robot asks the participant to name as many items as possible within a given category (e.g., animals or fruits, depending on the session) in one minute, following standard test procedures. After this initial minute, if pauses (> 4 seconds) are detected in the participant's speech, the robot provides hints about words not yet mentioned, based on predefined groups of animals or fruits (see Section 3.2.2).

Both tasks were designed to last up to five minutes, at which point the robot ended the interaction. The duration of each task was based on the typical duration of the task in a clinical setting, with additional time for interactive hints. We selected brief tasks to minimize participant fatigue. The robot-administered picture description tasks had an average duration of 4.9 minutes across the cohort, while the semantic fluency task averaged 5.2 minutes.

We designed interactions with motivational techniques (praise, positive feedback, and encouragement) as in previous work [19]. The robot praised the user with encouraging comments in response to their picture descriptions and the words recalled during the semantic task. The first and last study sessions, administered by a researcher, adhered to the same structure, using praise and hints similar to those used by the robot.

3.2.2 LLM Prompt Design. We designed two prompts for each task that were used as input to the OpenAI's GPT-4o model to generate contextualized and engaging responses during the interactive cognitive tasks in both English and Spanish. Each prompt was designed with key elements: the role of the LLM as a SAR, the target user demographic, the interaction context, the tone, the guidelines for initiating the interaction, the structure of the generated response with praise and hints, methods to clarify misunderstandings, the groups for hints, and the criteria for recognizing the end of the task (i.e., when all groups were covered or when explicitly expressed by the user). For the picture description task, the hints included five predefined areas or actions to describe. In the semantic fluency task, hints featured six predefined animal groups or seven fruit groups. The prompts were refined iteratively through in-lab pilot testing to ensure the reliability of generated responses. The LLM prompts and parameters used for both English and Spanish sessions, and the predefined groups used for hints are provided in Appendix A.1.

3.2.3 LLM-Driven Moderator. In addition to human-in-the-loop monitoring of HRI during the tasks, we designed an AI-driven moderator program as a preventive measure to ensure the quality of LLM-generated responses in our study. The moderator reviewed each generated response along with the preceding interaction context, filtering out any potentially problematic statements before the final response was converted into speech. The moderator prompt for GPT-4o was designed to output a simple "yes" or "no" (see Appendix A.2). If the moderator detected a response outside the desired context, it first regenerated the response. If the regenerated response fell outside acceptable parameters, the system automatically terminated the interaction. While no problematic instances occurred during our in-residence HRI experiments, we implemented the moderator as a pilot safety mechanism. The moderator provided real-time feedback to the researcher overseeing the HRI, who could review the moderator's decisions for potential false positives and intervene to stop the program if necessary.

3.3 User Study in Retirement Homes

We describe the user study conducted in two retirement homes, including our cohort of participants, the within-subject study design of five weekly sessions, and the procedure followed in each session. We then detail data collection and analysis methods.

3.3.1 Participants and Setting. Our study involved 22 participants aged 65 and older (mean age = 76.2, SD = 8) who lived in retirement homes in Los Angeles, CA. In order to better represent the population of older adults in the region, we recruited participants fluent in English (64%) or Spanish (36%). All study materials, including the consent form, questionnaires, and all interactions, were adapted to the participants' preferred language. The study excluded those who were unable to consent to the study information sheet or had impairments that precluded hearing or seeing the robot. Participants (13 females, 9 males) self-reported their races and ethnicities: 14

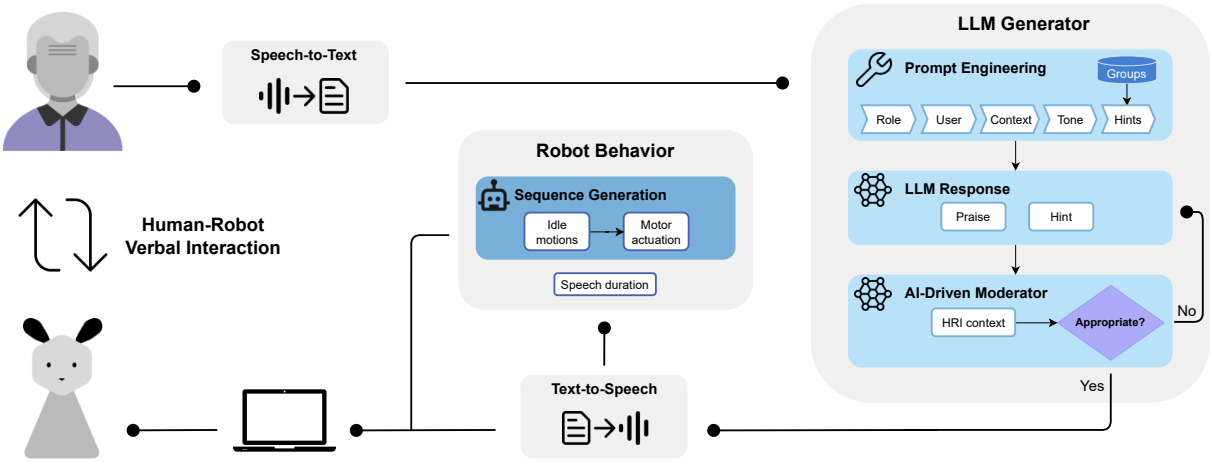


Figure 1: System implementation of an LLM-powered SAR administering cognitive tasks through verbal dialogue.



Figure 2: Experimental setup with a participant interacting with Blossom during a picture description task.

African-American, 7 white, 8 Hispanic Latino; 1 participant did not disclose. Participants completed the MMSE to assess their baseline cognitive abilities across five standardized domains: orientation, registration, attention and calculation, recall, and language (mean = 24.9, SD = 3.9). The participants could be classified into two cognitive groups: those experiencing normal aging (N = 8) and those with mild to moderate cognitive impairment (N = 14), using an MMSE cutoff of 26 as suggested in prior work [75]. Participants had varying levels of education (mean years = 10.5, SD = 4.9). Only two participants had some experience with robotics or AI, including the use of Amazon Alexa smart speaker, while the majority had no experience (N = 5) or very little experience (N = 15) with such technologies. Participants were compensated with a USD$50 gift card upon completing the study.

3.3.2 Study Design. To explore the research questions outlined in Section 3, we conducted a 5-week within-subjects study. Each session was followed by a one-week washout period. Due to health circumstances, some participants were unable to attend their sessions as scheduled. For those individuals, the average washout period was five days.

During each session, participants engaged in one picture description task and one semantic fluency task. To minimize task memorization, we alternated between two variations of each cognitive task, such that the first and last sessions utilized the same task variation, as illustrated in Figure 3. In the picture description task, participants described either the *Cookie Theft* picture from the Boston Diagnostic Aphasia Examination (BDAE) [26] or the *Picnic Scene* from the Western Aphasia Battery-Revised [36] (see Appendix A.3). The semantic fluency task, based on the Addenbrooke's Cognitive Examination [11], involved animals and fruits as categories. To ensure consistency and comparability of data, both human- and robot-administered sessions followed the same structure.

To establish an individual baseline, Session 1 was administered by a researcher. We used this baseline to address RQ2 (see Section 3) regarding task performance after three weekly robot-administered sessions. Our study included three consecutive robot-administered sessions to overcome the novelty effect, allowing participants to become familiar with Blossom. Additionally, at the end of the last robot-administered session, participants had the option to engage in an open dialogue with the robot on topics of their choice or those suggested by Blossom, which were unrelated to the cognitive tasks. This optional open dialogue had a predefined maximum duration of five minutes, after which the robot ended the interaction. The final session (Session 5) was administered by a researcher, allowing a direct comparison of performance and user perceptions of the system to those from the initial session.

3.3.3 Procedure. The study was approved by the University of Southern California Institutional Review Board (IRB UP-24-00154). Before the first session, each participant met with a researcher to review the informed consent form and receive a detailed explanation of the study. Upon providing informed consent and having the opportunity to ask questions, participants completed the Big

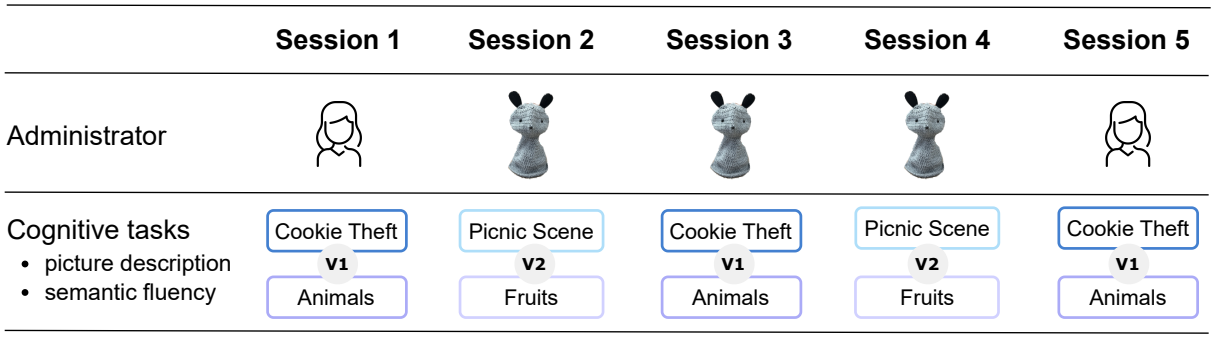


Figure 3: The within-subjects study design consisted of five sessions: two alternating versions of each cognitive task (V1 and V2), an initial session administered by the researcher used as a baseline, followed by three weekly robot sessions, and a final session conducted by the researcher.

Five Inventory (BFI) [71] and reported demographic information including age, gender, race, ethnicity, income, and highest level of education.

Each session lasted approximately 45 minutes. The two human-administered sessions included a brief introduction to the session, the two interactive cognitive tasks, and a follow-up semi-structured interview. The three robot-administered sessions included additional questionnaires, completed with the assistance of a researcher, to assess perceived system usability and robot acceptance (see Section 3.3.4). During the semi-structured interview, the researcher allowed time for open-ended responses, allowing participants to express their feedback about the session and the robot.

In the robot-administered sessions, we implemented *priming* to familiarize participants with Blossom, leveraging robot agency. Specifically, the robot introduced its interactive capabilities, drawing attention to the motor sounds and the time required for speech processing. We designed this priming phase to reduce the novelty effect [44] and prevent emotional deception [85], ensuring that participants were not interacting with the robot for the first time without prior exposure. By doing so, we aimed to create a more consistent HRI experience, allowing participants to focus on the tasks with minimal distractions due to absolute novelty of the experiment setup. We also used this pre-recorded robot priming to adjust the robot's speech volume before the tasks.

3.3.4 Data Collection. We collected quantitative data from task performance and questionnaires and qualitative data from interviews. We note that this study does not focus on the eye gaze data collected.

BFI. As part of the pre-study survey, participants completed the widely used BFI [71], which measures five personality traits: *extraversion*, *agreeableness*, *conscientiousness*, *neuroticism*, and *openness to experience*. To make the process more accessible for older adults, we administered the 10-item short version of the BFI, rated on a 5-point Likert scale.

MMSE. In the final session, we administered the MMSE to assess cognitive function across five domains: orientation, registration, attention and calculation, recall, and language. We chose the MMSE tool because it is shorter to administer than other cognitive tools (e.g., the Montreal Cognitive Assessment (MoCA) [32]) and makes the speech data collected comparable with benchmark DementiaBank datasets, which use MMSE scores as labels, enabling future predictive analysis.

SUS. After each robot-administered session, participants completed the 10-item System Usability Scale (SUS) [10], adapted for older adults and cognitively impaired users [29]. The SUS is scored out of 100 and rated on a 5-point Likert scale to evaluate system usability.

Robot Acceptance. Participants also completed a 14-item robot acceptance questionnaire tailored for SARs in elder care [28]. The questionnaire evaluates attitudes towards robots over five constructs: *sociability*, *enjoyment*, *usefulness*, *anxiety*, and *trust*, each rated on a 5-point Likert scale.

Interviews. Each session concluded with a semi-structured interview using a guide developed iteratively by the research team. The interview covered the participant's experience, comfort, enjoyment, preferences between robot and human-administered sessions, suggestions for robot improvements, and willingness to engage with the robot in the future.

Recordings. Sessions were audio- and video-recorded with consent for post-experiment analysis.

3.3.5 Data Analysis. Audio recordings of the cognitive tasks were segmented and transcribed with speech diarization techniques and OpenAI's Whisper speech recognition model [69], then verified for correctness by two researchers.

Statistical Analysis. The Shapiro-Wilk test was used to verify the normality of the dependent variables, determining if a parametric or nonparametric version of the test is appropriate. We performed paired Wilcoxon signed-rank tests to evaluate various task performance metrics and paired t-tests to compare SUS scores in the first and last study sessions. We also calculated Spearman's rank correlation to assess the relationships between task performance and MMSE scores, robot acceptance and personality traits, as well as robot acceptance and MMSE scores. We adjusted significance levels with Benjamini-Hochberg false discovery rate (FDR).

Thematic Analysis. Audio and video recordings of the interviews were transcribed complementary to on-site notes from a researcher. We conducted a thematic analysis following the steps outlined by Braun and Clarke [7] to identify recurring themes. We counted

how often each theme appeared in the transcripts to understand its prevalence, similar to the approach used in [4, 23, 42, 65, 81, 82]. First, two coders independently reviewed the interview data to familiarize themselves with its content. Following an inductive approach, the two coders independently developed a preliminary list of codes, then met to discuss their individual findings, identify emerging candidate themes, and create a unified codebook [7, 61]. The coders reviewed the interview transcripts, coding each sentence according to the unified codebook with inter-rater reliability (Cohen's $\kappa = 0.78$, indicating substantial agreement [16]). Finally, the candidate themes were refined into a structured set of themes and subthemes. The resulting themes, subthemes, and their frequencies are presented in Tables 2, 3, and 4.

3.4 Linguistic Analysis of Cognitive Tasks

We outline the features extracted from the transcripts of the picture description and semantic fluency tasks, which were used to analyze task performance over the five study sessions, addressing RQ2 (see Section 3). To further examine verbal engagement, three researchers annotated social comments made by participants beyond the tasks (e.g., sharing thoughts or personal experiences).

3.4.1 Picture Description Task. We extracted several features at the utterance level to assess:

- *Detail*. The level of detail in the first attempt (i.e., without hints, as described in Section 3.2.1), by counting the number of predefined areas/actions mentioned.
- *Words*. The total word count on the first attempt.
- *Hints*. The number of times the administrator provided hints to help the participant complete the task and elicit more details.
- *Completeness*. We used token embeddings to represent the pre-processed transcripts.[1] We compared participants' transcripts to a baseline generated by GPT-4o,[2] calculating cosine similarity between the pooled embeddings of both transcripts to assess how closely participants' descriptions were to the baseline. We hypothesized that higher similarity scores would indicate more semantically complete descriptions.
- *Use of language*: we extracted various lexical features to assess lexical diversity, the proportion of nouns, pronouns, verbs, and adjectives in participants' first attempt describing the picture (see details in Appendix B).

3.4.2 Semantic Fluency Task. We extracted the following linguistic features from the semantic task:

- *Unique correct words*. The number of unique animals or fruits named (depending on the task version; see Figure 3) within one minute during the first attempt.
- *Hints*. The number of times the administrator provided hints for additional words within the predefined categories to help the participant complete the task.
- *Repetitions*. The number of task-relevant words repeated during the entire task.

[1]We used the Multilingual-E5-large-instruct model from Hugging Face due to its superior performance on multilingual datasets compared to other models tested: OpenAI's GPT embeddings and E5-mistral-7b-instruct. The token embeddings were averaged (mean pooling) to create a single vector representation for each transcript.
[2]The GPT baseline covers all predefined picture areas and actions relevant to the task, which was verified for correctness by two researchers.

4 Results

In this section, we first report findings on task performance after robot-administered sessions (RQ2). Next, we compare social engagement in robot- and human-administered cognitive tasks. We then present results on perceived usability and robot acceptance from questionnaires (RQ3). Finally, we report the key themes from the interviews that explore older adults' perceptions and experiences interacting with a SAR for promoting cognitive health (RQ4).

4.1 Task Performance

4.1.1 Performance After Three Weekly Robot Sessions. We compared task performance on the same cognitive tasks in the first (Session 1) and last (Session 5) weekly study sessions using the paired Wilcoxon signed-rank test, considering independent linguistic features. To control for potential confounds, the same versions of the picture description and semantic fluency tasks were given by the same human administrator in Sessions 1 and 5. We used the first session as an individual baseline for how participants described the Cookie Theft picture and the number of animals mentioned in the semantic fluency task.

We found a significant increase in task performance across the cohort after the three robot-administered sessions. Specifically, in the picture description task, participants provided significantly more *detailed* descriptions in Session 5, as indicated by the number of picture areas and actions described in the first attempt, i.e., without hints ($W = 10$, $p < .01^{*}$). They also used a significantly higher number of *words* in their descriptions ($W = 18.5$, $p < .001^{***}$), and required fewer *hints* by the administrator to complete the task ($W = 0$, $p < .001^{***}$), as shown in Figure 4a. Additionally, when comparing with a GPT-generated baseline (detailed in Section 3.3.5), we found participants' verbal descriptions were semantically more *complete* in Session 5 compared to Session 1 ($W = 2.5$, $p < .001^{***}$). The analysis of lexical features revealed a significant improvement in lexical diversity across the cohort in Session 5 ($W = 33$, $p < .01$), indicating a less repetitive and more varied use of vocabulary (see Table 10 in Appendix B).

In the semantic fluency task, participants needed significantly fewer *hints* by the administrator to complete the task in Session 5 ($W = 6$, $p < .01^{**}$) and expressed fewer *repetitions* of task-relevant words ($W = 26.5$, $p < .01^{**}$), as shown in Figure 4b. Furthermore, participants recalled more *unique task-relevant words* (i.e., animals) in the last session, although this increase was not significant. Overall, task performance improvements after repeated robot-administered sessions support our primary goal of evaluating the feasibility of SARs for promoting cognitive health through verbal dialogue based on clinically validated tools.

4.1.2 Correlation with Cognitive Skills. We assessed the relationship between participants' cognitive skills, as measured by MMSE, and their performance in the interactive cognitive tasks in the last session. We calculated Spearman's rank correlation using the linguistic features selected (see Figure 4). The analysis revealed a moderate positive correlation between the number of unique task-relevant words on the semantic fluency task and MMSE scores ($r =$

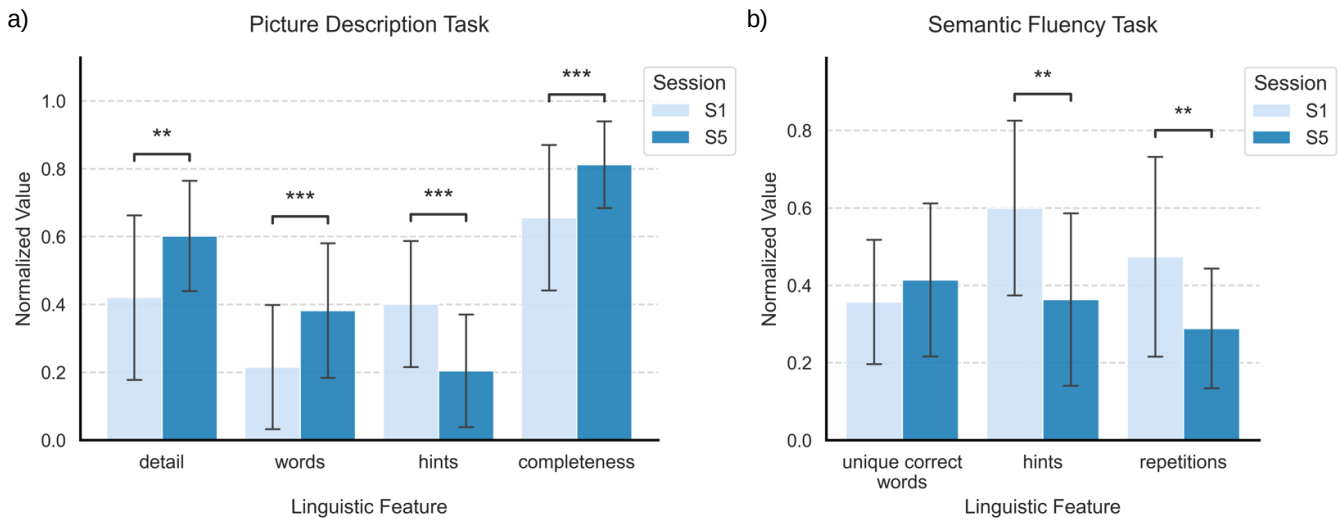


Figure 4: Task performance comparison between the first and last study sessions across the cohort for a) the picture description task; and b) the semantic fluency task.

Table 1: Demographics and number of animals named in the semantic fluency task in the first (Session 1) and last (Session 5) study sessions for the two cognitive groups — normal aging and cognitive impairment — presented as mean (standard deviation).

	Normal Aging (N = 8)	Cognitive Impairment (N = 14)
MMSE	28.8 (1)	22.7 (3)
Age	74 (5.9)	77.4 (8.9)
Sex (% male)	25%	50%
Education	11.4 (5.2)	10.1 (4.8)
Animals Session 1	17.3 (4.7)	13.5 (6.2)
Animals Session 5	18.3 (6.1)	16.1 (8)
Normative [15, 66]	18–21	8–15

0.5, $p < .05$), meaning participants with higher cognitive abilities mentioned more unique animals in one minute. Specifically, participants experiencing normal aging (MMSE > 26 [75]) named an average of 20.6 unique animals (SD = 6.6), compared to 17.4 (SD = 9.5) animals recalled by individuals with cognitive impairment (see Table 1). Notably, these values were higher than those obtained in the first study session for both cognitive groups. These findings align with normative data[3] for different age groups [15, 66], as well as recent psychology research suggesting the predictive value of verbal semantic fluency tests in screening early dementia [52]. Spearman's correlation between participants' cognitive skills and lexical features showed no significance (see Appendix B).

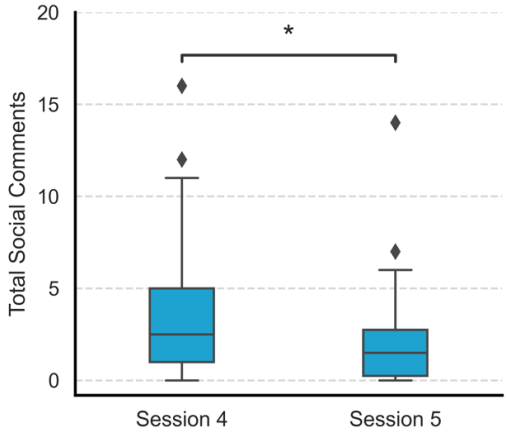


Figure 5: Number of verbal social comments during the last robot-administered (Session 4) and researcher-administered (Session 5) study sessions across the cohort.

4.2 Verbal Engagement

To eliminate absolute novelty, we compared the number of social comments participants made beyond the task (e.g., sharing thoughts or personal experiences) in the last robot and human sessions to consider overall engagement after participants became familiar with the robot. We measured social engagement per session by counting social comments in both the picture description and semantic fluency tasks. We found that social engagement was significantly higher in cognitive health-promoting tasks administered by the robot ($W = 44.5$, $p < .05^*$), as shown in Figure 5. Specifically, participants verbally engaged with Blossom beyond the task in Session 4 an average of 4.1 times (SD = 4.4) compared to 2.3 times (SD = 3.2)

[3]Note that we compare the normative data with the number of animals named in Session 1, where there was no risk of learning or memorizing the task.

in Session 5 with the researcher. For instance, a participant commented on the picture and recalled their young age while talking with Blossom, *"Oh that a... that's a beautiful scene there. I remember the time when I was young [...]"* (P7, S4), and another shared their shopping routine during the naming fruits fluency task, *"Uh...I- I can remember all them when I go to the store, I see them, I go right on to them and I pick them up when I- I go"* (P13, S2).

4.3 Perceived Usability and Robot Acceptance

We evaluated how perceived usability evolved over the three weekly robot-administered sessions. In the first robot-administered session, participants gave the robot a mean SUS score of 80.7, indicating "Good" usability. In the subsequent sessions, the SUS scores increased above 85.6, indicating "Excellent" usability [6]. We found that older adults rated the robot as increasingly easy to use over time, even after the initial novelty wore off. A paired t-test revealed a significant increase in SUS scores from the first to the last robot-administered session with a large effect size ($t(21) = -3.7$, $p < .05$, Cohen's $d = 0.8$).

Participants also completed a post-trial robot acceptance questionnaire. The acceptance scores for each session are provided in Appendix C. Overall, mean scores of 4.7 for *enjoyment* were the highest and consistent across all robot-administered sessions, indicating that participants found the interaction fun and engaging—an idea reinforced by the perspectives they shared in semi-structured interviews (see Section 4.4). The portion of participants who rated their feeling of *anxiety* as "strongly disagree" increased from 72.7% in the first robot-administered session to 87.8% in the final robot-administered session (Figure 6), suggesting that participants became more comfortable with HRI over time. Similar trends were observed in *sociability* or perception of the robot as pleasant, nice, and understanding (percentage of "strongly agree" ratings in Session 2: 65.2%, Session 4: 71.2%). *Usefulness* scores also increased across sessions (percentage of "strongly agree" ratings in Session 2: 51.2%, Session 4: 57.6%). While mean scores for *trust* were lower, ranging from 3.5 to 3.7, participants reported higher trust in the robot over time (percentage of "agree" ratings in Session 2: 13.6%, Session 4: 25%).

Additionally, to explore the relationship between cognitive abilities and robot acceptance, we computed Spearman's correlation between MMSE scores and robot acceptance ratings. This analysis revealed a moderate negative correlation between MMSE scores and *trust* (r = -0.5, $p < .05$), suggesting that participants with higher cognitive scores tended to have lower trust in the system. This trend may be due to greater awareness of privacy concerns with AI technology. Spearman's correlation results between participants' BFI personality traits and robot acceptance showed no significance (see Tables 12, 13 in Appendix C).

4.4 User Experience in Robot-Administered Cognitive Tasks

In this section, we report the themes that emerged from the interviews throughout the study. We found that participants' responses most commonly related to the robot's effect on their cognitive abilities, their perceptions of the robot and interaction experience, their motivations to use the robot, and whether they would be likely to continue using it over time. Tables 2, 3, and 4 list all sub-themes and their frequencies.[4]

Table 2: Perceived ability to focus and promote cognitive health.

Statement	Users
Helps with memory retention	12 (55%)
Enhances thinking	19 (86%)
Increases attention to the cognitive tasks and surroundings in daily living	12 (55%)
Helps with learning new information or skills	10 (45%)

4.4.1 Increased Ability to Focus and Promote Cognitive Health. Participants frequently reported experiencing cognitive benefits from interacting with Blossom. The recurring themes from their feedback are listed in Table 2. Participants (N=12) reported that interacting with Blossom improved their memory recall. P19 stated, *"It's beneficial for me mentally; I sometimes forget things, and that's where the robot would come in"* (S3). Most participants (N=18) expressed that completing the tasks with Blossom enhanced their thinking ability by challenging them to work their brains. For example, P12 reported benefiting from answering Blossom's questions about animals in the semantic fluency task: "*[...] she getting me to think more. Stuff like that is good for my brain*" (S3). P9 found that they were "*a little confused about some of the things Blossom mentioned, but that's good because it made me think, and I think that machines are supposed to help you focus and think*" (S4).

Participants (N=12) found that completing the picture description task with Blossom improved their awareness of their surroundings and ability to focus on details in their everyday lives. P9 explained their thought process during the activity: "*You may think you focused, you know, but sometimes you not focused. [...] Even with me looking straight at the picture, I still didn't see everything*" (S2). Several participants noted that their attention to small details in their everyday lives improved after interacting with Blossom: "*It makes me more alert*" (P23, S2); "It's *beneficial for me, [it helps me] notice more details, be more observant*" (P19, S4). Finally, participants (N=10) reported learning new information when interacting with Blossom. P22 stated, "*I enjoy it; every day I talk with the robot, I learn more about culture*" (S4), and P9 reported that "*[...] to me, has been a learning session*" (S5).

4.4.2 Experience Interacting with Blossom. Table 3 lists the common themes from participants' feedback about their experience interacting with the robot, perceptions of the robot, and suggestions to improve it. Participants praised Blossom's appearance and features but were critical of the robot's complexity and ability to communicate clearly and efficiently. Half of the participants (N=11) described Blossom as cute or friendly, comparing it to a toy with more advanced capabilities: "*Blossom was just like a talking toy*" (P10, S2). More than half of the participants (N=14) described Blossom's voice as "understandable, clear" (P24, S2). P2 and P5 complimented

[4]Note that all participant quotes have been translated to English for presentation in this section.

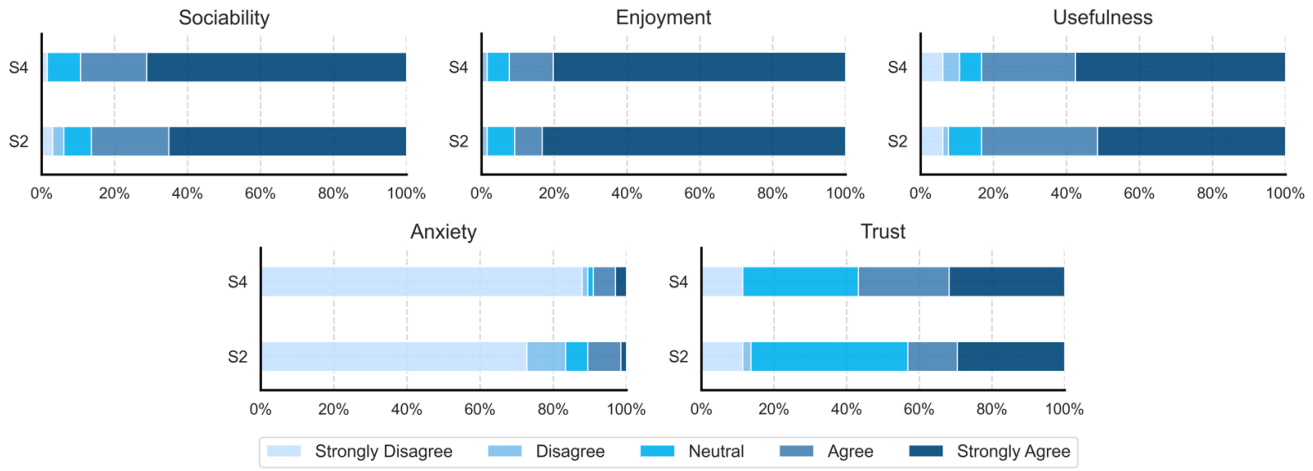


Figure 6: Results from acceptance questionnaires in the first (S2) and last (S4) robot-administered sessions.

its volume (P5, S2), cadence, and calming quality (P2, S4). Participants (N=5) also liked Blossom's movements.

All but one participant (N=21) described the interaction with Blossom as enjoyable. They commonly attributed their excitement to the robot's novelty ("*Because I have not seen any robot like this in my life. And the way it talks to me was different and kind of funny. And I like it*" (P16, S4), Blossom's responsiveness, and the thought-provoking nature of the tasks. They particularly enjoyed Blossom's ability to provide information, facilitate their comprehension, and converse; P7 stated, "*I enjoy when [Blossom]'s here 'cause he knows more things than I do. A lot of fun talking, listening, and saying things*" (S3).

Despite enjoying the interactions with Blossom, many (N=10) felt they would not have the technical skills or training to operate the robot alone. For example, P7 said, "*Personally [by] myself I wouldn't know how to operate*" (S4). Some suggested that they could operate the robot independently with enough training and practice: "*I need guidance, help at the beginning to start learning and gain practice*" (P20, S5). Upon inquiry, all participants affirmed that they felt comfortable interacting with Blossom or had become comfortable over the course of the study, with some even forming an emotional attachment: "*Today, [I felt] more confident with the robot, it's like we're friends now*" (P24, S4). Conversely, some participants (N=5) expressed anxiety and nervousness during the interactions due to their skepticism of technology and the duration of the exercises: "*Sometimes I was not very comfortable. I didn't know what else to say [about the picture]*" (P22, S3).

Participants found that although they were initially skeptical of Blossom and AI, they became more comfortable in general as the sessions progressed: "*At first, I was a little shy about dealing with her, with Blossom, and I would've preferred [the human administrator], but now it just doesn't make a difference*" (P2, S4); "*AI was like frightening for me at first because of the things I was seeing on social media, but now [...] I'm more comfortable*" (P2, S4). One participant who had an Alexa smart speaker also reported feeling more comfortable exploring a broader range of its applications after their HRI experience with Blossom.

Participants (N=8) appreciated how Blossom's communication was "*clear, precise*" (P18, S3) and "*straight to the point*" (P14, S2). Conversely, other participants (N=6) reported facing challenges communicating with Blossom, recalling times that they didn't understand the robot or the robot misunderstood them, "*I don't think the robot can understand feelings. I can't see that it understands me.*" (P4, S3). In addition, participants (N=6) noted the latency in Blossom's responses. They expressed frustration when the expected response was significantly delayed or failed to appear: "*I'm impatient when I'm talking to someone [...] she's too slow*" (P12, S2). Participants also interpreted Blossom's slow response time as indifference, assuming that "*it's not interested in my question*" (P23, S2) or that "*it lacks thought*" (P17, S4).

During the two interactive cognitive tasks, Blossom provided hints. Most participants (N=16) reported that the hints improved their focus, memory, and observation, "*It made me see what I wasn't seeing [in the picture] and remember things*" (P21, S4). In contrast, some participants (N=6) who found Blossom unhelpful during the tasks stated that they could have completed the task independently and that the hints provided by the robot were too vague: "*He didn't say look over here, look over there*" (P6, S3). For the semantic task, participants suggested that hints focused on specific fruits or animals would have been more useful than broad categories.

4.4.3 Suggested Improvements. Participants provided various suggestions for Blossom's conversational abilities. The most common (N=7) was to make Blossom more interactive by enabling two-way open conversations ("*I'd like to kind of ask her questions*" (P2, S3)), expanding topic variety, and extending interaction duration. P9 explained, "*If [Blossom's vocabulary] was wider or better, it would be even more entertaining*" (S5). Some participants (N=3) suggested personalizing Blossom to engage in conversations tailored to their interests and "*fit to [their] personality*" (P4, S2). P9 stated that personalizing the dialogue would increase user comfort "*because you would engage with it more, and it would put your mind into ease*" (S4). P4 was concerned that without personalized knowledge of their unique health history, Blossom may provide inaccurate healthcare

Table 3: HRI experience, perceptions of the robot, and suggested improvements.

Statement	Users
HRI Experience	
Found the tasks with the robot fun or enjoyable	20 (91%)
Found Blossom helpful for task completion	16 (73%)
Found Blossom unhelpful for task completion	6 (27%)
Felt comfortable interacting with Blossom	22 (100%)
Felt uncomfortable or anxious	5 (23%)
Noticed improvements in task performance, confidence, or comfort over time	7 (32%)
Perceptions	
Viewed as cute, friendly, or toy-like	11 (50%)
Enjoyed Blossom's voice	14 (64%)
Enjoyed Blossom's movements	5 (23%)
Easy to understand and communicate with	8 (36%)
Faced communication challenges with Blossom	6 (27%)
Complicated to set up and use alone	10 (46%)
Suggestions	
Shorten response time	6 (27%)
Make interactions more conversational	7 (32%)
Personalize responses	3 (14%)
Add/modify physical design features	5 (23%)
Enhance body movement	4 (18%)

recommendations, "*If [it's] giving me advice on what to do to help my health issues, it got to be personalized.*" (S5).

Participants suggested adding or modifying Blossom's physical appearance and capabilities. A commonly shared criticism (N=5) was the ambiguity of the robot's appearance and its lack of human-like characteristics: "*[. . .] might've been more comfortable if it looked like something that you've seen*" (P6, S2). Suggestions included "*a more humanized voice, less robotic, with more inflections*" (P23, S4) and facial expressions to make the robot appear friendlier and mimic human emotional intelligence: "*I would like the robot to show a little smile. [...] So that there is some empathy, something that is friendly*" (P23, S2). In addition to Blossom's physical features, participants (N=4) shared the desire to increase the robot's movement capabilities: "*Move around. Move the ears*" (P10, S5).

4.4.4 Perceived Usefulness and Interest in Future Use. Participants provided feedback on the usefulness of Blossom if integrated into their daily lives. Table 4 lists the common themes that emerged from this feedback. All but one participant (N=21) believed the robot would be beneficial for setting reminders and providing information. Many suggested that Blossom could remind them of daily tasks, such as taking medications and exercising: "*Remind me of tasks, time, and appointments [. . .] or where I leave things and give me information like an assistant*" (P23, S2); "*would be good for motivating me to exercise, telling me 'go out for a walk.'*" (P19, S4).

Participants also expressed interest in using Blossom to learn more about current news, health, and technology. As noted by P9, "*It could make my life easier by imparting information to me. It gives me an opportunity to expand my thinking if it can answer questions.*" (S3). P14 compared Blossom to a teacher: "*I can learn, like if I go to school or something, the robot can show me something [...] giving me [new] information.*" (S4). They speculated that future systems with these capabilities would enable older adults to live more independently, stating, *"There is hope for tomorrow for us"* (P24, S5).

Table 4: Perceived usefulness and interest in continued use.

Statement	Users
Useful for reminders and information	21 (96%)
Useful as a conversational partner at home	15 (84%)
Useful to support daily living and well-being	21 (96%)
Engaged in optional open dialogue	16 (73%)
Interested in interacting again due to novelty	10 (46%)
Uncertain about future use due to distrust	9 (41%)
Uncertain about future use due to lack of perceived usefulness	6 (27%)
Curious about AI and future robot capabilities	9 (41%)

Beyond providing reminders and assisting with daily tasks, more than half of the participants (N=15) remarked on Blossom's usefulness as an in-home companion or conversational partner. Many emphasized Blossom's role as entertainment to alleviate boredom and loneliness: "*It would be a good tool for people that don't really interact with human people*" (P4, S3). P22, in particular, preferred having Blossom as a conversational partner since "*there is dialogue [and] reciprocity*" (S3). Similarly, P14 mentioned using Blossom to improve their social skills and help them "*get around, [...] talk better, understand people better"* (S5).

In the last robot-administered session, most participants (N=16) chose to engage in an optional open dialogue with the robot, showing their interest in interacting with Blossom outside the study tasks. The majority (N = 10) suggested topics of conversation and engaged with Blossom to learn about geography and technology, discuss activities of interest, past trips, and music, share childhood memories, and reflect on family and friends.

Many participants (N=10) expressed interest in future interactions due to Blossom's novelty: "*I just want to know what she's got to say. I've never talked to a robot before*" (P10, S2). They hoped HRI would better prepare them for the increasing role of AI in the future, "*In the future, we will have to get used to living with [robots and AI], and this is an experience to get used to*" (P17, S3).

Despite general interest, some participants (N=9) hesitated to continue interacting with Blossom due to distrust in its capabilities and privacy concerns, "*I'm afraid of being listened to*" (P24, S2). These participants doubted the accuracy of Blossom's speech: "*I'm not for sure if it's gonna give me the right answer. [...] it is man-made and it can make mistakes*" (P10, S3). Some participants mentioned that they "*would have to be familiar with it*" (P8, S3) for future interactions. Additionally, hesitant participants (N=6) felt that Blossom would not be useful in their daily lives, often remarking that they might forget about Blossom due to other responsibilities, "*Days might go by [I would] forget about he even there*" (P7, S3).

5 Discussion

This work introduced LLM-powered SARs in the context of promoting and assessing cognitive health in elder care. In a 5-week study, 22 older adults engaged in three weekly robot-administered sessions to evaluate task performance, engagement, and user experience beyond the novelty effect. We alternated two task variations to minimize task memorization and established an individual baseline through an initial human-administered session. Our findings demonstrated the efficacy of SARs for in-residence cognitive support. Relative to each research question (see Section 3), we found that:

(1) The conversational SAR successfully engaged older adults with varying cognitive abilities in repeated weekly sessions of cognitive health-promoting tasks;
(2) Participants improved task performance after three robot-administered sessions, with a significantly enhanced level of detail, increased word count, and improved semantic completeness in their verbal picture descriptions, fewer word repetitions in semantic fluency, and a reduced need for hints by the administrator to complete the tasks;
(3) Participants accepted the robot and found it increasingly easy to use over time, giving it an average SUS score above 85.6 after the initial novelty wore off;
(4) Participants perceived the SAR as a valuable tool to enhance thinking, increase focus, and help them learn new skills; they were willing to interact with it in the future and, in some cases, enthusiastic to have their own robot.

These contributions aim to inform the design of more accessible, personalized, and autonomous robots that can effectively promote and assess cognitive health in elder care in real-world settings beyond clinics.

5.1 Robots as Tools to Promote Cognitive Health

Our study demonstrated the efficacy of SARs in engaging older adults in cognitive health-promoting tasks. The role of the SAR was to act as a delivery mechanism for these interactive cognitive tasks aimed at verbally engaging target populations. A critical factor for the successful home deployment of SARs is their accessibility. We chose Blossom as the platform for this study because it is affordable and could be quickly fabricated at the scale needed for future deployments, making it a viable tool for broader use in elder care. Blossom is an embodied robot that combines voice interaction with additional interactive capabilities that can enhance user engagement compared to technologies that rely solely on voice.

Our qualitative insights on user impressions and experience can inform the design of future SARs, targeting long-term acceptance, engagement, and comfort among older adults. Participants frequently reported cognitive benefits from their interactions with Blossom. Particularly, they perceived the robot as useful for enhancing thinking and memory recall, learning new tasks, and maintaining focus—not only during study tasks but also in their awareness and attention to detail in daily living. These self-reported benefits underscore the potential of SARs to keep older adults cognitively engaged while supporting well-being through continuous natural language interactions.

While some participants initially felt uncomfortable engaging with Blossom, their discomfort decreased with repeated exposure (as seen in the anxiety ratings in Figure 6). After repeated interactions, some participants also expressed that they developed an emotional connection with Blossom. There were mixed perceptions about trust in the robot's advice or suggestions, with participants with higher cognitive function (i.e., higher MMSE scores) showing lower trust and more hesitant attitudes, possibly due to greater awareness of AI and consequent privacy concerns. Nevertheless, trust scores increased over the study sessions across the cohort (see Figure 6) and 75% of participants in the normal aging cognitive group chose to engage in optional open dialogue with the robot. This finding highlights the importance of promoting familiarity and personalizing interactions between SARs and older adults. Tailored interactions that align with users' interests and cognitive abilities can increase trust and enhance the perceived cognitive health benefits of the technology. This approach is particularly important for overcoming the challenges associated with introducing voice technology to current older generations and individuals with cognitive impairment. As suggested in previous research [45], a training period can prevent confusion and discomfort. In our study, we leveraged the robot's agency to promote familiarity and transparency of its interactive capabilities and limitations, including the motor sounds and the time required for speech processing.

5.2 Considerations for in-Home Cognitive Assessment

Dementia is a growing concern worldwide as the population ages. Our study was not focused on clinical screening, hence we did not measure pre/post-study cognitive outcomes. However, our findings suggested the potential for using SARs as a delivery mechanism for cognitive assessment in residential settings. We implemented two tasks based on clinically validated cognitive tools, allowing us to collect data on spontaneous speech and verbal fluency, which have predictive value for early dementia screening [3, 39, 52]. The data collected were comparable to existing normative data across different age and cognitive groups (i.e., normal aging and cognitive impairment).

Furthermore, we found that repeated robot-administered sessions, conducted at weekly intervals with alternated task versions, led to overall improvements in verbal fluency of both cognitive groups. This finding highlights future opportunities for SARs to not only support assessments in residential settings but also help to maintain cognitive health in elder care through regular, personalized verbal HRI.

5.3 Design Implications

This work leveraged the linguistic abilities of LLMs to enable a SAR to administer interactive cognitive health-promoting tasks. Compared to rule-based chatbots, the use of LLMs offers dialogue flexibility, a higher variability of responses, and the ability to understand and adapt to the nuances of multilingual contexts. This capability makes the verbal interaction more contextually relevant, engaging, and adaptable during the cognitive-health promoting tasks. However, the integration of LLMs into SARs for real-world deployments, particularly in unconstrained living environments,

presents several design challenges and implications. While LLMs excel at generating contextualized natural language responses, they can also produce unpredictable or inappropriate content, especially in long-term, real-world living contexts where noise, data variability, and errors in speech recognition increase. Ensuring the safe and reliable use of LLMs in these contexts is essential.

5.3.1 Safe and Reliable Use of LLMs. Our study implemented an AI-driven moderator as a preventive measure designed to review and filter the quality of LLM-generated responses before they were delivered to participants via TTS. The purpose of this pilot safety mechanism was to ensure that all interactions remained appropriate and constrained within the cognitive tasks. After checking post-interaction transcripts, we confirmed that the LLM-generated responses were effectively constrained, with no problematic statements. However, this AI-driven approach relied on the same LLM using a different prompt, hence it may suffer from the same shortcomings. To mitigate risks, we followed a human-in-the-loop approach, where a researcher oversaw the sessions, with the ability to review the moderator's output in real time and interrupt the program if necessary. While effective in our study, this approach is not feasible for long-term in-home applications. Therefore, it is important for future research to develop robust methods to monitor LLM-generated speech and minimize problematic responses, particularly among vulnerable populations.

5.3.2 Speech Interaction Dynamics. Designing conversational AI for older adults requires careful consideration of speech interaction dynamics, including speech recognition, processing time, and communication breakdowns. These challenges are compounded in older populations with cognitive impairments, with frequent hesitations, pauses, and slower speech, all of which complicate turn-taking. It is important to avoid interruptions during user speech, especially when users pause to find the right words.

We identified latency in verbal interactions stemming from the use of cloud-based text-to-speech Whisper speech recognition, LLM API calls, and connectivity issues in the retirement homes where the study was conducted. These factors occasionally disrupted conversations, with some participants (27%) interpreting it as a sign that the robot was indifferent or unresponsive to their queries. These findings suggest that extended response times and interruptions can lead to disengagement, highlighting the need for robust turn-taking, accurate speech recognition in multilingual contexts, and adequate network infrastructure. Despite these challenges, user engagement and high Enjoyment scores (see Figure 6) indicate that latency did not affect overall robot acceptance. We also observed that participants learned to use the robot over the study sessions and became more accepting of its limitations, including processing time, aided by priming which included transparency about the robot's speech processing time (detailed in Section 3.3.3).

Response delays in verbal HRI due to issues with automatic speech recognition and response generation are a well-recognized research challenge [50, 62, 92]. Effective spoken interaction with SARs requires dialogue systems capable of handling multiple communication channels and quickly adapting to dynamic environments. Current cloud-based TTS and STT systems, while more accurate and superior in vocabulary size than local speech recognizers, often suffer from high latency, which disrupts verbal interactions. To enhance human-robot engagement among older adults, SARs need to be more responsive in real-time, with interruptible, adaptable speech processes.

5.3.3 Addressing Privacy Concerns and Facilitating Independent Use. Despite general interest in interacting with Blossom, 41% of participants hesitated to continue using the robot due to distrust in its capabilities and concerns about privacy, particularly regarding who might access and read their personal voice data. Interviews indicated that these mistrustful attitudes were shaped by media coverage, and participants trusted the robot more when they had a clear understanding of its capabilities and purpose.

Privacy concerns related to voice data collection remain significant among older adults [34, 47, 77, 79]. Addressing both research ethics (e.g., informed consent and data protection) and societal ethics (e.g., integration of nonintrusive SARs at home) requires transparent communication about the robot's functions, limitations, and data access protocols to foster long-term engagement, trust, and acceptance. Strategies could include priming with robot's agency to set realistic expectations (as incorporated in our study design), and stakeholder involvement throughout the design, initial evaluation, and deployment phases. Furthermore, ensuring feature anonymity in voice analysis can prevent privacy threats associated with speech processing [41].

Despite overall enjoyment and perceived usefulness, some participants reported lacking the technical skills or training to operate the robot independently. To address this, we recommend that SARs include accessible training resources and consider the role of family members or caregivers in facilitating interactions.

5.4 Limitations and Future Work

We acknowledge several limitations in our study. First, despite targeting a diverse and multilingual population, our sample size was relatively small. To improve the generalizability of our findings, future studies should include larger, more culturally diverse cohorts to evaluate the feasibility of SARs for in-home cognitive support across varied populations and settings. Second, while the 5-week period was sufficient for evaluating performance changes, observing engagement, and capturing user impressions beyond the novelty effect, it did not evaluate the long-term efficacy of SARs in daily, unsupervised living contexts. Future research should focus on longer-term deployments to assess the role of SARs in continuous cognitive health monitoring. Integrating SARs with additional home monitoring technologies could further enhance their ability to support safety and well-being by tracking behavior, monitoring cognitive health, detecting potential risks, and providing real-time updates to caregivers. Our study focused on two SAR-administered cognitive tasks based on validated tools: picture description and semantic fluency. Future studies could target a broader range of cognitive functions, including memory, attention, executive function, and problem-solving.

To minimize task memorization by older adult participants, our design alternated two variations of each task across five weekly sessions. While a 2-week washout period was included between tasks of the same version (see Figure 3), we acknowledge that task

performance is subject to learning effects from prior exposure. Our primary analysis compared task performance between the first (individual baseline) and the last session, after three weekly robot-administered sessions. We stress that our study did not aim to assess the effect of SARs on cognitive outcomes but rather evaluate their feasibility as engaging tools to promote cognitive health among older adults.

To explore learning effects, a secondary analysis examined change trends in task performance across the study sessions (see Appendix D). For consistency, we only compared sessions conducted by the same administrator (human or robot) given the potential motivational effects of the robot on task performance, which were outside the scope of this study. This analysis suggested task-specific learning effects, particularly when comparing sessions with the same task version (e.g., S1 vs. S5 and S2 vs. S4). While learning effects were present in individual cognitive tasks, they did not impact the primary goal and contribution of this study – to validate the feasibility of SARs as tools to promote cognitive health through human-robot verbal dialogue. Furthermore, there was general cognitive engagement across the three robot-administered sessions, with gradual improvements in some of the extracted features, indicating the capacity of SARs to keep users engaged in regular cognitive tasks.

Task difficulty presents an additional consideration in the interpretation of performance changes over time, as different pictures can influence lexicosemantic processing [76], and the semantic search for animals/fruits can reflect varying levels of difficulty [24]. While such analysis was beyond the scope of this study, future research could explore within- and between-task performance, extending the duration of the study, incorporating additional task variations, and addressing task difficulty as a confounding factor.

The weekly interactions lasted only 10 minutes, for a total of 30 minutes with the robot. We believe this was sufficient to overcome novelty effects and enable participants to evaluate the robot. However, many participants expressed interest in more interactions with the robot. Future work could incorporate a more diverse set of cognitive tasks to provide more opportunities for interaction and explore the relationship between interaction time, robot acceptance, and cognitive task performance.

Another limitation highlighted by participants was the latency in verbal interaction with Blossom. While this is a well-recognized challenge in spoken language interactions with robots [50], these technological issues are being addressed by new tools such as OpenAI's Realtime API[5], which reduces latency by streaming audio inputs and outputs directly with a persistent WebSocket connection to access GPT-4o, a promising area for future work. Future study iterations with older populations could further incorporate the suggested design features for Blossom. These include more expressive facial features, ear movements, and personalized interactions with an appropriate tone and topics of interest. Lastly, our future work will analyze the eye gaze data collected during picture descriptions to explore participants' focus during the task.

[5]https://openai.com/index/introducing-the-realtime-api/

6 Conclusion

With the world population of older adults growing at an unprecedented rate, there is a pressing need for accessible technologies that promote cognitive health and detect early signs of cognitive decline. This study contributes novel insights into the efficacy of LLM-powered SARs to promote cognitive health in elder care within residential settings. Our findings show that: 1) the conversational robot engaged older adults in repeated verbal HRI sessions aimed at promoting cognitive health, with increased social engagement in tasks administered by the robot; 2) task performance increased after repeated weekly robot-administered sessions, with participants providing more detailed picture descriptions, making fewer word repetitions in semantic fluency, and requiring fewer hints by the administrator to complete both tasks; 3) older adults accepted the SAR in this context; and 4) participants expressed interest in future use and recognized the utility of SARs for enhancing cognitive abilities, including focus, thinking, and learning new skills. Additionally, the speech data collected with our LLM-powered SAR system implementation are based on clinically validated cognitive tools. Therefore, we argue the potential of LLM-powered SARs for in-home cognitive support in elder care. With consistent use, for which accessibility and personalized interactions are key, such technologies could ultimately monitor cognitive health and identify individuals who may require more comprehensive diagnostic evaluation.

Acknowledgments

This work was supported by the National Science Foundation grant NSF IIS-1925083, the National Institutes of Health grant NIH 5-P30-AG-073105-02, and the UK Dementia Research Institute award number UK DRI-7003 through UK DRI Ltd, principally funded by the Medical Research Council, and additional funding partner Alzheimer's Society.

References

[1] Jordan Abdi, Ahmed Al-Hindawi, Tiffany Ng, and Marcela P Vizcaychipi. 2018. Scoping review on the use of socially assistive robot technology in elderly care. *BMJ open* 8, 2 (2018), e018815.

[2] Hojjat Abdollahi, Ali Mollahosseini, Josh T Lane, and Mohammad H Mahoor. 2017. A pilot study on using an intelligent life-like robot as a companion for elderly individuals with dementia and depression. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. IEEE, 541–546.

[3] Samad Amini, Boran Hao, Jingmei Yang, Cody Karjadi, Vijaya B Kolachalama, Rhoda Au, and Ioannis C Paschalidis. 2024. Prediction of Alzheimer's disease progression within 6 years using speech: A novel approach leveraging language models. *Alzheimer's & Dementia* (2024).

[4] Victor Nikhil Antony, Sue Min Cho, and Chien-Ming Huang. 2023. Co-designing with older adults, for older adults: robots to promote physical activity. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 506–515.

[5] Stefanie Baisch, Thorsten Kolling, Arthur Schall, Saskia Rühl, Stefanie Selic, Ziyon Kim, Holger Rossberg, Barbara Klein, Johannes Pantel, Frank Oswald, et al. 2017. Acceptance of social robots by elder people: does psychosocial functioning matter? *International Journal of Social Robotics* 9 (2017), 293–307.

[6] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human-Computer Interaction* 24, 6 (2008), 574–594.

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[8] Elizabeth Broadbent, Kathy Peri, Ngaire Kerse, Chandimal Jayawardena, IHan Kuo, Chandan Datta, and Bruce MacDonald. 2014. Robots in older people's homes to improve medication adherence and quality of life: a randomised cross-over trial. In *Social Robotics: 6th International Conference, ICSR 2014, Sydney, NSW, Australia, October 27-29, 2014. Proceedings 6*. Springer, 64–73.

[9] Henry Brodaty and Marika Donkin. 2022. Family caregivers of people with dementia. *Dialogues in clinical neuroscience* (2022).

[10] J Brooke. 1996. SUS: A "quick and dirty" Usability Scale. *Usability Evaluation in INdustry/Taylor and Francis* (1996).

[11] Diana Bruno and Sofia Schurmann Vignaga. 2019. Addenbrooke's cognitive examination III in the diagnosis of dementia: a critical review. *Neuropsychiatric Disease and Treatment* (2019), 441–447.

[12] Clare Carroll, Catherine Chiodo, Adena Xin Lin, Meg Nidever, and Jayanth Prathipati. 2017. Robin: enabling independence for individuals with cognitive disabilities using voice assistive technology. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 46–53.

[13] Chitat Chan and Feng Li. 2023. Developing a natural language-based AI-chatbot for social work training: an illustrative case study. *China Journal of Social Work* 16, 2 (2023), 121–136.

[14] Yu-Ling Chang, Di-Hua Luo, Tsung-Ren Huang, Joshua OS Goh, Su-Ling Yeh, and Li-Chen Fu. 2022. Identifying Mild Cognitive Impairment by Using Human-Robot Interactions. *Journal of Alzheimer's Disease* 85, 3 (2022), 1129–1142.

[15] Linda J Clark, Margaret Gatz, Ling Zheng, Yu-Ling Chen, Carol McCleary, and Wendy J Mack. 2009. Longitudinal verbal fluency in normal aging, preclinical, and prevalent Alzheimer's disease. *American Journal of Alzheimer's Disease & Other Dementias®* 24, 6 (2009), 461–468.

[16] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[17] Eric Deng, Bilge Mutlu, Maja J Mataric, et al. 2019. Embodiment in socially interactive robots. *Foundations and Trends® in Robotics* 7, 4 (2019), 251–356.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[19] Juan Fasola and Maja J Mataric. 2012. Using socially assistive human-robot interaction to motivate physical exercise for older adults. *Proc. IEEE* 100, 8 (2012), 2512–2526.

[20] David Feil-Seifer and Maja J Matarić. 2011. Socially assistive robotics. *IEEE Robotics & Automation Magazine* 18, 1 (2011), 24–31.

[21] Marshal F Folstein, Susan E Folstein, and Paul R McHugh. 1975. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research* 12, 3 (1975), 189–198.

[22] Emil Fristed, Caroline Skirrow, Marton Meszaros, Raphael Lenain, Udeepa Meepegama, Stefano Cappa, Dag Aarsland, and Jack Weston. 2022. A remote speech-based AI system to screen for early Alzheimer's disease via smartphones. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 14, 1 (2022), e12366.

[23] Yajie Gao, Yaping Chang, Tangwutu Yang, and Zhihao Yu. 2025. Consumer acceptance of social robots in domestic settings: A human-robot interaction perspective. *Journal of Retailing and Consumer Services* 82 (2025), 104075.

[24] Carla Giacominelli, Cybelle Diniz, and Paulo Bertolucci. 2012. P2-263: Qualitative analysis of clustering in semantic fluency fruit tasks. *Alzheimer's & Dementia* 8, 4S_Part_10 (2012), P354–P354.

[25] Susel Góngora Alonso, Sofiane Hamrioui, Isabel de la Torre Díez, Eduardo Motta Cruz, Miguel López-Coronado, and Manuel Franco. 2019. Social robots for people with aging and dementia: a systematic review of literature. *Telemedicine and e-Health* 25, 7 (2019), 533–540.

[26] Harold Goodglass, Edith Kaplan, and Sandra Weintraub. 2001. *BDAE: The Boston diagnostic aphasia examination.* Lippincott Williams & Wilkins Philadelphia, PA.

[27] Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. 2009. Influence of social presence on acceptance of an assistive social robot and screen agent by elderly users. *Advanced Robotics* 23, 14 (2009), 1909–1923.

[28] Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. 2010. Assessing acceptance of assistive social agent technology by older adults: the almere model. *International Journal of Social Robotics* (2010).

[29] Richard J Holden. 2020. A simplified system usability scale (SUS) for cognitively impaired and older adults. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, Vol. 9. SAGE Publications Sage CA: Los Angeles, CA, 180–182.

[30] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. 2024. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review* 57, 7 (2024), 175.

[31] Bahar Irfan, Sanna Kuoppamäki, and Gabriel Skantze. 2024. Recommendations for designing conversational companion robots with older adults through foundation models. *Frontiers in Robotics and AI* 11 (2024), 1363713.

[32] Nayaar Islam, Rola Hashem, Maryse Gad, Aime Brown, Brooke Levis, Christel Renoux, Brett D Thombs, and Matthew DF McInnes. 2023. Accuracy of the Montreal Cognitive Assessment Tool for detecting mild cognitive impairment: a systematic review and meta-analysis. *Alzheimer's & Dementia* 19, 7 (2023), 3235–3243.

[33] Angela M Jensen, Helen J Chenery, and David A Copland. 2006. A comparison of picture description abilities in individuals with vascular subcortical lesions and Huntington's disease. *Journal of Communication Disorders* 39, 1 (2006), 62–77.

[34] Yucheng Jin, Wanling Cai, Li Chen, Yizhe Zhang, Gavin Doherty, and Tonglin Jiang. 2024. Exploring the Design of Generative AI in Supporting Music-based Reminiscence for Older Adults. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

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

[36] Andrew Kertesz. 1982. The western aphasia battery. *(No Title)* (1982).

[37] Cory D Kidd and Cynthia Breazeal. 2004. Effect of a robot on user perceptions. In *2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)(IEEE Cat. No. 04CH37566)*, Vol. 4. IEEE, 3559–3564.

[38] Sunyoung Kim and Abhishek Choudhury. 2021. Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study. *Computers in Human Behavior* 124 (2021), 106914.

[39] Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H Robert, et al. 2015. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 1, 1 (2015), 112–124.

[40] Agnieszka Korchut, Sebastian Szklener, Carla Abdelnour, Natalia Tantinya, Joan Hernández-Farigola, Joan Carles Ribes, Urszula Skrobas, Katarzyna Grabowska-Aleksandrowicz, Dorota Szczęśniak-Stańczyk, and Konrad Rejdak. 2017. Challenges for service robots—requirements of elderly adults with cognitive impairments. *Frontiers in neurology* 8 (2017), 228.

[41] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Philip Raschke. 2020. Privacy implications of voice and speech analysis–information disclosure by inference. *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14* (2020), 242–258.

[42] Sandeep Kaur Kuttal, Bali Ong, Kate Kwasny, and Peter Robe. 2021. Trade-offs for substituting a human with an agent in a pair programming context: the good, the bad, and the ugly. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–20.

[43] Jieon Lee, Daeho Lee, and Jae-gil Lee. 2024. Influence of rapport and social presence with an AI psychotherapy chatbot on users' self-disclosure. *International Journal of Human–Computer Interaction* 40, 7 (2024), 1620–1631.

[44] Iolanda Leite, Carlos Martinho, and Ana Paiva. 2013. Social robots for long-term interaction: a survey. *International Journal of Social Robotics* 5 (2013), 291–308.

[45] Maria R Lima, Sophie Horrocks, Sarah Daniels, Moesha Lamptey, Matthew Harrison, and Ravi Vaidyanathan. 2023. The Role of Conversational AI in Ageing and Dementia Care at Home: A Participatory Study. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 571–578.

[46] Maria R Lima, Ting Su, Melanie Jouaiti, Maitreyee Wairagkar, Paresh Malhotra, Eyal Soreq, Payam Barnaghi, and Ravi Vaidyanathan. 2023. Discovering behavioural patterns using conversational technology for in-home health and well-being monitoring. *IEEE Internet of Things Journal* (2023).

[47] Maria R Lima, Maitreyee Wairagkar, Manish Gupta, Ferdinando Rodriguez y Baena, Payam Barnaghi, David J Sharp, and Ravi Vaidyanathan. 2021. Conversational affective social robots for ageing and dementia support. *IEEE Transactions on Cognitive and Developmental Systems* 14, 4 (2021), 1378–1397.

[48] Maria R Lima, Maitreyee Wairagkar, Nirupama Natarajan, Sridhar Vaitheswaran, and Ravi Vaidyanathan. 2021. Robotic telemedicine for mental health: a multimodal approach to improve human-robot engagement. *Frontiers in Robotics and AI* 8 (2021), 618866.

[49] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, Vol. 2023. American Medical Informatics Association, 1105.

[50] Matthew Marge, Carol Espy-Wilson, Nigel G Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gil Blankenship, Joyce Chai, Hal Daumé III, Debadeepta Dey, et al. 2022. Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language* 71 (2022), 101255.

[51] Laura Martinengo, Elaine Lum, and Josip Car. 2022. Evaluation of chatbot-delivered interventions for self-management of depression: Content analysis. *Journal of affective disorders* 319 (2022), 598–607.

[52] Michelle McDonnell, Lauren Dill, Stella Panos, Stacy Amano, Warren Brown, Shadee Giurgius, Gary Small, and Karen Miller. 2020. Verbal fluency as a screening tool for mild cognitive impairment. *International Psychogeriatrics* 32, 9 (2020), 1055–1062.

[53] Wendy Moyle. 2019. The promise of technology in the future of dementia care. *Nature Reviews Neurology* 15, 6 (2019), 353–359.

[54] United Nations. 2023. World Social Report 2023: Leaving No One Behind In An Ageing World. https://desapublications.un.org/publications/world-social-report-2023-leaving-no-one-behind-ageing-world.

[55] Toshiaki Nishio, Yuichiro Yoshikawa, Kazuki Sakai, Takamasa Iio, Mariko Chiba, Taichi Asami, Yoshinori Isoda, and Hiroshi Ishiguro. 2021. The effects of physically

embodied multiple conversation robots on the elderly. *Frontiers in Robotics and AI* 8 (2021), 633045.

[56] Amy O'Connell, Ashveen Banga, Jennifer Ayissi, Nikki Yaminrafie, Ellen Ko, Andrew Le, Bailey Cislowski, and Maja Mataric. 2024. Design and Evaluation of a Socially Assistive Robot Schoolwork Companion for College Students with ADHD. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 533–541.

[57] Richelle ACM Olde Keizer, Lex van Velsen, Mathieu Moncharmont, Brigitte Riche, Nadir Ammour, Susanna Del Signore, Gianluca Zia, Hermie Hermens, and Aurèle N'Dja. 2019. Using socially assistive robots for monitoring and preventing frailty among older adults: a study on usability and user experience challenges. *Health and technology* 9 (2019), 595–605.

[58] Ronan Peter Daniel O'Malley, Bahman Mirheidari, Kirsty Harkness, Markus Reuber, Annalena Venneri, Traci Walker, Heidi Christensen, and Dan Blackburn. 2021. Fully automated cognitive screening tool based on assessment of speech and language. *Journal of Neurology, Neurosurgery & Psychiatry* 92, 1 (2021), 12–15.

[59] Batyrkhan Omarov, Zhandos Zhumanov, Aidana Gumar, and Leilya Kuntunova. 2023. Artificial intelligence enabled mobile chatbot psychologist using AIML and cognitive behavioral therapy. *International Journal of Advanced Computer Science and Applications* 14, 6 (2023).

[60] World Health Organization. 2022. Ageing and health (2021). https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

[61] Michael Quinn Patton. 2005. Qualitative research. *Encyclopedia of statistics in behavioral science* (2005).

[62] Hannah Pelikan and Emily Hofstetter. 2023. Managing delays in human-robot interaction. *ACM Transactions on Computer-Human Interaction* 30, 4 (2023), 1–42.

[63] Sandra Petersen, Susan Houston, Huanying Qin, Corey Tague, and Jill Studley. 2017. The utilization of robotic pets in dementia care. *Journal of alzheimer's disease* 55, 2 (2017), 569–574.

[64] Chloé Pou-Prom, Stefania Raimondo, and Frank Rudzicz. 2020. A conversational robot for older adults with Alzheimer's disease. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 3 (2020), 1–25.

[65] Rhian C Preston, Kenna Dinsdale, Madison R Shippy, and Naomi T Fitter. 2024. Robot-mediated nudges for workplace health: not a one-size-fits-all modeling problem. *International journal of social robotics* 16, 5 (2024), 899–918.

[66] Sarah E Price, Glynda J Kinsella, Ben Ong, Elsdon Storey, Elizabeth Mullaly, Margaret Phillips, Lanki Pangnadasa-Fox, and Diana Perre. 2012. Semantic verbal fluency strategies in amnestic mild cognitive impairment. *Neuropsychology* 26, 4 (2012), 490.

[67] Lihui Pu, Wendy Moyle, Cindy Jones, and Michael Todorovic. 2019. The effectiveness of social robots for older adults: a systematic review and meta-analysis of randomized controlled studies. *The Gerontologist* 59, 1 (2019), e37–e51.

[68] Kun Qian, Zixing Zhang, Yoshiharu Yamamoto, and Björn W Schuller. 2021. Artificial intelligence internet of things for the elderly: From assisted living to health-care monitoring. *IEEE Signal Processing Magazine* 38, 4 (2021), 78–88.

[69] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[70] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[71] Beatrice Rammstedt, Christoph J Kemper, Mira Céline Klein, Constanze Beierlein, and Anastassiya Kovaleva. 2013. A short scale for assessing the big five dimensions of personality: 10 item big five inventory (BFI-10). *methods, data, analyses* 7, 2 (2013), 17.

[72] Hayley Robinson, Bruce MacDonald, and Elizabeth Broadbent. 2014. The role of healthcare robots for older people at home: A review. *International Journal of Social Robotics* 6 (2014), 575–591.

[73] Silvia Rossi, Gabriella Santangelo, Mariacarla Staffa, Simone Varrasi, Daniela Conti, and Alessandro Di Nuovo. 2018. Psychometric evaluation supported by a social robot: Personality factors and technology acceptance. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 802–807.

[74] Frank Rudzicz, Rosalie Wang, Momotaz Begum, and Alex Mihailidis. 2015. Speech interaction with personal assistive robots supporting aging at home for individuals with Alzheimer's disease. *ACM Transactions on Accessible Computing (TACCESS)* 7, 2 (2015), 1–22.

[75] Francesco Salis, Diego Costaggiu, and Antonella Mandas. 2023. Mini-mental state examination: Optimal cut-off levels for mild and severe cognitive impairment. *Geriatrics* 8, 1 (2023), 12.

[76] Mustafa Seçkin and Merve Savaş. 2023. Picnic, accident or cookies? A systematic approach to guide the selection of the picture definition tasks in linguistic assessment. *Archives of Clinical Neuropsychology* 38, 2 (2023), 236–246.

[77] William Seymour, Xiao Zhan, Mark Cote, and Jose Such. 2023. A systematic review of ethical concerns with voice assistants. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 131–145.

[78] Zhonghao Shi, Amy O'Connell, Zongjian Li, Siqi Liu, Jennifer Ayissi, Guy Hoffman, Mohammad Soleymani, and Maja J Matarić. 2024. Build Your Own Robot Friend: An Open-Source Learning Module for Accessible and Engaging AI Education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23137–23145.

[79] Jianna So, Samantha Estrada, Matthew Jörke, Eva Bianchi, Maria Wang, Nava Haghighi, Kristen L Fessele, James A Landay, and Andrea Cuadra. 2024. " They Make Us Old Before We're Old": Designing Ethical Health Technology with and for Older Adults. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–30.

[80] Unreal Speech. 2024. Unreal Speech API. https://unrealspeech.com/.

[81] Vivian P Ta-Johnson, Carolynn Boatfield, Xinyu Wang, Esther DeCero, Isabel C Krupica, Sophie D Rasof, Amelie Motzer, and Wiktoria M Pedryc. 2022. Assessing the topics and motivating factors behind human-social chatbot interactions: Thematic analysis of user experiences. *JMIR human factors* 9, 4 (2022), e38876.

[82] Marion Lara Tan, Raj Prasanna, Kristin Stock, Emma EH Doyle, Graham Leonard, and David Johnston. 2020. Modified usability framework for disaster apps: a qualitative thematic analysis of user reviews. *International Journal of Disaster Risk Science* 11 (2020), 615–629.

[83] Adriana Tapus, Cristian Tapus, and Maja J Mataric. 2009. The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia. In *2009 IEEE international conference on rehabilitation robotics*. IEEE, 924–929.

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

[85] Anouk Van Maris, Nancy Zook, Praminda Caleb-Solly, Matthew Studley, Alan Winfield, and Sanja Dogramadzi. 2020. Designing ethical social robots— A longitudinal field study with older adults. *Frontiers in Robotics and AI* 7 (2020), 1.

[86] Ryan Van Patten, Amber V Keller, Jacqueline E Maye, Dilip V Jeste, Colin Depp, Laurel D Riek, and Elizabeth W Twamley. 2020. Home-based cognitively assistive robots: maximizing cognitive functioning and maintaining independence in older adults without dementia. *Clinical interventions in aging* (2020), 1129–1139.

[87] Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. 2006. The role of physical embodiment in human-robot interaction. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 117–122.

[88] WHO. 2023. Dementia key facts. https://www.who.int/news-room/fact-sheets/detail/dementia.

[89] Yasunori Yamada, Kaoru Shinkawa, Miyuki Nemoto, Kiyotaka Nemoto, and Tetsuaki Arai. 2023. A mobile application using automatic speech analysis for classifying Alzheimer's disease and mild cognitive impairment. *Computer Speech & Language* 81 (2023), 101514.

[90] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–35.

[91] Yang Ye, Hengxu You, and Jing Du. 2023. Improved trust in human-robot collaboration with ChatGPT. *IEEE Access* (2023).

[92] Ji Won Yoon, Beom Jun Woo, and Nam Soo Kim. 2022. Hubert-ee: Early exiting hubert for efficient speech recognition. *arXiv preprint arXiv:2204.06328* (2022).

A Cognitive Tasks

A.1 LLM Prompt for Cognitive Tasks

Table 5 describes the prompt for robot-administered cognitive tasks and Table 6 lists the parameters used with OpenAI API. For Spanish sessions, the prompt had an additional instruction: "For this interaction, reply in Spanish. Use *<gender>* for the user". The prompts underwent in-lab pilot testing to ensure the reliability of GPT automated translations from English to Spanish before responses were delivered to participants via TTS.

Table 5: Prompt used in LLM-powered robot-administered cognitive tasks.

Element	Prompt
Role	You are Blossom a friendly social robot acting as a coach in a verbal interactive task to promote cognitive skills.
User	Older adults who are unfamiliar with robots and need time to think.
Context	Guide users to describe *<picture name OR semantic category>* from *<neuropsychological test>*. Interaction is limited to this task.
Tone	Encourage participants to keep engaging with supportive comments and short hints. Use friendly language be patient and engaging. Don't use emojis.
Start	Greet the user and ask for their name. Use name throughout. Ask if they are ready to play a game. Introduce the task with *<example of introduction>*.
Turns	Give turns and wait for user responses.
Hints	Use the groups for hints without naming specific *<picture actions OR category words>*. Give one hint per turn *<example of hint>*.
Groups (picture)	*<areas of Cookie Theft OR areas of Picnic Scene>*.
Groups (semantic)	*<groups of animals OR groups of fruits>*.
Track (picture)	Track what the user described.
Track (semantic)	Track *<animals OR fruits>* mentioned in each category. If user asks for a hint prompt about a group not already mentioned.
End (picture)	After all picture areas are mentioned end the task with *<example of end message>*.
End (semantic)	After at least two *<animals OR fruits>* from each group are mentioned end the task with *<example of end message>*.

Table 6: Input parameters for OpenAI's API call.

Parameter	Value
model	GPT-4o
messages	*<transcript>*
stop	null
temperature	1
frequency_penalty	0
presence_penalty	0
max_tokens	256
top_p	1

Table 7: Groups used in picture description tasks with pre-defined areas and actions. The content units were selected based on literature about neuropsychological evaluations [33].

Groups	Examples of content units
Picture: *Cookie Theft*	
Boy	brother, stool, falling, reaching, cookies, jar
Girl	sister, laughing, reaching, help
Woman	mother, sink, washing, dishes, water, overflowing
Atmosphere	disaster, mess, chaos, family
Feelings	excited, determined, distracted, anxious, sneaky, curious
Picture: *Picnic Scene*	
Couple	picnic, reading, book, drink, basket, radio
Boy	kite, dog, water, edge
Child	beach, playing, sandcastle, spade
People	fishing, jetty, pier, sailing, water, boat
Objects	car, garage, flag, flying, tree, house

Table 8: Groups used in semantic fluency tasks.

Groups	Examples of words
Semantic Category: *Animals*	
Pet	Dog, cat, hamster, guinea pig, fish, rabbit
Jungle	Lion, tiger, elephant, wolf, bear, gorilla
Can fly	Bat, eagle, sparrow, owl, falcon, pigeon
Ocean	Whale, shark, dolphin, octopus, sea turtle, squid
Farm	Cow, sheep, goat, horse, pig, chicken
North pole	Polar bear, penguin, seal, walrus, reindeer
Semantic Category: *Fruits*	
Citrus	Orange, lemon, lime, tangerine, clementine
Yellow	Banana, mango, lemon, pineapple
Tropical	Mango, pineapple, coconut, passionfruit, kiwi
Berries	Strawberry, raspberry, blackberry, blueberry, cranberry
With a pit	Peach, plum, cherry, apricot, mango
Used for juices	Strawberry, blueberry, grape, orange, mango, apricot

A.2 Moderator Prompt

Table 9: Prompt for the moderator to determine if a response is within the desired context.

Element	Prompt
Role	You are moderating responses generated by a robot interacting with older adults.
Context	*<task>*. The robot is acting as coach prompting users with hints to *<goal of task>*. In the end the robot praises the user and moves to the next task. The response must be in this context.
Output	Given this interaction context, is the response appropriate? Reply with just Yes or No.

A.3 Picture Description Task

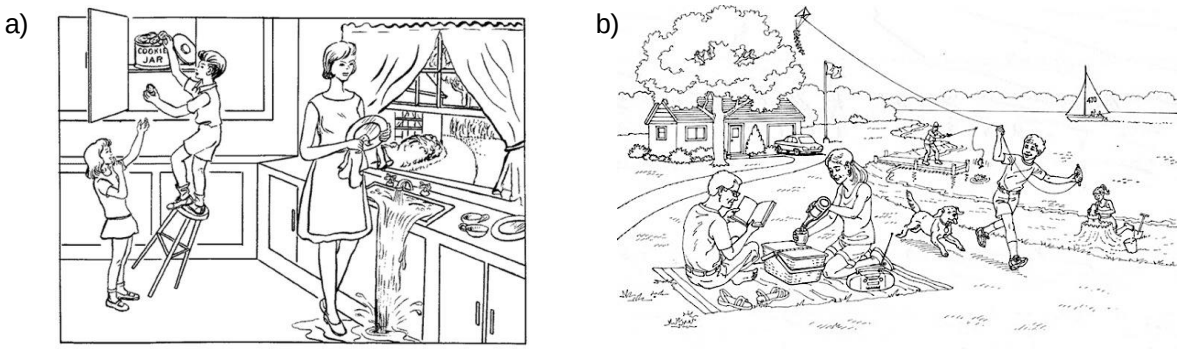


Figure 7: Pictures used: a) Cookie Theft from the Boston Diagnostic Aphasia Examination, b) Picnic Scene from the Western Aphasia Battery-Revised.

B Use of Language in Picture Descriptions

Table 10: Spearman's correlation results between traditional linguistic features and MMSE scores, as well as comparison results between Session 1 and Session 5 across the cohort using the paired Wilcoxon signed-rank test.

		MMSE	Session 1 vs. Session 5	
Feature	**Description**	$r\ (p, p_{corrected})$	$W\ (p, p_{corrected})$	Cohen's d
Type-Token Ratio	The ratio of unique words to total words (tokens) in a transcript corrected for total length, measuring lexical diversity.	0.416 (0.054, 0.220)	**33 (0.001*, 0.009*)**	**0.65**
Propositional Idea Density	The number of distinct propositions (i.e. distinct facts or notions contained in a text) divided by total words, assessing vocabulary richness.	-0.227 (0.31, 0.465)	84 (0.273, 0.273)	0.29
Noun Rate	The proportion of nouns relative to total words in the transcript.	0.137 (0.543, 0.652)	68 (0.167, 0.2)	0.4
Pronoun Rate	The proportion of pronouns relative to relative to total words.	0.008 (0.973, 0.973)	67.5 (0.095, 0.142)	0.41
Verb Rate	The proportion of verbs relative to relative to total words.,	-0.245 (0.272, 0.465)	54.5 (0.034*, 0.102)	0.5
Adjective Rate	The proportion of adjectives relative to total words.	0.389 (0.073, 0.220)	44.5 (0.074, 0.142)	0.57

C Robot Acceptance

Table 11: Items used in the robot acceptance questionnaire.

Construct	Description	Items
Perceived sociability	The perceived ability of the robot to perform sociable behavior.	The robot is a pleasant conversational partner. I think the robot is nice. I feel the robot understands me.
Perceived enjoyment	The degree to which the user finds interacting with the robot enjoyable or fascinating.	I enjoyed interacting with the robot. I find the robot enjoyable. I find the robot boring.
Perceived usefulness	The degree to which one believes the robot would be assistive in the daily context at home.	I would find the robot useful in my life. It would be convenient for me to have the robot. I think the robot could help promote cognitive health.
Anxiety	Evoking anxious or emotional reactions when using the robot.	If I were to use the robot at home, I would be afraid to make mistakes with it. I find the robot scary. I find the robot intimidating.
Trust	The belief that the system performs with personal integrity and reliability.	I would trust the robot if it gave me advice. I would follow the advice the robot gives me.

Table 12: Acceptance scores for robot-administered sessions, presented as mean (standard deviation).

Session	Sociability	Enjoyment	Usefulness	Anxiety	Trust
S2	4.42 (0.66)	4.73 (0.51)	4.21 (0.80)	4.44 (0.64)	3.48 (1.18)
S3	4.48 (0.60)	4.73 (0.57)	4.21 (1.02)	4.68 (0.44)	3.48 (1.19)
S4	4.58 (0.56)	4.71 (0.58)	4.24 (0.98)	4.65 (0.51)	3.66 (1.21)

Table 13: Spearman's correlation between robot acceptance ratings and MMSE scores in the last robot-administered session, presented as r (p, $p_{corrected}$).

	Sociability	Enjoyment	Usefulness	Anxiety	Trust
MMSE	-0.308 (0.163, 0.288)	-0.077 (0.733, 0.733)	-0.118 (0.600, 0.733)	0.302 (0.173, 0.288)	**-0.540 (0.009*, 0.047*)**

Table 14: Spearman's correlation between robot acceptance ratings and BFI personality traits considering mean acceptance scores from three robot-administered sessions, presented as r (p, $p_{corrected}$).

	Sociability	Enjoyment	Usefulness	Anxiety	Trust
Extraversion	0.381 (0.080, 0.670)	0.119 (0.598, 0.830)	0.427 (0.047*, 0.590)	0.078 (0.732, 0.858)	0.490 (0.021*, 0.518)
Agreeableness	0.141 (0.532, 0.830)	0.158 (0.482, 0.830)	0.120 (0.594, 0.830)	0.132 (0.558, 0.830)	0.191 (0.394, 0.830)
Conscientiousness	-0.177 (0.431, 0.830)	0.170 (0.448, 0.830)	-0.076 (0.737, 0.858)	0.315 (0.153, 0.763)	-0.031 (0.890, 0.923)
Neuroticism	0.225 (0.313, 0.830)	0.022 (0.923, 0.923)	0.170 (0.449, 0.830)	-0.049 (0.829, 0.901)	0.271 (0.223, 0.797)
Openness	-0.072 (0.750, 0.858)	-0.215 (0.336, 0.830)	0.071 (0.755, 0.858)	0.323 (0.142, 0.763)	-0.279 (0.209, 0.797)

D Performance Comparison Across Sessions

To explore learning effects, we present a secondary analysis of change trends in performance across sessions. For consistency and to control for potential motivational influences from the SAR, comparisons were restricted to sessions conducted by the same administrator (human or robot). Within-task comparisons involved sessions with the same task version (S1 vs. S5 and S2 vs. S4), and a between-task comparison (S3 vs. S4) assessed performance changes of different task versions after prior exposure to both.

The picture description task demonstrated significant within-task improvements across all performance metrics (see S1 vs. S5 and S2 vs. S4 in Figure 8). Additionally, gradual improvements were observed for *words* and *completeness* across the three robot-administered sessions (S2 to S3 to S4), suggesting general cognitive engagement even between different task versions. In particular, participants used a significantly higher number of words in their descriptions in S4 than in S3. For the semantic fluency task, Figure 9 suggests task-dependent performance on *unique correct words* and the number of *hints* required to complete the task, yet *repetitions* showed a more gradual trend with minimal differences between the three robot-administered sessions.

Overall, this secondary analysis indicates that both learning effects and general cognitive engagement contributed to performance improvements. Future studies could assess within- and between-task performance in greater detail by extending the number of weeks and incorporating additional task variations. Overall, the performance improvements observed between the first and last session (our primary analysis to address RQ2) support the feasibility of SARs as engaging tools to promote cognitive health through repeated robot-administered verbal interactions based on clinically validated tools.

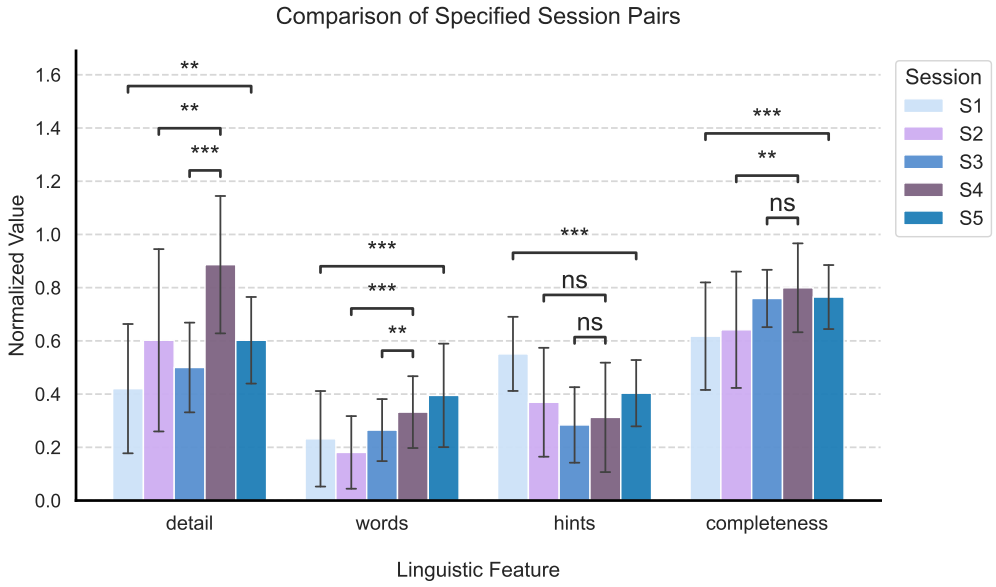


Figure 8: Task performance across the sessions for the picture description task. For consistency, comparisons are limited to sessions conducted by the same administrator (human or robot): within-task (S1 vs. S5 and S2 vs. S4), between-task (S3 vs. S4, after prior exposure to both versions). Refer to Figure 3 for details on the task variations and session structure.

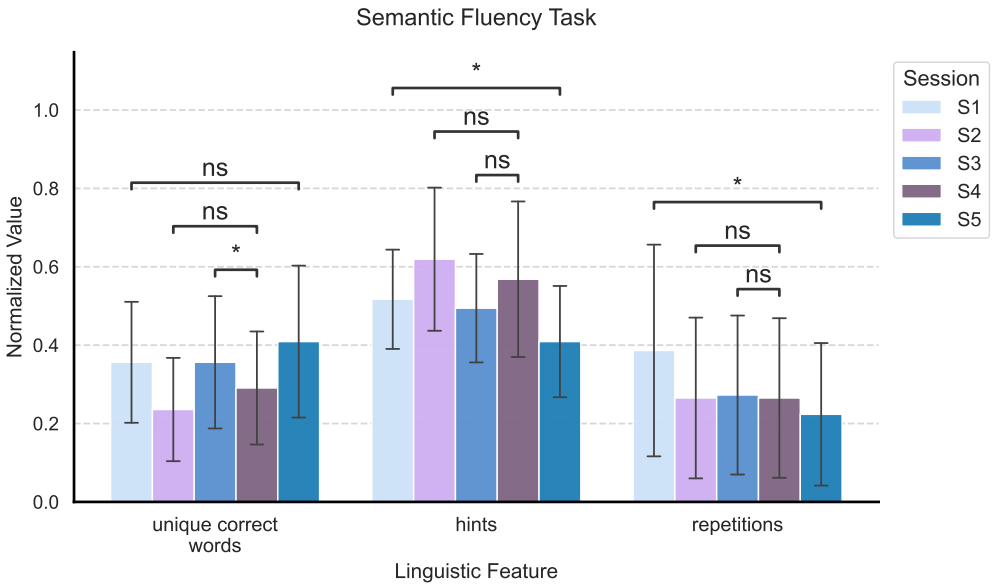


Figure 9: Task performance across the sessions for the semantic fluency task.