OUKA
The University of Osaka
Institutional Knowledge Archive

Title	Focus Group-Led Refinement of an LLM-Enabled Companion Robot for Older People
Author(s)	Satake, Yuto; Nygren, Minna; Yu, Clare et al.
Citation	International Journal of Social Robotics. 2026, 18, p. 67
Version Type	VoR
URL	https://hdl.handle.net/11094/105902
rights	This article is licensed under a Creative Commons Attribution 4.0 International License.
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

International Journal of Social Robotics (2026) 18:67
https://doi.org/10.1007/s12369-026-01407-w

Focus Group-Led Refinement of an LLM-Enabled Companion Robot for Older People

Yuto Satake[1,2] · Minna Nygren[3] · Clare Yu[1] · Akira Tsuboi[4] · Ippei Endo[4] · Kazuhiro Umemura[4] · Kunihiko Katsuki[2] · Daiki Ishimaru[2,5] · Penny Rapaport[1] · Nadia Bianchi-Berthouze[3] · Manabu Ikeda[2] · Robert Howard[1]

Received: 20 October 2025 / Revised: 6 March 2026 / Accepted: 15 May 2026
© The Author(s) 2026

Abstract
Loneliness in later life is a public-health concern, and large language model (LLM)-enabled conversational robots have potential to supplement social connection at scale by facilitating everyday conversations. However, evidence on the development and acceptability of such systems for older people remains scarce. We conducted a two-round mixed methods study with older community-dwelling participants, to test whether pipeline refinements—streaming automatic speech recognition, text-to-speech, and simplified turn-taking cues—improved conversational experience. A tabletop conversational robot (Sota, Vstone), powered by Generative Pre-trained Transformer (GPT)-4o was used. Eighteen individuals participated in round one; paired analyses were conducted on the 17 eligible participants remaining in round two. We collected interaction logs (response latency, LLM-level latency, and deletion rate [proportion of words in the manual transcript omitted in the ASR output]), technical failure events, and participant questionnaires (Godspeed Questionnaire Series; System Usability Scale; a customised five-item scale assessing conversational experience). Qualitative data were analysed using a framework approach. Median response latency decreased from 9.8 to 4.7s, deletion rate from 0.37 to 0.07, accompanied by improved perceived usability and conversational naturalness, while questionnaire indices of companionship changed little. Qualitative accounts converged: interactions felt smoother and more in-context, yet interruptions, perceived superficiality of conversation, and only a modest sense of companionship were still experienced. Findings indicate that improving basic conversational fluency enhances naturalness and usability but is insufficient for delivering truly relational experience outcomes. Progress toward better companionship requires an additional relational layer—shared memories, age-attuned personas, and calibrated reciprocity.

Keywords Companion robot · Loneliness · Large language model · AI · Focus group · Mixed methods

1 Introduction

Loneliness and lack of social connection in later life are associated with adverse outcomes including depression [1], anxiety [2], and increased mortality [3], and are now recognised as major public health concerns. In 2023, the World Health Organization launched the Commission on Social Connection [4], and several countries, including the UK and Japan, appointed ministers for loneliness, highlighting growing policy attention to this issue [5]. Randomised controlled trials (RCTs) and meta-analyses indicate that interventions such as group-based activities and social prescribing, volunteer befriending, telephone and online support, psychosocial interventions such as cognitive behavioural therapy and behavioural activation, and

✉ Yuto Satake
y.satake@psy.med.osaka-u.ac.jp

1 Division of Psychiatry, University College London, London, UK

2 Department of Psychiatry, The University of Osaka Graduate School of Medicine, Suita, Japan

3 UCL Interaction Centre, University College London, London, UK

4 NTT WEST, Inc., Osaka, Japan

5 Department of Medical Technology, The University of Osaka Hospital, Suita, Japan

Published online: 29 May 2026

activity engagement programs can reduce loneliness and related outcomes in the short term [6, 7]. However, challenges remain in terms of training requirements, scalability, reach, and the durability of effects. Against this backdrop, there is increasing interest in scalable digital interventions that can supplement social connection through facilitation of everyday conversation [8].

Conversational agents (CAs) have attracted growing attention as a potential way to address these challenges. Our recent systematic review and meta-analysis suggested that, despite variability and bias across studies, such agents can have beneficial effects on loneliness and depression among older people without dementia. However, most of the included systems were rule-based or domain-limited, raising concerns about limited flexibility and contextual continuity in real-world use [9]. In our prior fieldwork deploying a rule-based conversational robot in older people's homes, participants reported dissatisfaction with delayed responses and superficial dialogue [10]. The advent of large language models (LLMs) has made open-domain conversation increasingly feasible, but early reports emphasise that conversational characteristics such as response timing, the handling of utterance boundaries during turn-taking, and the frequent use of repetitive fixed phrases, critically influence perceived naturalness and trust [11]. These issues are particularly salient in socially situated interactions with older people. Building on this earlier work, we initiated a project using a conversational robot equipped with Generative Pre-trained Transformer (GPT) to alleviate loneliness in older people. GPT is a type of LLM, trained to generate human-like replies by next-token prediction and to adapt to new tasks from natural language instructions [12]. As a first step, we worked with community-dwelling older people without dementia to examine acceptability, identify design challenges, and refine the dialogue design in both the UK and Japan [13].

Building on this groundwork, this study examines whether pipeline refinements to a GPT-4o–enabled conversational robot for older people (shorter response latency and more stable speech recognition), improved perceived conversational naturalness, depth, and companionship in the same participants across two rounds. Our contributions are threefold. First, to our knowledge, this is among the earliest in-situ evaluations of a GPT-4o–enabled conversational robot with community-dwelling older people. Second, we integrate distributional, log-based conversation metrics with qualitative accounts. Third, we link improvements in basic conversational fluency to user experience while highlighting limits in relational outcomes.

2 Related Work

2.1 Conversational Agents for Loneliness in Older People

2.1.1 Background and Rationale

Digital interventions addressing loneliness grew rapidly between 2010 and 2022, with companion robots and conversational agents (CAs) forming a distinct cluster [14]. However, most studies have been short-term, small-scale, and concentrated within institutional settings, with limited attention to cost-effectiveness, adverse effects, or longer follow-up. Systematic reviews report small-to-moderate benefits on measures of loneliness but note that many trials are non-randomised or single-arm, carrying high risk of bias [9, 15]. Broader mental health meta-analyses similarly suggest short-term improvements in depression, anxiety, stress, and quality of life, while emphasizing heterogeneity, limited durability, and the importance of personalization and empathic responses [16].

2.1.2 Chatbots, Personal Voice Assistants (PVAs) and Screen-Based Embodied Conversational Agents (ECAs) (Including Digital Humans)

Evidence specific to chatbots for loneliness in older people remains sparse [17]. For example, a four-week study in psychiatric outpatients aged ≥ 65 reported reduced loneliness with high engagement; however, the participants were not severely lonely at baseline, which may limit the generalizability of such findings to lonely older adults [18, 19]. The rise of LLM-integrated chatbots (e.g., ChatGPT) may facilitate more rigorous studies; nonetheless, extrapolation from general populations to older people should be made cautiously.

In community-dwelling older people living alone, video-enabled PVAs yielded larger reductions in loneliness and improved perceived social support more than audio-only versions over 12 weeks [20]. Another study in adults aged ≥ 75, found that total time of engagement, rather than input frequency, predicted loneliness reduction, suggesting a potential "dose–time" relationship [21]. Whether this relationship holds for LLM-empowered PVAs—where conversational depth and variability are higher—remains to be tested.

Screen-based ECAs have been evaluated in both community and institutional contexts. Randomised feasibility work indicates general good acceptability among independently living older people, but no consistent short-term effects on loneliness in care facilities [22]. Proactive ECAs that initiate conversation appear more effective than passive ones

in home settings, implicating agent proactivity as a design lever [23].

2.1.3 Robots

Socially assistive robots (SARs) have been widely studied in long-term care settings. A 2024 meta-analysis of randomised trials reported reductions in depression and loneliness among institutionalised residents [24]. However, many of the robots studied, such as PARO [25], are non-conversational, and their effects cannot be directly generalised to conversational systems. Reviews in dementia care emphasise feasibility and acceptance but highlight mixed findings for cognition, neuropsychiatric symptoms, and quality of life [26]. In community settings, early studies of rule-based conversational robots (e.g., doll-type systems), suggest short-term benefits, such as reduced depression at four months, and demonstrate feasibility for extended use [27, 28], yet a six-month mixed methods study found that benefits had attenuated by 6 months, highlighting novelty effects and challenges in everyday use [29]. Users also reported one-way communication and operational issues as barriers; these implementation hurdles are likely to generalise across platforms, particularly in populations with varying levels of digital literacy.

2.2 Use of Human–Robot Interaction in Older People

2.2.1 Settings and Implementation

Alongside growth in interventions, recurring implementation barriers are documented—digital access, connectivity, affordability, and governance/privacy [14]. A randomised, waitlist-controlled pilot trial evaluated a screen-based digital human that delivered a one-week loneliness and stress programme of about fifteen minutes per day to at-risk adults. Feasibility and acceptability were demonstrated among older people living independently, whereas delivery in nursing homes was less feasible and typically required caregiver assistance [22]. These constraints apply equally to robots; initial setup, onboarding, and connectivity often determine sustained engagement.

2.2.2 Embodiment, Copresence, and Companionship Cues

Across controlled comparisons, copresence—defined as sharing the same physical space that enables active, reciprocal perception between human and agent—tends to increase trust, engagement, and willingness to follow suggestions relative to screen/telepresence agents; nonverbal behaviours can amplify these effects [30]. Among older people, perceived social presence is strongly associated with both enjoyment and intention to use robots [31], and physically embodied robots elicit greater conversational engagement than on-screen agents when the dialogue proceeds successfully [32]. These embodiment and copresence effects matter because engagement time is associated with loneliness reduction in adjacent modalities [21].

Companionship perceptions depend partly on appearance and behaviour. Older people tend to prefer familiar, friendly designs and "aliveness" cues (gaze, blinking, breathing-like motion, warmth), with brief, clear utterances and consistency without monotony [33]. Abstract/novel designs can be seen as childish. While these features foster social presence and acceptance, which conversational capabilities strengthen companionship remains under-specified, warranting further study.

2.2.3 Dialogue Mechanics Tailored to Older Users

Conversation design should consider sensory/cognitive characteristics: clear speech, adjustable speech rate, and robust error-repair (repetition/clarification), while avoiding excessive repetitiveness [23, 31]. Automatic speech recognition (ASR) accuracy can decline with age [34, 35]; in care-facility robot studies, word error rates (WER; the proportion of substitutions, deletions and insertions divided by the number of reference words) as high as 0.78 have been reported [36]. Practical designs should anticipate ASR failures and provide visible turn-taking cues and repair strategies.

2.3 The Promise of LLMs and Advanced ASR/TTS for Social Robotics

2.3.1 LLMs in Conversation Content

LLMs enable flexible, context-aware dialogue beyond scripted systems, with advantages reported in richness, perceived empathy, and some reasoning tasks [37–39]. These findings, however, come largely from non-clinical, younger samples and from task-specific evaluations. Real-time social dialogue also imposes tight timing constraints (typical human gaps < 300 ms) [40, 41]. While recent models (e.g., GPT-4o) can yield low audio-to-response latencies under optimal conditions [42], end-to-end timing in robots depends critically on ASR/TTS pipelines and Input/Output (I/O) orchestration.

2.3.2 Advances of ASR and TTS

Deep learning markedly reduced ASR error rates and improved robustness to noise and accent [43, 44].

RNN-Transducer (RNN-T) style architectures can enable practical end-to-end streaming, while modeling output dependencies [45, 46]. Self-supervised pretraining (e.g., wav2vec 2.0, HuBERT) and large multilingual models (e.g., Whisper), further improve accuracy and stability for low-latency streaming use [47–49]. In TTS, Tacotron-based seq2seq improved naturalness but was slow due to autoregressive decoding [50]. Non-autoregressive Transformer models (e.g., FastSpeech) and neural vocoders (HiFi-GAN, Parallel WaveGAN) can now support near real-time synthesis with high quality [51–53].

2.3.3 Use of LLMs with Advanced ASR/TTS for Companion Robots

LLM-based robots offer greater flexibility and contextual adaptation than scripted systems, improving perceived naturalness and enjoyment. While task efficiency is often comparable; the trade-off is higher latency/energy and the need to manage failure modes (repetition, premature termination) [11, 54]. Integrating gaze/speech cues with LLM inference can improve responsiveness and bidirectional smoothness [55]. Embodiment and I/O pathways matter: studies report higher failure rates in voice-only conditions than text or embodied agents, underlining the centrality of input–output design [56]. Evidence specific to older people remains limited and constitutes a key gap that this study addresses.

2.3.4 Concerns in Using LLMs

Systematic analysis highlights risks with LLMs—hallucinations, opacity, bias, privacy—and emphasises risk-calibrated human oversight [57]. A case study further shows an LLM-equipped care robot misrepresenting capabilities (e.g., medication reminders) and approving unsafe drug combinations [58]. These concerns are especially salient for older people and argue for transparency, data minimization, fail-safes, and escalation pathways by design.

2.4 Summary

Across modalities, conversational agents show short-term potential to reduce loneliness/depression in older people, but the evidence base is constrained by short follow-up, small samples, limited controls, and lack of blinded outcomes. For deployments with older people, implementation should be tailored to users' abilities and usage contexts, and embodied robots may offer an advantage through the benefits of co-presence. While approachable, user-friendly design appears important, it remains uncertain which kinds of dialogue best foster a genuine sense of companionship; age-related reductions in ASR accuracy must also be considered. Against this background, combining LLMs with advanced ASR/TTS offers the possibility of moving beyond scripted systems to achieve practical conversational fluency and continuity of context on robots. Nevertheless, evidence specifically involving community-dwelling older people using LLM-enabled robots remains scarce.

3 Study Design

This was a predominantly qualitative mixed methods study, examining whether improvements to the support of conversational dynamics in an LLM-enabled conversational robot, specifically response latency, speech recognition quality, voice tone and LED turn-taking indicators, would improve older peoples' experiences of social interaction with the robot. We focused on experiential outcomes most likely to be influenced by these changes, principally, perceived conversational naturalness, conversational depth/richness, and sense of companionship.

The study integrated qualitative focus group interviews, quantitative UX measures, quantitative dynamic system performance metrics and system-level LLM interaction metrics. We conducted two rounds of focus groups (four sessions per round): the first round captured impressions of an original system, and the second round evaluated an improved version (updated Sota and dialogue system). 18 people participated in the first round and 17 of the same participants in the second round, one participant having been removed as she was found to have been ineligible for the study.

Qualitative data were analysed using framework analysis [59], enabling structured comparison across the two rounds while remaining open to inductive insights. Quantitative measures (questionnaire scores and interaction metrics) were collected in both rounds and compared within participants to complement and contextualise the qualitative findings.

4 Robot and Dialogue System: Original Setup and Modifications

The robot used in this study was Sota, a compact humanoid communication robot developed by Vstone (Osaka, Japan. Supplementary Fig. S1). Standing 28 cm tall, Sota features a microphone array with noise cancellation, an internal speaker, LED indicators embedded in its eyes and mouth, and capable of basic gestures such as nodding and arm gestures.

The dialogue system was designed as a modular pipeline, consisting of automatic speech recognition (ASR), an LLM

to generate appropriate responses, and speech synthesis (Fig. 1A). User speech was captured by the robot's microphones and processed by a cloud-based ASR engine, after which the transcribed text was sent to GPT-4o via Azure OpenAI. Responses were then synthesised into speech processed by a cloud-based TTS engine and played through the robot. This enabled real-time, user-initiated conversation, guided by a structured prompt that defined the robot's role as being a friendly companion for older people. This primarily user-initiated interaction model (with no robot-initiated proactivity in the present implementation) was intentionally adopted for this early-stage, focus-group-based formative evaluation, to prioritise stabilising the conversational pipeline and interaction flow before introducing proactive behaviours that could disrupt group discussion. To initiate dialogue, the experimenter pressed a button on the back of the robot. The robot's eye LEDs signalled readiness to receive input, helping participants recognise when to begin speaking.

Following the first round of focus groups, it became evident that the robot's slow response time in conversation was a prominent barrier to engagement and acceptability. Therefore, the robot's system was modified to facilitate faster responses. The robot's internal processor was upgraded from Intel Edison to Raspberry Pi 4 Model B to enhance processing speed. The TTS and ASR engines were migrated from services of NTT TechnoCross to Azure AI Speech of Microsoft, enabling streaming communication and reducing latency (Fig. 1B). To simplify turn-taking cues and to reduce cognitive load, the robot's eye LEDs were set to remain blue throughout interaction. The synthetic voice was also changed to that of a British English child's voice ("Maisie") from the Azure Speech Studio gallery (Microsoft), as migration to the Azure-based speech synthesis platform necessitated the use of a new voice template. To accommodate the new processor, the base of the robot was slightly enlarged using a 3D-printed enclosure (Table 1). Changes were prioritised to address interaction-breaking bottlenecks identified in the first round (especially latency and ASR-related breakdowns), to be feasible within the study timeline, and to avoid design choices that could disrupt focus-group discussion dynamics. All modifications were implemented by the development team at NTT West.

Further technical details, system architecture, and implementation notes are provided in the Supplementary Note S1.

5 Methods

5.1 Participants and Recruitment

Participants were recruited between January and March 2025 via direct outreach to individuals who had previously taken part in ageing or dementia-prevention studies, and through participant-led referrals. Inclusion criteria were: (1) age 60 years or older (2), living independently in the community (3), able to provide written informed consent, and (4) sufficient English proficiency to participate in a focus group. Exclusion criteria were: (1) diagnosed with dementia (2), uncontrolled physical illness.

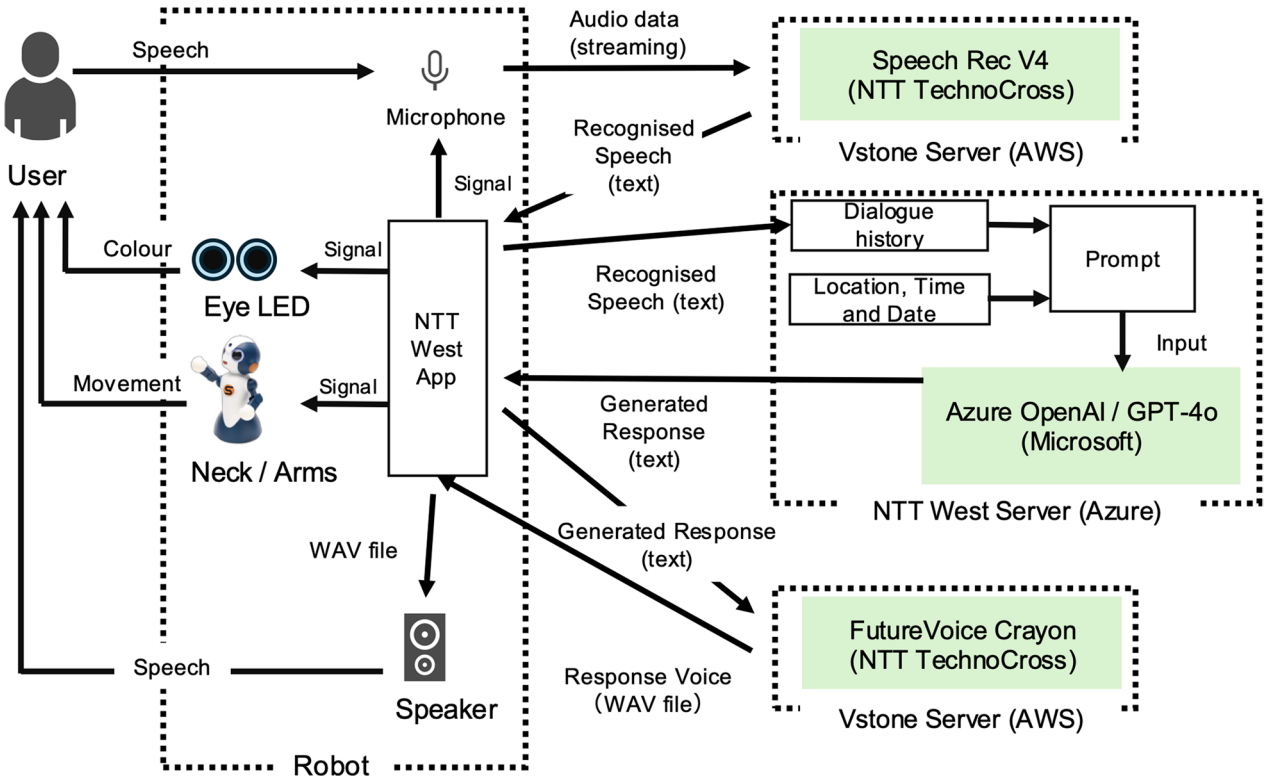
Eighteen participants (aged 49–86) took part in the first-round focus groups (February–March 2025). One participant aged 49 (ID: P4) was inadvertently included due to a scheduling error and was excluded from second-round sessions. In June 2025, all eligible first-round participants ($n=17$, aged 61–86) were contacted and agreed to take part in the second round. Consequently, the analytic sample for paired comparisons comprised the 17 eligible participants. The mean age was 72.5 years (SD = 7.1), and the average years of education was 14.6 (SD = 3.4). The sample included 10 women and 7 men, and 10 participants were living alone. All participants owned a smartphone; in addition, 14 owned a personal computer, 13 had a tablet device, and 6 reported owning a conversational agent, including PVAs. Three participants (P1, P12, and P16) reported hearing problems; two used hearing aids. P1 appeared to require clearer, more deliberate speech for everyday conversation. To minimise potential impacts on ASR/TTS timing and interaction flow, sessions were conducted in a quiet room, participants were seated close to the robot, and the robot's speech output volume was adjusted when needed. Further demographic details are presented in Supplementary Table S1.

5.2 Procedure and Data Collection

Four focus group sessions were held in each round at University College London (UCL), typically with 4–6 participants per session; in the second round, two sessions had 3 participants due to last-minute cancellations. The overall structure and number of sessions were determined based on best practice in qualitative focus group design, which emphasises conceptual depth and data adequacy over statistical generalizability [60, 61]. First-round sessions lasted approximately two hours (including a short break), and second-round sessions around 90 min. A photograph of a typical focus group setting is provided in Fig. 2 to illustrate the study environment and arrangement.

In both rounds, participants were seated in a semicircle facing the robot, which was connected via a portable Wi-Fi

A. first round (baseline pipeline)



B. second round (updated pipeline)

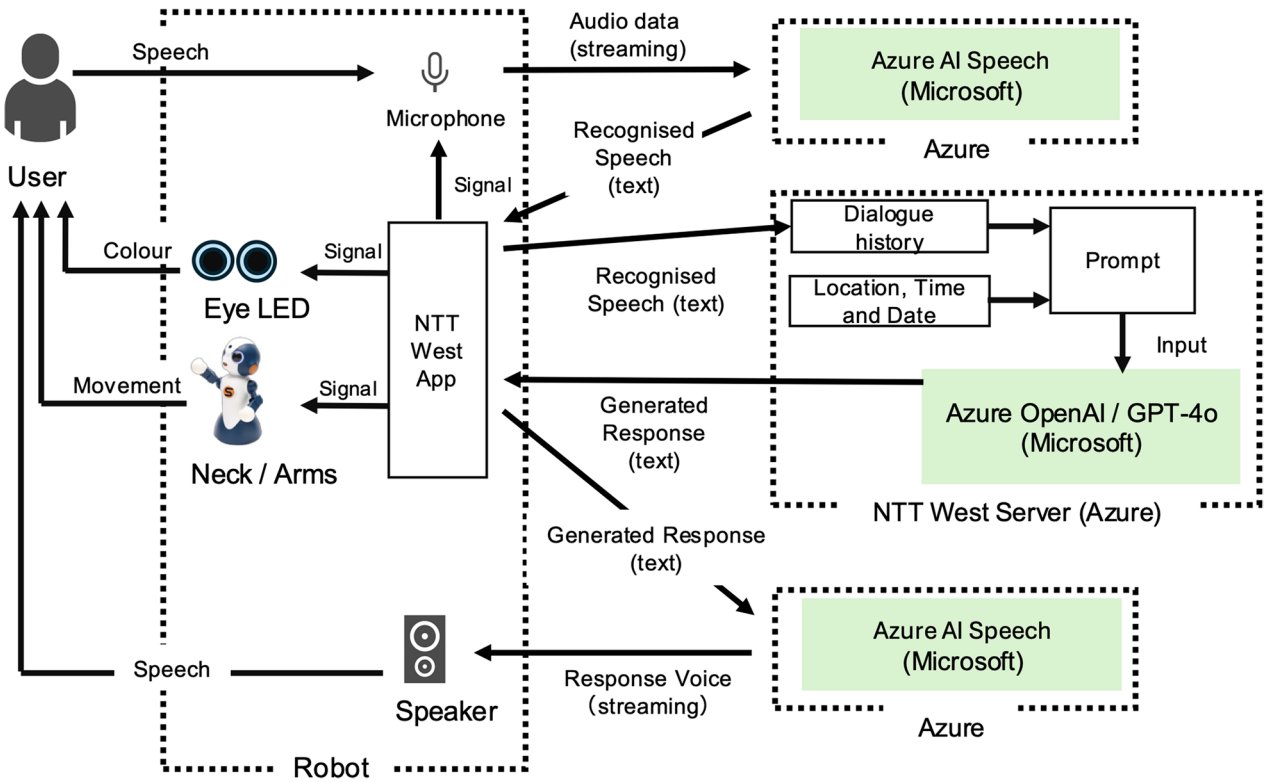


Fig. 1 Dialogue system pipeline (first round vs. second round). Green boxes denote ASR, TTS, and text generation services; arrows indicate data flow. (**A**) first round (baseline): User speech was transcribed by Speech Rec V4 (NTT TechnoCross) and processed via the NTT West server, with responses generated by Azure OpenAI (GPT-4o) and synthesised by FutureVoice Crayon (NTT TechnoCross) as WAV files. (**B**) second round (updated): Both automatic speech recognition (ASR) and text-to-speech (TTS) were migrated to Azure AI Speech (Microsoft, streaming), with dialogue still routed via the NTT West server and responses generated by Azure OpenAI (GPT-4o). LED cues were simplified (constant blue) and the output voice changed to "Maisie."

hotspot. Network speed was tested prior to each session, though upload bandwidth occasionally fell below 5 Mbps. Sessions were facilitated by YS with at least one other team member present as co-moderator. In one second-round session, YS joined remotely due to illness while on-site facilitation and robot operation were handled by another researcher. Two semi-structured interview guides were developed a priori—one per round—based on the study aims and qualitative best practice. The first round used a discovery-oriented guide to elicit older people's expectations and concerns about using an AI companion robot for loneliness through hands-on interaction. The second round used a distinct, comparison-focused guide to appraise the updated system against the earlier version, with prompts targeting response speed, conversational naturalness, depth/richness, perceived companionship, and overall satisfaction; the same round-specific guide was used in all sessions of that round. Sessions began with a demonstration of the robot's capabilities, followed by structured and unstructured interactions. Audio recordings (and video recordings where consent was obtained) were made using a dedicated digital recorder and Microsoft Teams under a UCL institutional licence. System-level interaction logs (e.g., speech recognition outputs, generated text, LLM timestamps) were recorded on a secure NTT West server. Occasional field notes were taken during or immediately after sessions.

In first-round focus groups, all participants engaged in at least one individual interaction with the robot, followed by group discussions. Interactions included free-form dialogue and three predefined emotional scenarios (neutral, boredom or grief). Two different voice settings (default and lower-pitched/slower-tempo) were trialled, with the preferred voice used for subsequent interactions. Topics included perceptions of conversational quality, perceived shortcomings that emerged, and potential uses of AI robots to address loneliness in older people.

In second-round focus groups, the sessions followed the same general format but were centred on direct comparisons between the improved robot and its earlier version. The updated system incorporated prompt and dialogue management changes aimed at improving response latency, speech recognition, and conversational engagement. The semi-structured guide emphasised three key interactional dimensions—conversational naturalness, conversational depth or richness, and perceived sense of companionship—selected because first-round participant feedback indicated that improvements in response speed were most likely to influence these aspects of conversational quality and relational experience. Additional prompts explored overall satisfaction and perceived changes across versions.

5.3 Outcome Measures

Outcomes comprised post-session UX questionnaires and log-/audio-derived performance metrics. Questionnaires were administered after the focus-group discussion to capture session-level evaluations based on the complete interaction and the overall session experience. Participants completed the Godspeed Questionnaire Series (GQS) to capture subjective impressions of the robots. The scale comprises five subscales—Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety—rated on five-point semantic differentials; we analysed subscale means, with higher scores indicating more positive impressions [62]. Usability was measured with the ten-item System Usability Scale (SUS), adapted by replacing "system" with "robot" consistent with prior HRI studies that contextualise SUS item wording to the target artefact [63]. SUS scores were converted to the standard 0–100 metric, where higher values indicate better perceived usability [64]. We report internal consistency (Cronbach's alpha) for the adapted SUS items separately for first and second rounds. Both instruments have widely reported validity and reliability in HRI and usability research. In the second round, participants also completed a customised five-item scale (Quickness, Naturalness, Depth/Richness, Satisfaction, Companionship) to rate perceived changes in the updated robot (1 = "strongly disagree" to 5 = "strongly agree"). First-round questionnaires also included demographic items.

For log-based metrics, response latency was defined as the interval between the end of the participant's item of speech and the start of the robot's verbal response. LLM-level latency was defined from system log timestamps as the time from sending the transcribed speech item to the LLM to receiving the generated text. Speech recognition accuracy was quantified using Deletion Rate (DR), defined as the proportion of reference words in the manual transcript omitted in the ASR output (DR = D/N, where N is the total number of reference words and D is the number of omitted words). Technical difficulty events were also counted and categorised according to Irfan et al. [11]: turn-taking violations, repetitive responses, hallucinations, empty responses, and premature termination due to misinterpreting terminating cue phrases.

Table 1 Summary of specification changes in the improved robot

Category	Change	first-round issue addressed	Basis	Rationale
Internal processor	Intel Edison→Raspberry Pi 4 Model B	Slow response time	Design inference (participant-reported problem)	Enhance processing speed and overall responsiveness
Speech recognition	Speech Rec V4 (NTT TechnoCross, Streaming)→ Azure AI Speech (Microsoft, Streaming)	Insufficient speech recognition	Design inference (log-informed and participant-reported problem)	Improve recognition robustness and reduce breakdowns (and secondarily latency)
Speech synthesis	FutureVoice Crayon (NTT TechnoCross, WAV file download)→ Azure AI Speech (Microsoft, Streaming)	Slow response time	Design inference (participant-reported problem)	Reduce latency and improve conversational flow
LED eye cues	Dynamic colour changes during turn-taking→Constant blue throughout interaction	Disjointed interaction	Participant suggestion	Simplify cues and reduce cognitive load
Synthetic voice	Default child-like voice (Vstone)→ British English child-like voice "Maisie" (Azure Speech Studio)	Insufficient speech intelligibility	Design inference (participant-reported problem)	Increase perceived naturalness and intelligibility
Physical casing	Original base (factory-made)→ Slightly enlarged base (3D-printed enclosure)	Not a user-facing issue	Technical constraint	Accommodate upgraded processor

All entries summarise changes implemented between the first round and the second round. "first-round issue addressed" reflects the primary issue motivating each change (participant-reported and/or log-informed). "Basis" distinguishes changes driven by explicit participant suggestions versus author design inference (informed by participant-reported problems and/or log-based evidence) versus technical constraints. "Streaming" denotes incremental processing/output for ASR/TTS. Abbreviations: ASR, automatic speech recognition; TTS, text-to-speech



Fig. 2 Study setting and focus-group interaction with the conversational robot. The photograph illustrates the typical arrangement used in the sessions. The individual on the left is the researcher, and participants' faces have been anonymised by applying white rectangular masks

5.4 Data Analysis

Response latency was measured for the second user–robot exchange in each interaction by YS from audio recordings using ELAN 7.0. LLM-level latency was calculated from system log timestamps. For DR calculations, manual transcripts were compared with ASR outputs after excluding filler words (e.g. uh, um, er, ah, eh, oh, yeah, mm-hmm, huh, well, you know, I mean). Substitutions and insertions were not counted because overlapping and interruptive speech made their classification ambiguous; deletions were counted by comparing words between the transcript and the ASR output while ignoring word order. Conversation episodes that were terminated after the first user–robot exchange did not contain a second exchange and were therefore excluded from these analyses. Technical difficulty events were coded

by YS from system logs and audio recordings following Irfan et al. [11]

Response latency, LLM-level latency, DR, GQS and SUS scores from the first and the second rounds were compared using Wilcoxon signed-rank tests. To address multiple comparisons, Benjamini–Hochberg (BH) correction was applied within two pre-defined families of tests: (A) system-performance metrics (response latency, LLM-level latency, DR) and (B) UX measures (GQS subscales and SUS). Two-tailed tests were used, and results are reported with BH-adjusted q-values ($q < 0.05$). Custom impression items were analysed descriptively.

Exploratory association analyses were conducted to examine whether participant-level changes in log-based metrics were associated with changes in subjective user experience. Change scores were defined as post minus pre. Spearman rank correlations were computed between changes in system-performance metrics (response latency and DR) and changes in subjective measures (SUS and selected GQS subscales: Perceived Intelligence and Likeability). To limit multiplicity, six correlations were pre-specified and BH adjustment was applied within this correlation set; BH-adjusted q-values (two-tailed) are reported.

All tests and plotting of metrics data were performed using Microsoft Excel and R (version 4.5.1). Calculations for DR were implemented in Python (v3.10) using pandas, with scripts and logs archived for reproducibility.

Verbatim transcripts from both rounds were analysed in NVivo (versions 14 and 15), using framework analysis. Analysis proceeded through familiarisation, development of the analytic framework, indexing, charting, and mapping/interpretation. Following familiarisation with the first-round data, before the collection of second-round data, we developed the three key interactional categories—conversational naturalness, conversational depth and richness, and perceived sense of companionship— judged by the research team to best capture participants' concerns and most likely to be influenced by improvements in response speed and related system refinements. For the purposes of analysis, we used the following working definitions. Conversational naturalness captured the extent to which the interaction resembled a human-like flow of conversation, including smooth timing, coherent turn-taking, and continuity across turns. Conversational depth captured the extent to which the conversation went beyond a surface exchange, including the ability to explore a topic further as well as the presence of subjective viewpoints, unexpected or non-obvious responses, and a sense of substance or nuance. Conversational richness referred to the amount, specificity, and variety of information provided in responses. Perceived companionship referred to the extent to which the robot was experienced as a companion-like social presence rather than a purely functional device (e.g., feelings of connection and reciprocity). They were therefore adopted as the analytic framework for data from both rounds, with scope left for inductive additions (none were ultimately identified). Transcripts were coded in NVivo and charted into a case-by-category matrix to enable within-person, cross-round comparison. Coding and interpretation were conducted by the research team collectively, with regular discussion and consensus-building meetings.

Descriptive statistics for the first round (e.g., session time, episode counts) were based on data from all 18 participants. However, data from the ineligible participant (P4) were excluded for the calculation of response latency, LLM-level latency, and DR. Both inferential and qualitative analyses were conducted using the data from the 17 eligible participants.

6 Results

6.1 Qualitative Analysis

Participants' reflections on their interactions with the robot were analysed in three interrelated categories: naturalness, depth/ richness, and companionship. Within each category, accounts from the first and second rounds were compared, and changes within the same individual were noted where present. These categories were pre-specified rather than derived inductively. Representative quotations are presented to illustrate typical or particularly salient perspectives within each category.

6.1.1 Naturalness

In the first round, participants described the interactions as slow, disjointed, and lacking in contextual continuity:

> …the only trouble, sleepy… when you said, oh, you know, I'm buried, the natural human response is "What?" "What about?". And it was far too long. I would end up… I'm getting more pleasure out doing that [elsewhere]. (P2, 1st)

> I still think that I'm talking to a machine that doesn't really understand what I'm saying. And it's saying more like you said. Scripted responses. (P14, 1st)

These accounts suggest that long response gaps and weak contextual continuity disrupted turn-taking expectations and reduced the sense of being heard. Such experiences may reflect constraints across multiple parts of the interaction pipeline. Participants sometimes compensated by supplying

their own follow-up responses, indicating that latency could shift effort from the robot to the user and undermine perceived conversational flow. In the second round, participants described smoother, more natural exchanges, while also noting that uneven timing and interruptions persisted:

> I think it was more natural, I mean, yeah, there are lots of reservation(s). (P2, 2nd)

> More natural in terms of just all the dynamics and the speed. But actually the responses that like he [robot]'s more in context it seems. (P14, 2nd)

> The timing was a bit jerky… I was trying to speak, and then he interrupted… It's not a conversation, really. It's a question and answering. (P11, 2nd)

Together, these comments indicate that the refinements were perceived as improving interaction speed and the contextual fit of responses. However, faster replies did not necessarily result in smoother turn-taking, as interruptions and uneven timing could still disrupt the flow and limit perceived conversational naturalness. One participant also noted that the robot often repeated the same fixed phrase at the start of an exchange, saying "I'm here to help," which made it feel as though a new conversation was beginning rather than continuing from the previous turn.

Overall, the improvements were generally perceived as enhancing contextual continuity and the smoothness of exchanges, thereby contributing to greater conversational naturalness. Nevertheless, participants pointed out that elements undermining naturalness persisted, including turn-taking interruptions and the recurrence of fixed or repetitive phrases.

6.1.2 Depth/ Richness

In the first round, participants described the conversation as tending to remain on the surface, using terms such as "diplomatic" and "superficial," and noting that responses could feel formulaic or "generic"; some also expressed dissatisfaction with the amount of information provided:

> I think it's very diplomatic." "I mean, I suppose with this robot is quite good to have it with you just for a very superficial conversation. (P9, 1st)

> But sometimes his responses are a bit generic, especially with some of the answers about, you know, references to things (P14, 1st)

> It [robot] doesn't have enough information behind it for me… I would appreciate something that would give me any information at any time. (P5, 1st)

Taken together, these comments suggest that the interaction did not consistently meet participants' expectations for a more engaging exchange. In the second round, participants acknowledged notable improvements in the level of detail and variety in responses:

> He or she (robot) then (at the first round) was very hesitant, slow to respond and very limited suggestion and conversation. But this time I'm quite impressed because she has suggested a lot of things I could do, which I didn't think about and it's quite, quite, very logical. (P9, 2nd)

> They seem to have much more information available, which was good. (P13, 2nd)

These comments indicate that participants noticed an increase in informational content and option generation. However, improvements in depth tended not to be recognised or valued:

> It's still superficial. That's why I'm thinking that if we were drilling down to find out whether it could really help us make decisions, that's then becomes more deeper and we didn't do that yet. (P14, 2nd)

> very superficial" "It doesn't deep enough. Yeah, but it gives you a suggestion. It's quite a lot of suggestions. (P17, 2nd)

Taken together, these accounts suggest that the interaction often broadened into multiple suggestions without deepening around a single topic. Participants speculated that the reason might be the robot's lack of knowledge about the user or inability to access sufficient information to deepen the conversation:

> It doesn't have access to enough information to be a conversation. (P5, 2nd)

> I think it can only be superficial if it doesn't know you, and it cannot read your surroundings or you as a person. (P17, 2nd)

Overall, participants generally perceived the improved system as providing more varied and detailed responses, which in some cases enhanced the perceived richness of the conversation. Nevertheless, the interaction was often still

viewed as lacking genuine depth, attributing this to insufficient background knowledge about the user and limited information available for delivery.

6.1.3 Companionship

In the first round, participants positively perceived the robot functioning as a conversational partner:

> It makes you feel like you're talking to somebody. (P1, 1st)

> I think it's all right. It's like a partner. Like keeping you busy, keeping you focused. (P3, 1st)

These positive accounts framed companionship primarily as having someone to talk to and as momentary engagement during the session. In contrast, other participants evaluated companionship against a higher relational standard—what would be required for a genuine companion—and therefore judged the robot as falling short:

> It worries me that it looks a little bit too much like a toy. (P5, 1st)

> I think it was reading the script you programmed: "Oh, I'm sorry to hear all that." I mean, that's sort of typical psychological talk basically: "Hmm, the situation wasn't well." It's a machine… It didn't feel sincere. Genuine. It didn't really. (P18, 1st)

These critiques framed companionship as more than functional conversation. Toy-like appearance and scripted-sounding responses were described as cues of limited sincerity, making it difficult for some participants to view the robot as a genuine companion. In the second round, in contrast, a small set of accounts described a stronger sense of companion-like connection:

> I left feeling a bit, you know, it was just a robot toy [about the first round]. This is not… get less of a feeling that it's a toy [about the second round]. (P7, 2nd)

This suggests that improved interaction fluency can, for some participants, shift the robot from "toy-like" toward a more acceptable conversational presence. Nevertheless, for others, the companionship judgment appeared anchored in whether the robot could demonstrate reciprocity, continuity of shared history, and a sense of agency, which were not substantially altered by the technical refinements:

> I think I used the word conversation, but it couldn't really have a conversation not like human. So, if I could say I was a person who couldn't go out and didn't have any threats, I don't think it would be good enough for me. (P11, 2nd)

> I think it more as a tool companion to ask questions and get responses. There's no emotion there. It's straightforward, which you would expect from a machine. (P12, 2nd)

> It's like I said, randomised… it's not real then. With friends, they already know your history… so what they're saying is, you know… (P14, 2nd)

> It's like being in an echo chamber… it will always agree and say, 'oh, that's a good choice.' A real friend puts their personality into it… they'll come up with some reason why they don't like something. (P14, 2nd)

In addition, conditional or context-dependent views were also expressed:

> In my current situation, I don't think I need it, but for someone that is lonely… The answers you get to your questions is good. So, I think it would work for someone that needed it. (P10, 2nd)

> Because a lot of the responses are all the same… If I had one, I had a Sota and a friend had a Sota… we're getting the same responses… it wouldn't seem as if it was meeting the need. (P13, 2nd)

Overall, a smaller set of accounts reported a stronger sense of connection with the robot following the improvements, describing it as more engaging or more useful as a companion. Some also acknowledged its potential value for specific groups, such as individuals experiencing loneliness. However, the majority of those who had been skeptical in the first round maintained their negative stance, citing the robot's continued lack of emotional reciprocity, shared history, and individual agency, which they regarded as essential for genuine companionship.

6.2 Quantitative Analysis

Table 2 summarises the interaction data. In the first round, the total session time was 450.0 min, with 75.9 min of active conversation over 27 episodes; in the second round, session time was 327.5 min, with 79.8 min of active conversation over 40 episodes. The number of turns per minute during

Table 2 Interaction data

	First round	Second round	W	q	*r*
Session, n	4	4		–	–
Conversation episodes, n	27	40		–	–
Recorded time, min	450.0	327.5		–	–
Interaction time, min	75.9	79.8		–	–
Turn-taking, n	141	210		–	–
Text exchange with LLM, n	253	336		–	–
Response latency, sec	9.8 (5.5)	4.7 (0.9)	0	**<0.001**	0.88
LLM-level latency, sec	1.2 (1.4)	1.0 (0.2)	37	0.064	0.45
DR	0.37 (0.19)	0.07 (0.05)	1	**<0.001**	0.87
Turn-taking violation	19.9% [13.6–27.4]	9.5% [5.9–14.3]		–	–
Repetitive response	0.0% [0.0–2.6]	1.9% [0.5–4.8]		–	–
Hallucinations	2.8% [0.8–7.1]	1.4% [0.3–4.1]		–	–
Empty response	0.0% [0.0–2.6]	0.5% [0.0–2.6]		–	–
Mistaking termination phrase	0.0% [0.0–2.6]	2.4% [0.8–5.5]		–	–

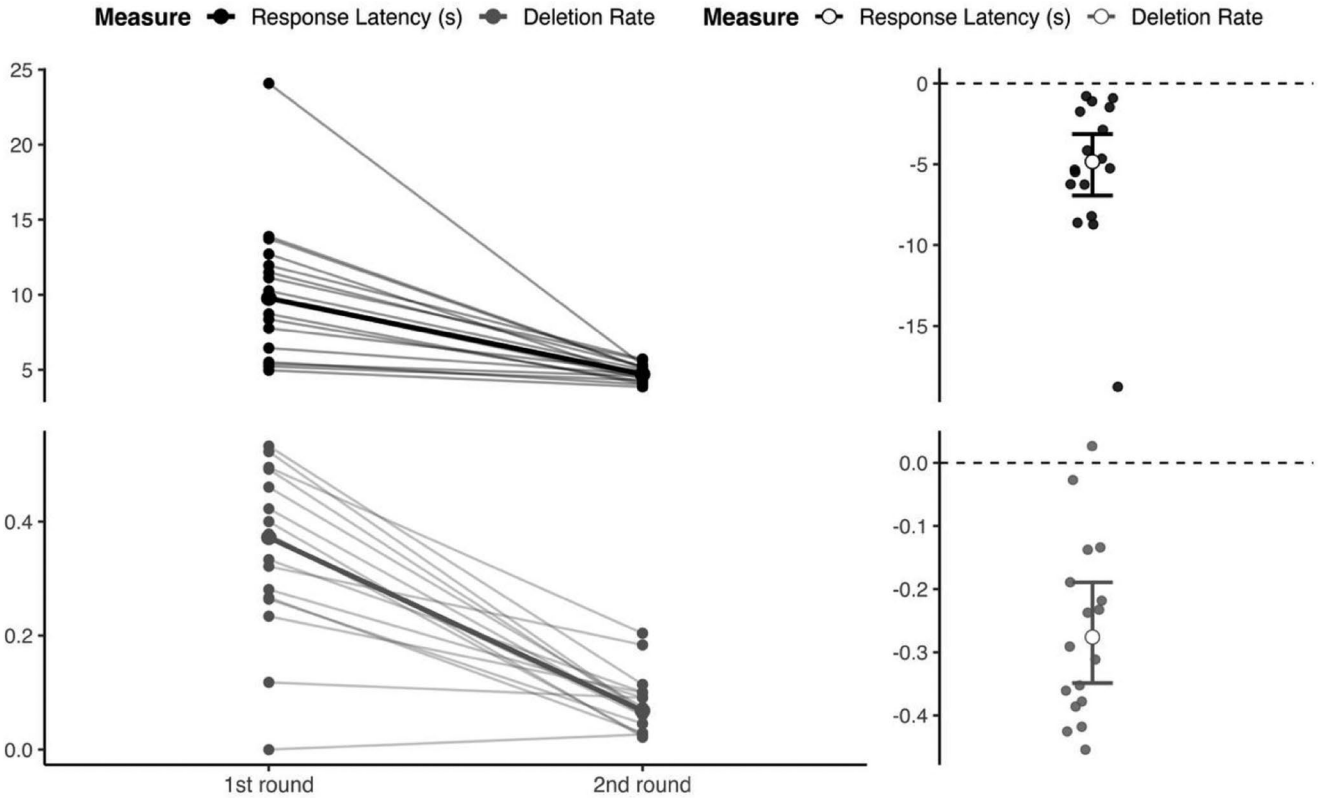


Fig. 3 Participant-level paired changes in response latency and deletion rate. Left-sided panels show participant-level medians for each round; dots are individual participants and lines connect the same participant across rounds. Right-sided panels display the paired differences (Post − Pre) for each participant. The open circle marks the paired Hodges–Lehmann median difference, and the horizontal bar indicates its 95% confidence interval; the dashed line at 0 denotes no change. Negative values indicate reductions (improvements)

active conversation episodes increased from 1.86 to 2.63. Median response latency at the participant level decreased markedly from 9.8 s (IQR = 5.5) to 4.7 s (IQR = 0.9), while median LLM-level latency changed from 1.2 (1.4) to 1.0 (0.2) sec without significance. Median DR also significantly decreased from 0.37 (IQR = 0.19) to 0.07 (IQR = 0.05). Response latency and DR are visualised by the paired plot (Fig. 3), which shows each participant's pre–post change together with the Hodges–Lehmann paired estimates. The Hodges–Lehmann estimates for post–pre were negative and their 95% CIs did not cross zero, indicating consistent reductions across participants. These estimates were based on 27 and 36 analysable exchanges in the first and second round, respectively. Technical issues were observed in both

Table 3 Perceived user experience measures

	First round	Second round	W	q	*r*
GQS					
Anthropomorphism	2.9 (1.1)	2.9 (0.8)	48.0	0.844	0.02
Animacy	3.2 (1.0)	3.3 (0.7)	76.5	0.844	0.09
Likability	3.8 (0.9)	4.0 (0.8)	44.5	0.656	0.24
Intelligence	3.4 (0.8)	3.6 (0.9)	60.5	0.656	0.29
Safety	3.9 (0.6)	3.9 (0.8)	42.0	0.844	0.04
SUS	56.2 (11.4)	66.9 (10.4)	145.0	**0.002**	0.79
Original questionnaire					
Quickness	-	4.2 (1.0)		-	-
Naturalness	-	4.1 (0.9)		-	-
Depth/Richness	-	3.7 (0.8)		-	-
Satisfaction	-	4.1 (1.1)		-	-
Companionship	-	3.3 (1.0)		-	-

rounds, with turn-taking violations most frequent (28 in the first round; 20 in the second). The rate of turn-taking violations decreased from 19.9% to 9.5%. Other issues included repetitive responses (0 vs. 4), hallucinations (4 vs. 3), empty responses (0 vs. 1), and premature terminations due to misinterpreted farewell phrases (0 vs. 5). We also provide episode-level distributions of response latency and DR as violin plots in Supplementary Fig. S2 to illustrate variability across individual interactions.

The Mean of GQS subscale scores showed minimal changes between rounds: Anthropomorphism was 2.9 in both rounds, Animacy increased slightly from 3.2 in the first round to 3.3 in the second, Likability from 3.8 to 4.0, Intelligence from 3.4 to 3.6, and Safety remained at 3.9. No statistically significant differences were detected for these measures. In contrast, the mean SUS score improved significantly from 56.2 (SD = 11.4) in the first round to 66.9 (SD = 10.4) in the second (Table 3). Internal consistency of the adapted SUS was modest in the first round (Cronbach's $\alpha = 0.44$) and low in the second round ($\alpha = 0.24$). Item-level diagnostics (corrected item–total correlations and alpha-if-item-deleted) are provided in Supplementary Table S3.

Exploratory Spearman correlations were conducted to examine whether participant-level changes in log-based metrics were associated with changes in subjective user experience (pre-specified six correlations; BH-adjusted within this set). No statistically reliable associations were observed between changes in response latency or DR and changes in SUS, GQS Intelligence, or GQS Likeability (all $q > 0.05$; Supplementary Table S4).

Supplementary Fig. S3 and S4 provide illustrative examples of the human–robot dialogue alongside the corresponding speech-recognition output and language-model responses, making speech segmentation and response generation transparent.

First-round descriptive totals include all 18 participants, however, the ineligible participant's data was removed to calculate response latency, LLM (Large Language Model)-level latency and Deletion Rate (DR). Conversation episodes: discrete participant–robot conversation sessions initiated and completed within the focus group; Turn-taking: total number of individual speaking turns between participant and robot; Text exchange with LLM: number of text messages sent to and received from the LLM. Response latency: time between the end of the participant's speech item and the start of the robot's verbal response; LLM-level latency: time between sending a transcribed speech item to the LLM and receiving the generated text; DR: proportion of user transcript words missing in the speech-to-text output, calculated as D/N (D = number of missing words, N = total words in the user transcript). Those three values are participant-level medians computed within each round (i.e., for each participant we first took the median across that participant's eligible episodes). The table reports the group median (IQR) across participants of these participant-level medians. Paired comparisons use the Wilcoxon signed-rank test; q-values are Benjamini–Hochberg adjusted across the three paired comparisons reported in Table 2. $Q < 0.05$ are highlighted in bold to indicate statistical significance. Effect size $r = Z/\sqrt{N}$ (Rosenthal's r). Turn-taking violation: instances where the robot interrupted the participant; Repetitive response: episodes sequentially repeating sentences within the same conversation episode; Hallucinations: factually incorrect statements; Empty response: no LLM output with a fallback phrase; Misinterpreting termination phrase: premature conversation end after misinterpreting phrases such as "see you" or "goodbye" as termination cues. All event rates are expressed per turn (denominator = total number of turns in each round) with exact 95% CIs (Clopper–Pearson intervals).

GQS, Godspeed Questionnaire Series; SUS, System Usability Scale. All p-values were jointly corrected using the Benjamini–Hochberg procedure across the six paired comparisons reported in Table 3; values shown are adjusted

q-values, with $q < 0.05$ considered significant. R, effect size (Rosenthal's r). In the SUS, the word "system" was replaced with "robot." GQS scores for each subscale were calculated as the mean of the item scores within that subscale for each participant and then averaged across participants; item-level scores are provided in Supplementary Table S2. SUS scores were calculated in accordance with the standard method, taking into account reverse-scored items. The original questionnaire was a custom five-item scale administered only in the second round, rated on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree), with higher scores indicating a more favourable impression.

7 Discussion

7.1 Summary of Findings and Orientation

This two-round, mixed methods study with community-dwelling older people found that participants experienced smoother, more manageable exchanges after tightening the conversational mechanics, particularly in terms of conversational flow and usability. At the same time, qualitative accounts indicated that limitations in conversational depth and companionship persisted. Median response latency roughly halved (from 9.8 to 4.7 s), DR fell (from 0.37 to 0.07), and SUS increased (from 56.2 to 66.9). By contrast, changes in GQS subscales were minimal, and the Companionship item of the custom questionnaire averaged 3.3/5. Taken together, these findings suggest that while conversational fluency at the technical level may have been achieved, the level of relational readiness remained insufficient.

7.2 What Changed and Why it Matters: Latency and ASR as Prerequisites

Most improvement stemmed from streaming ASR/TTS and revised I/O orchestration, rather than changes in LLM generation time itself. Reduced deletions eased the experience of being misheard and supported topical continuity; turns per minute rose accordingly. This pattern aligns with prior work showing that real-time dialogue hinges on tight turn-timing and input/output design, not only language generation quality [11, 40, 41, 56]. Usability ratings (SUS) increased from the first to the second round; however, given the context-specific wording adaptation and low internal consistency in our sample, SUS should be interpreted cautiously and triangulated with log-based indicators and qualitative accounts. Qualitative accounts consistently highlighted long waits and disrupted turn-taking as key barriers to conversational flow. In addition, although recognition errors were not a dominant topic in participants' comments, the very high baseline deletion rate indicates that ASR reliability was insufficient, and we infer that improving recognition robustness is a prerequisite for trustworthy, sustainable dialogue. Overall, these findings suggest that stabilising latency and ASR is a prerequisite for smoother interaction, but further refinements (e.g., reducing button operations and improving operational simplicity) remain important for future implementations [11, 56]. Exploratory correlations did not show statistically reliable associations between participant-level changes in log-based metrics and changes in questionnaire scores (Supplementary Table S4), suggesting heterogeneity in how technical improvements translate into perceived experience.

7.3 Naturalness and Richness Improved, but Depth and Companionship did Not

Reductions in latency and stabilization of ASR improved perceived naturalness and usability, yet they did not translate into clear gains in anthropomorphism or companionship. Across accounts, perceived naturalness was shaped mainly by interaction coordination (timing, turn-taking, and continuity across turns), whereas perceived richness was judged by the specificity and variety of information and suggestions provided. For example, participants described the second-round interactions as "more natural" and "more in context," but interruptions and fixed phrase reuse occasionally persisted; the opening phrase "I'm here to help" sometimes made exchanges feel reset, and therefore discontinuous and less conversational. Perceived richness increased for some, yet depth remained limited where user-specific knowledge and reciprocity were lacking. This non-symmetric pattern is consistent with evidence that embodiment and copresence increase social presence and acceptance [30, 31], while enhancements in naturalness and usability alone may not yield relational gains. Although agent proactivity can help reduce loneliness [23], our primarily user-initiated design may have limited opportunities for the robot to re-engage participants, which could have contributed to the modest gains observed in perceived companionship. At the same time, our qualitative findings suggest that durable companionship is more likely to require personalization and empathic responses, as suggested by prior work [16], together with continuity behaviours, such as referencing prior interactions and calibrated reciprocity, as emphasised in relational-agent design [65, 66].

Our qualitative accounts help specify what participants meant by these relational ingredients. Emotional reciprocity was framed as more than polite empathic phrases: some

participants described scripted-sounding sympathy as lacking sincerity, whereas companionship was associated with responses that feel contingent and personally meaningful. Shared history was discussed as continuity grounded in knowing the person over time; participants contrasted this with responses feeling "randomised," noting that real friends "already know your history," and that such background shapes what a companion can say. Individual agency was often inferred from an independent personality—such as offering non-obvious perspectives or reasons for disagreement—while consistently agreeable replies were described as an "echo chamber." This expectation of an independent personality was also relevant to perceived conversational depth, as participants valued responses that could add non-obvious perspectives or reasons rather than simply extending a list of agreeable suggestions. Together, these accounts suggest that improved fluency can make interaction easier, but companionship judgements remained anchored in whether the robot could sustain relational continuity and demonstrate reciprocal, agentive engagement. In short, basic conversational fluency appears necessary but insufficient; additional ingredients—shared history, reciprocity and a sense of agency/independent personality—are required for relational outcomes.

Embodiment and non-verbal behaviour may also have shaped interaction quality. Some participants responded positively to Sota's gestures and eye/LED cues as contributing to approachability and engagement, consistent with prior work highlighting the importance of socially legible cues such as eye contact/gaze behaviour for acceptability of companion robots among older people [33]. Our prior home-based RoBoHoN studies likewise suggest that "cute" looks and actions can support engagement alongside conversational interaction [67]. However, several participants also described the robot's toy-like appearance as potentially undermining the seriousness or social weight of the interaction. At the same time, in a short focus-group setting such features may primarily affect initial social presence rather than durable companionship; this further supports evaluating conversational refinements together with systematic manipulation of non-verbal and embodiment features in longer deployments.

7.4 Companion or Catalyst? Two Design Strategies

Drawing on participants' views about whether the robot felt like a companion, we consider two pragmatic design strategies for addressing loneliness in older people. These are not mutually exclusive in principle; however, we present them as distinct options to clarify mechanisms, evaluation, and risks.

(A) Companion-oriented design

Mechanisms and requirements. Prioritise continuity of interaction and a stable consistent conversational style, with age-attuned error repair and expectation management. Personalization and empathic responses are likely prerequisites for perceived relationship quality [16].
Evaluation. Primary outcomes should eventually be validated subjective loneliness measures (e.g., UCLA Loneliness Scale [19]) and related perceptions of social connection; short-term UX and conversational metrics can serve as mediators.
Risks. Personalised, long-term operation raises privacy, misrepresentation, and over-reliance concerns; LLM-specific risks such as hallucinations and opacity require calibrated oversight and escalation pathways [53, 54]. Residual failure modes observed here (repetition, premature termination), indicate the need for guardrails and monitoring [11, 56].

(B) Catalyst-oriented design

Mechanisms and requirements. Emphasise orchestration of human contact—prompts to join activities, lightweight scheduling or introductions, and family check-ins—supported by integration with community services or social prescribing pathways. Evidence from SARs suggests potential to facilitate social interaction in care settings, even without rich conversation [26]. Such a design may also help overcome the difficulty of training and sustaining human providers required for other non-pharmacological interventions [68].
Evaluation. Treat subjective loneliness and social isolation indices as primary endpoints (e.g., change in loneliness and perceived social connection), complemented by objective social-contact indicators such as call/visit frequency or group participation; UX remains secondary.
Risks. Success depends on the individual and contextual social capacity of the user—including physical ability, existing relationships, and available social support. Moreover, prompting greater human contact may, in some cases, undermine self-esteem or reduce willingness to continue using the robot.
A hybrid stance—companion-like micro-interactions coexisting with catalyst functions that link people to people—may be useful in some contexts. The immediate implication of the present results is that improvements in basic conversational fluency will be the basic requirements for both routes.

7.5 Strengths and Limitations

Strengths. First, a within-group, two-round design with the same community-dwelling older people enabled paired comparisons while controlling for between-person variability. Second, we integrated log-level interaction metrics (latency distributions, DRs, turn-taking violations) with standardised questionnaires and framework-guided qualitative analysis, allowing convergent interpretation. Third, we identified a clinically relevant asymmetry—naturalness/usability improved whereas depth/companionship lagged—which directly informs design trade-offs and outcome selection.

Limitations. Several constraints affect study interpretation. (1) Multiple system aspects were changed between rounds (ASR/TTS, LED cues, synthetic voice, casing), and network conditions were not standardised across sessions, limiting causal attribution. (2) Response latencies were calculated for second-turn interactions only; early terminations and first-turn phenomena were not captured. (3) Session lengths differed between rounds, which may have affected opportunity counts; although we emphasised rate-based indicators (e.g., turns per minute), residual bias is possible. (4) Our recognition metric focused on deletions only (not substitutions/insertions); with overlapping speech, this may understate perceived errors. (5) GQS may be insensitive to micro-persona changes without embodiment updates. (6) Given the small sample size and the context-specific wording adaptation with relatively small Cronbach's alpha, SUS results should be interpreted as descriptive indicators of perceived usability rather than as a psychometrically validated measure in this setting. (7) Focus-group opinions are susceptible to expectations and group dynamics. (8) Repeated participation and involvement in suggesting improvements may have introduced familiarity and commitment effects. (9) Administering SUS/GQS after the group discussion may have shaped ratings via shared framing; thus, they should be interpreted as situated evaluations. (10) One ineligible younger participant was present in the first round; although excluded from paired analyses, their presence could have influenced group dynamics in qualitative discussions in the first round. (11) Participants were active and had relatively high levels of digital literacy, and sessions were conducted in a research-facility setting rather than at home; generalisability to less connected or home-use contexts is limited. (12) We did not examine moderators of response such as individual differences (e.g., personality traits including the Big Five, technology familiarity, communication preferences, and sensory status), and cultural differences in norms and expectations may further shape interaction preferences and perceived companionship. Future work could also examine linguistic markers of relational attribution, such as shifts in pronoun use when referring to the robot (e.g., "it" versus more agentive framing), and how these relate to perceived companionship, toy-likeness, and processes of subjectivation over longer deployments. These constraints motivate factorial/AB designs to disentangle component effects, longer deployments in real homes, inclusion of diverse participants, richer ASR metrics, and assessments of potential moderators such as personality traits and cultural norms and expectations in future work.

8 Conclusion

In the same participants, pipeline refinements halved median response latency (from 9.8 to 4.7 s), reduced DR (from 0.37 to 0.07). Perceived usability and conversational naturalness improved, whereas GQS subscales and perceived companionship changed little. These findings indicate that conversational mechanics improved whereas relational qualities did not. Immediate priorities are to remove residual failure modes (interruptions, fixed-phrase resets, premature terminations), simplify operation, and standardise I/O timing; in parallel, progress toward companionship will require a relational layer—consented memory and continuity across encounters, a stable age-attuned persona, and calibrated reciprocity—so that exchanges accumulate rather than reset. The next phase should test these adjustments in longer home deployments and evaluate them with validated loneliness and social-isolation measures as primary outcomes, using UX and log metrics as process indicators under explicit risk controls for LLM-based systems. Taken together, these steps can move the system from fluent demonstration toward measurable social benefit for older people.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12369-026-01407-w.

Acknowledgements We used ChatGPT (GPT-5 and GPT-5 Pro; OpenAI) to assist with drafting, language refinement, and cross-section consistency checks during manuscript preparation. The authors reviewed and edited all suggestions and take full responsibility for the content. We greatly thank all participants for their time and valuable contributions to the study.

Author Contributions YS conceived the study, recruited participants, collected and analysed the data, and drafted the manuscript. MN and CY contributed to data collection and cleaning, with CY also assisting in recruitment. AT, IE and KU developed the robot's dialogue system and improved its user experience. DI and KK contributed to the enhancement of dialogue system design. Qualitative data analysis was discussed primarily among YS, NB, MN, PR and RB. PR supervised qualitative data collection and analysis. NB contributed to the ethical approval process and the refinement of robot UX design. MI secured funding and provided supervision on study design. RH managed and supervised the overall project and guided manuscript preparation. All authors reviewed the manuscript and approved the final version.

Funding Open Access funding provided by The University of Osaka. This work was supported by the SENSHIN Medical Research Foundation (Overseas Study grant), Daiichi-Sankyo "Habataku" Support Program for the Next Generation of Researchers and EPSRC-funded Sensing to Collaboration Programme Grant (EP/V000748/1). Microsoft Azure OpenAI Service and server infrastructure were provided in kind by NTT West.

Data Availability The datasets generated and analysed during the current study are available from the corresponding author upon reasonable request. Owing to privacy and ethics constraints, raw audio and full transcripts cannot be shared publicly.

Declarations

Compliance with Ethical Standards Ethical approval was obtained from the University College London Interaction Centre Local Research Ethics Committee (ID: 0416). All participants provided written informed consent prior to participation.

Competing Interests Authors AT, IE and KU are employees of NTT West, which provided Azure services and server infrastructure but had no role in the data analysis, the decision to submit the manuscript, or the writing of the manuscript. The remaining authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Lee SL, Pearce E, Ajnakina O, Johnson S, Lewis G, Mann F et al (2021) The association between loneliness and depressive symptoms among adults aged 50 years and older: a 12-year population-based cohort study. Lancet Psychiatry 8(1):48–57. https://doi.org/10.1016/S2215-0366(20)30383-7
2. Domènech-Abella J, Mundó J, Haro JM, Rubio-Valera M (2019) Anxiety, depression, loneliness and social network in the elderly: Longitudinal associations from The Irish Longitudinal Study on Ageing (TILDA). J Affect Disord 246:82–88. https://doi.org/10.1016/j.jad.2018.12.043
3. Holt-Lunstad J, Smith TB, Baker M, Harris T, Stephenson D (2015) Loneliness and Social Isolation as Risk Factors for Mortality: A Meta-Analytic Review. Perspect Psychol Sci 10(2):227–237. https://doi.org/10.1177/1745691614568352
4. World Health Organization. WHO Commission on Social Connection [Internet] (2024) [cited 2024 Nov 26]. Available from: https://www.who.int/groups/commission-on-social-connection
5. Holt-Lunstad J (2024) Social connection as a critical factor for mental and physical health: evidence, trends, challenges, and future implications. World Psychiatry 23(3):312–332. https://doi.org/10.1002/wps.21224
6. Hoang P, King JA, Moore S, Moore K, Reich K, Sidhu H et al (2022) Interventions Associated With Reduced Loneliness and Social Isolation in Older Adults: A Systematic Review and Meta-analysis. JAMA Netw Open 5(10):e2236676. https://doi.org/10.1001/jamanetworkopen.2022.36676
7. Gilbody S, Littlewood E, McMillan D, Atha L, Bailey D, Baird K et al (2024) Behavioural activation to mitigate the psychological impacts of COVID-19 restrictions on older people in England and Wales (BASIL+): a pragmatic randomised controlled trial. Lancet Healthy Longev 5(2):e97–107. https://doi.org/10.1016/S2666-7568(23)00238-6
8. World Health Organization. Digital interventions for reducing social isolation and loneliness in older adults. [cited 2024 Nov 28]. Available from: https://www.who.int/initiatives/decade-of-healthy-ageing/evidence-gap-map/sil-digital
9. Satake Y, Costello H, Naran N, Ishimaru D, Ikeda M, Howard R (2026) Autonomous conversational agents for loneliness, social isolation, depression and anxiety in older people without cognitive impairment: Systematic review and meta-analysis. Psychol Med 56:e27. https://doi.org/10.1017/S0033291725103073
10. Nagata Y, Satake Y, Yamazaki R, Nishio S, Suzuki M, Kanemoto H et al (2025) A Conversational Robot for Cognitively Impaired Older People Who Live Alone: An Exploratory Feasibility Study. Psychogeriatrics 25(5):e70076. https://doi.org/10.1111/psyg.70076
11. Irfan B, Kuoppamäki S, Hosseini A, Skantze G (2025) Between reality and delusion: challenges of applying large language models to companion robots for open-domain dialogues with older adults. Auton Robots 49(1):9. https://doi.org/10.1007/s10514-025-10190-y
12. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P et al (2020) Language Models are Few-Shot Learners. arXiv. https://doi.org/10.48550/arXiv.2005.14165
13. Satake Y, Ishimaru D, Katsuki K, Rapaport P, Berthouze N, Howard R et al (2025) A Week With a Conversational Large Language Model Companion Robot. Am J Geriatr Psychiatry 33(7):799–800. https://doi.org/10.1016/j.jagp.2025.03.010
14. Welch V, Ghogomu ET, Barbeau VI, Dowling S, Doyle R, Beveridge E et al (2023) Digital interventions to reduce social isolation and loneliness in older adults: An evidence and gap map. Campbell Syst Rev 19(4):e1369. https://doi.org/10.1002/cl2.1369
15. Sha S, Loveys K, Qualter P, Shi H, Krpan D, Galizzi M (2024) Efficacy of relational agents for loneliness across age groups: a systematic review and meta-analysis. BMC Public Health 24(1):1802. https://doi.org/10.1186/s12889-024-19153-x
16. He Y, Yang L, Qian C, Li T, Su Z, Zhang Q et al (2023) Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. J Med Internet Res 25:e43862. https://doi.org/10.2196/43862
17. Zhang Q, Wong AKC, Bayuo J (2024) The Role of Chatbots in Enhancing Health Care for Older Adults: A Scoping Review. J Am Med Dir Assoc 25(9):105108. https://doi.org/10.1016/j.jamda.2024.105108
18. Chou YH, Lin C, Lee SH, Lee YF, Cheng LC (2024) User-Friendly Chatbot to Mitigate the Psychological Stress of Older Adults During the COVID-19 Pandemic: Development and Usability Study. JMIR Form Res 8:e49462. https://doi.org/10.2196/49462
19. Russell DW (1996) UCLA Loneliness Scale (Version 3): Reliability, validity, and factor structure. J Pers Assess 66(1):20–40. https://doi.org/10.1207/s15327752jpa6601_2
20. Jones VK, Yan C, Shade MY, Boron JB, Yan Z, Heselton HJ et al (2024) Reducing Loneliness and Improving Social Support among Older Adults through Different Modalities of Personal

Voice Assistants. Geriatr (Basel) 9(2):22. https://doi.org/10.3390/geriatrics9020022
21. Yan C, Johnson K, Jones VK (2024) The Impact of Interaction Time and Verbal Engagement with Personal Voice Assistants on Alleviating Loneliness among Older Adults: An Exploratory Study. Int J Environ Res Public Health 21(1):100. https://doi.org/10.3390/ijerph21010100
22. Loveys K, Sagar M, Pickering I, Broadbent E (2021) A Digital human for delivering a remote loneliness and stress intervention to at-risk younger and older adults during the COVID-19 pandemic: Randomized pilot trial. JMIR Ment Health 8(11):e31586. https://doi.org/10.2196/31586
23. Ring L, Shi L, Totzke K, Bickmore T (2015) Social support agents for older adults: longitudinal affective computing in the home. J Multimodal User Interfaces 9(1):79–88. https://doi.org/10.1007/s12193-014-0157-0
24. Yen HY, Huang CW, Chiu HL, Jin G (2024) The Effect of Social Robots on Depression and Loneliness for Older Residents in Long-Term Care Facilities: A Meta-Analysis of Randomized Controlled Trials. J Am Med Dir Assoc 25(6):104979. https://doi.org/10.1016/j.jamda.2024.02.017
25. Wada K, Shibata T, Saito T, Sakamoto K, Tanie K (2005) Psychological and social effects of one year robot assisted activity on elderly people at a health service facility for the aged. In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA), pp. 2785–90
26. Yu C, Sommerlad A, Sakure L, Livingston G (2022) Socially assistive robots for people with dementia: Systematic review and meta-analysis of feasibility, acceptability and the effect on cognition, neuropsychiatric symptoms and quality of life. Ageing Res Rev 78:101633. https://doi.org/10.1016/j.arr.2022.101633
27. Lee OE, Nah KO, Kim EM, Choi NG, Park DH (2024) Exploring the Use of Socially Assistive Robots Among Socially Isolated Korean American Older Adults. J Appl Gerontol 43(9):1295–1304. https://doi.org/10.1177/07334648241236081
28. Lee OE, Lee H, Park A, Choi NG (2024) My Precious Friend: Human-Robot Interactions in Home Care for Socially Isolated Older Adults. Clin Gerontol 47(1):161–170. https://doi.org/10.1080/07317115.2022.2156829
29. Lee OEK, Nam I, Chon Y, Park A, Choi N (2023) Socially Assistive Humanoid Robots: Effects on Depression and Health-Related Quality of Life among Low-Income, Socially Isolated Older Adults in South Korea. J Appl Gerontol 42(3):367–375. https://doi.org/10.1177/07334648221138283
30. Li J (2015) The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. Int J Hum-Comput Stud 77:23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001
31. Heerink M, Kröse B, Evers V, Wielinga B (2009) Influence of Social Presence on Acceptance of an Assistive Social Robot and Screen Agent by Elderly Users. Adv Robot 23(14):1909–1923. https://doi.org/10.1163/016918609X12518783330289
32. Nishio T, Yoshikawa Y, Sakai K, Iio T, Chiba M, Asami T et al (2021) The Effects of Physically Embodied Multiple Conversation Robots on the Elderly. Front Robot AI 8:633045. https://doi.org/10.3389/frobt.2021.633045
33. Bradwell HL, Edwards KJ, Winnington R, Thill S, Jones RB (2019) Companion robots for older people: importance of user-centred design demonstrated through observations and focus groups comparing preferences of older people and roboticists in South West England. BMJ Open 9(9):e032468. https://doi.org/10.1136/bmjopen-2019-032468
34. Pellegrini T, Trancoso I, Hämäläinen A, Calado A, Dias MS, Braga D et al (2012) Impact of age in ASR for the elderly: preliminary experiments in European Portuguese. In: Torre Toledano D, Ortega Giménez A, Teixeira A, González Rodríguez J, Hernández Gómez L, San Segundo Hernández R, editors. Advances in Speech and Language Technologies for Iberian Languages [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 139–47. (Communications in Computer and Information Science). https://doi.org/10.1007/978-3-642-35292-8_15
35. Vipperla R, Renals S, Frankel J (2008) Longitudinal study of ASR performance on ageing voices. In: Interspeech 2008. ISCA, pp 2550–2553. https://doi.org/10.21437/Interspeech.2008-632
36. Iio T, Yoshikawa Y, Chiba M, Asami T, Isoda Y, Ishiguro H (2020) Twin-Robot Dialogue System with Robustness against Speech Recognition Failure in Human-Robot Dialogue with Elderly People. Appl Sci 10(4):1522. https://doi.org/10.3390/app10041522
37. Strachan JWA, Albergo D, Borghini G, Pansardi O, Scaliti E, Gupta S et al (2024) Testing theory of mind in large language models and humans. Nat Hum Behav 8(7):1285–1295. https://doi.org/10.1038/s41562-024-01882-z
38. Welivita A, Pu P (2024) Is ChatGPT more empathetic than humans? arXiv. https://doi.org/10.48550/arXiv.2403.05572
39. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R et al (2025) Towards conversational diagnostic artificial intelligence. Nature 642(8067):442–450. https://doi.org/10.1038/s41586-025-08866-7
40. Meyer AS (2023) Timing in Conversation. J Cogn 6(1):20. https://doi.org/10.5334/joc.268
41. Levinson SC, Torreira F (2015) Timing in turn-taking and its implications for processing models of language. Front Psychol 6:731. https://doi.org/10.3389/fpsyg.2015.00731
42. OpenAI, Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A et al (2024) GPT-4o system card. arXiv. https://doi.org/10.48550/arXiv.2410.21276
43. Hinton G, Deng L, Yu D, Dahl G, Mohamed AR, Jaitly N et al (2012) Deep neural networks for acoustic modeling in speech recognition
44. Amodei D, Anubhai R, Battenberg E, Case C, Casper J, Catanzaro B et al (2015) Deep speech 2: end-to-end speech recognition in English and Mandarin. arXiv. https://doi.org/10.48550/arXiv.1512.02595
45. Graves A (2012) Sequence transduction with recurrent neural networks. arXiv. https://doi.org/10.48550/arXiv.1211.3711
46. Li J, Zhao R, Hu H, Gong Y, Improving (2019) RNN Transducer Modeling for End-to-End Speech Recognition. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) [Internet]. SG, Singapore: IEEE;, pp 114–121. https://doi.org/10.1109/ASRU46091.2019.9003906
47. Baevski A, Zhou H, Mohamed A, Auli M (2020) wav2vec 2.0: a framework for self-supervised learning of speech representations. arXiv. https://doi.org/10.48550/arXiv.2006.11477
48. Hsu WN, Bolte B, Tsai YHH, Lakhotia K, Salakhutdinov R, Mohamed A (2021) HuBERT: self-supervised speech representation learning by masked prediction of hidden units. arXiv. https://doi.org/10.48550/arXiv.2106.07447
49. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I (2022) Robust speech recognition via large-scale weak supervision. arXiv. https://doi.org/10.48550/arXiv.2212.04356
50. Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z et al (2018) Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. arXiv. https://doi.org/10.48550/arXiv.1712.05884
51. Ren Y, Ruan Y, Tan X, Qin T, Zhao S, Zhao Z et al (2019) FastSpeech: fast, robust and controllable text to speech. arXiv. https://doi.org/10.48550/arXiv.1905.09263
52. Kong J, Kim J, Bae J (2020) HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. arXiv. https://doi.org/10.48550/arXiv.2010.05646
53. Yamamoto R, Song E, Kim JM (2020) Parallel WaveGAN: a fast waveform generation model based on generative adversarial

networks with multi-resolution spectrogram. arXiv. https://doi.org/10.48550/arXiv.1910.11480
54. Schreiter T, Rüppel JV, Hazra R, Rudenko A, Magnusson M, Lilienthal AJ (2025) Evaluating efficiency and engagement in scripted and LLM-enhanced human-robot interactions. arXiv. https://doi.org/10.48550/arXiv.2501.12128
55. Rüppel JV, Rudenko A, Schreiter T, Magnusson M, Lilienthal AJ (2025) Gaze-supported large language model framework for Bi-directional human-robot interaction. arXiv. https://doi.org/10.48550/arXiv.2507.15729
56. Kim CY, Lee CP, Mutlu B (2024) Understanding large-language model (LLM)-powered human-robot interaction. In: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, pp. 371–80. https://doi.org/10.1145/3610977.3634966
57. Haltaufderheide J, Ranisch R (2024) The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). Npj Digit Med 7(1):183. https://doi.org/10.1038/s41746-024-01157-x
58. Ranisch R, Haltaufderheide J (2025) Rapid Integration of LLMs in Healthcare Raises Ethical Concerns: An Investigation into Deceptive Patterns in Social Robots. Digit Soc 4(1):7. https://doi.org/10.1007/s44206-025-00161-2
59. Gale NK, Heath G, Cameron E, Rashid S, Redwood S (2013) Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 13(1):117. https://doi.org/10.1186/1471-2288-13-117
60. Krueger RA, Casey MA (2015) Focus group interviewing. In: Handbook of Practical Program Evaluation. John Wiley & Sons, Ltd, pp 506–534. https://doi.org/10.1002/9781119171386.ch20
61. Braun V, Clarke V (2021) Thematic analysis: a practical guide. SAGE, London; Thousand Oaks (CA), p 376
62. Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. Int J Soc Robot 1(1):71–81. https://doi.org/10.1007/s12369-008-0001-3
63. Seaborn K, Sekiguchi T, Tokunaga S, Miyake NP, Otake-Matsuura M (2023) Voice Over Body? Older Adults' Reactions to Robot and Voice Assistant Facilitators of Group Conversation. Int J Soc Robot 15(2):143–163. https://doi.org/10.1007/s12369-022-00925-7
64. Brooke J (1996) SUS - a quick and dirty usability scale. In: Usability Evaluation in Industry. Taylor & Francis, pp 189–194
65. Bickmore TW, Caruso L, Clough-Gorr K, Heeren T (2005) It's just like you talk to a friend' relational agents for older adults. Interact Comput 17(6):711–735. https://doi.org/10.1016/j.intcom.2005.09.002
66. Bickmore TW, Puskar K, Schlenk EA, Pfeifer LM, Sereika SM (2010) Maintaining reality: Relational agents for antipsychotic medication adherence. Interact Comput 22(4):276–288. https://doi.org/10.1016/j.intcom.2010.02.001
67. Nagata Y, Satake Y, Yamazaki R, Nishio S, Ishiguro H, Ikeda M (2026) Case report: A five-case series of 18-month home use of a conversational companion robot for psychological support in older people with mild cognitive impairment or late-onset psychosis. Front Psychiatry 16:1700340. https://doi.org/10.3389/fpsyt.2025.1700340
68. Wang HI, Gilbody S, Littlewood E, Baird K, Ekers D, McMillan D et al (2025) Cost-utility of behavioural activation for mitigating psychological impacts of COVID-19 on socially isolated older adults with depression and multiple long-term conditions compared with usual care: results from a pragmatic randomised controlled trial. BMJ Ment Health 28(1):e301270. https://doi.org/10.1136/bmjment-2024-301270

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dr Yuto Satake is a geriatric psychiatrist affiliated with the Division of Psychiatry, University College London, and the Department of Psychiatry, The University of Osaka Graduate School of Medicine. His research focuses on geriatric psychiatry, late-life neuropsychiatric disorders, and the development and evaluation of conversational robots and LLM-enabled agents to support older people.

Dr Minna Nygren Minna Nygren is a Research Fellow at the UCL Interaction Centre, University College London. Her interests include human–robot interaction , embodied interaction, and the design and evaluation of interactive technologies in real-world contexts.

Clare Yu is a PhD candidate at the Division of Psychiatry, University College London. Her research focuses on dementia and mental health, with a particular interest in the design and evaluation of digital and robotic interventions for older adults.

Akira Tsuboi is an engineer at NTT WEST, Inc. His work focuses on the development and implementation of conversational and interactive systems, including the technical design and deployment of LLM-enabled robot platforms.

Ippei Endo is an engineer at NTT WEST, Inc. deployed conversational technologies, including dialogue-system implementation, performance monitoring, and operational support.

Kazuhiro Umemura is a data analyst at NTT WEST, Inc. His work focuses on the analysis of large-scale datasets, including sleep-related data, and the development of data-driven approaches for real-world healthcare and technology applications.

Kunihiko Katsuki is an occupational therapist affiliated with the Department of Psychiatry, The University of Osaka Graduate School of Medicine. His work relates to clinical research in psychiatry and the application of digital technologies to support older people in clinical and community contexts.

Dr Daiki Ishimaru is an academic occupational therapist affiliated with the Department of Psychiatry, The University of Osaka Graduate School of Medicine, and the Department of Medical Technology, The University of Osaka Hospital. His interests include dementia research, activities of daily living, and the application of quantitative approaches to psychiatric and behavioural research.

Penny Rapaport is a Professor of Psychological Interventions in Dementia at the Division of Psychiatry, University College London. Her research focuses on developing and evaluating interventions to support people living with dementia and memory problems, including work with older people and marginalised populations.

Prof Nadia Bianchi-Berthouze is a Professor at the UCL Interaction Centre, University College London. She is an HCI researcher whose work explores affective and embodied interaction, with interests in designing interactive systems that support engagement, communication, and wellbeing.

Prof Manabu Ikeda is a Professor of Psychiatry at the Department of Psychiatry, The University of Osaka Graduate School of Medicine. His research focuses on geriatric psychiatry, dementia, and neuropsychiatric symptoms, including clinical and translational approaches to Alzheimer's disease, frontotemporal dementia, and dementia with Lewy bodies.

Prof Robert Howard is a Professor of Old Age Psychiatry at the Division of Psychiatry, University College London. His research focuses on late-life psychiatric disorders, dementia, psychopharmacology, and clinical trials of interventions for older people with mental health conditions and neurodegenerative disorders.