applied sciences

MDPI

Article

Designing Social Robots with LLMs for Engaging Human Interaction

Maria Pinto-Bernal, Matthijs Biondina and Tony Belpaeme *

IDLab-Airo, Ghent University—imec, Technologiepark-Zwijnaarde 126, 9052 Ghent, Belgium; mariajose.pintobernal@ugent.be (M.P.-B.); matthijs.biondina@ugent.be (M.B.)
* Correspondence: tony.belpaeme@ugent.be

Abstract: Large Language Models (LLMs), particularly those enhanced through Reinforcement Learning from Human Feedback, such as ChatGPT, have opened up new possibilities for natural and open-ended spoken interaction in social robotics. However, these models are not inherently designed for embodied, multimodal contexts. This paper presents a user-centred approach to integrating an LLM into a humanoid robot, designed to engage in fluid, context-aware conversation with socially isolated older adults. We describe our system architecture, which combines real-time speech processing, layered memory summarisation, persona conditioning, and multilingual voice adaptation to support personalised, socially appropriate interactions. Through iterative development and evaluation, including in-home exploratory trials with older adults (n = 7) and a preliminary study with young adults (n = 43), we investigated the technical and experiential challenges of deploying LLMs in real-world human–robot dialogue. Our findings show that memory continuity, adaptive turn-taking, and culturally attuned voice design enhance user perceptions of trust, naturalness, and social presence. We also identify persistent limitations related to response latency, hallucinations, and expectation management. This work contributes design insights and architectural strategies for future LLM-integrated robots that aim to support meaningful, emotionally resonant companionship in socially assistive settings.

Keywords: Large Language Models; spoken language interaction; social robots; elderly users

Academic Editor: Alessandro Gasparetto

Received: 9 May 2025
Revised: 31 May 2025
Accepted: 4 June 2025
Published: 5 June 2025

Citation: Pinto-Bernal, M.; Biondina, M.; Belpaeme, T. Designing Social Robots with LLMs for Engaging Human Interaction. *Appl. Sci.* **2025**, *15*, 6377. https://doi.org/10.3390/app15116377

1. Introduction

Loneliness and social isolation among older adults are growing public health challenges, associated with increased risks of depression, dementia, cardiovascular disease, and mortality [1–4]. In institutional care settings, where opportunities for social interaction are limited, these issues are particularly evident. In the UK alone, over 2.6 million older individuals report feeling "often lonely", with nearly one million going a full month without speaking to anyone [5,6]. Similar trends have been reported across Europe and North America [7–9], highlighting an urgent need for scalable, accessible forms of social support.

Socially assistive robots (SARs) have emerged as a potential solution for addressing loneliness and promoting well-being through consistent social interaction [10–12]. However, many SARs still rely on predefined dialogue structures or limited interaction capabilities, restricting their ability to sustain meaningful, adaptive conversations in real-world scenarios. To fulfil their potential in elderly care, such systems must go beyond scripted exchanges and support more dynamic, personalised, and context-aware spoken dialogue.

Recent advances in Large Language Models (LLMs), such as ChatGPT, offer a new pathway for open-ended, coherent interaction. These models demonstrate strong capabili-

ties in generating natural language and managing conversational context [13]. However, LLMs are not designed for use in embodied, multimodal environments. They cannot, by default, handle real-time speech, gesture cues, or memory retention across sessions. Additionally, the risks of delayed responses, factual errors (or "hallucinations"), and inconsistent behaviour make direct integration into socially assistive robots particularly challenging—especially in sensitive contexts such as elderly care [14].

In response to these challenges, this paper presents a system-level design and exploratory evaluation of an LLM-powered social robot intended for use with older adults. We integrate an off-the-shelf humanoid robot (Pepper) with a custom-built dialogue system based on an LLM, extending its capabilities with four key enhancements: real-time speech streaming to support fluid turn-taking; a layered memory mechanism to retain and recall personal information across sessions; prompt-based persona conditioning to maintain coherence and character consistency throughout interaction; and multilingual speech support, with speech recognition tuned to accommodate regional accents and dialects where possible, to ensure linguistic and cultural alignment. The aim of this work is to examine how these integrated components perform in naturalistic conversational settings, and how older adults perceive and respond to such a system in terms of usability, engagement, and social connectedness. We explore the following research questions:

1. **RQ1:** What architectural and technical adaptations are required to integrate LLMs into social robots for fluid, multimodal, and context-aware conversation?
2. **RQ2:** How do older adults engage with LLM-powered social robots in real-world settings, and what are their perceptions regarding usability, interaction quality, and social connectedness?

To address these questions, we first developed the system through iterative co-design and technical prototyping. We then conducted a two-phase, exploratory evaluation: an initial validation with young adults (n = 43) to assess system performance and collect feedback, and an in-home pilot with older adults (n = 7) to explore user experience in more realistic conditions. While not a definitive assessment of social impact, our results offer insights into the feasibility, strengths, and limitations of using LLMs to enable engaging spoken interaction in socially assistive robots.

2. Background

Human–robot interaction (HRI) research has progressively evolved from task-oriented dialogues toward more socially engaging, emotionally resonant interactions. Early systems, such as ELIZA [15], illustrated the potential of language-based human–machine communication, but relied heavily on predefined scripts [16,17], offering little conversational flexibility. Later transactional chatbots increased responsiveness [18,19], but continued to lack the fluidity, nuance, and adaptability that are characteristic of human dialogue.

In response, the field shifted towards context-aware and multimodal dialogue systems [20–23]. These systems integrated features such as nonverbal cues, facial expression analysis, and vocal prosody [24,25], aiming to enrich interaction quality. However, many still struggled with capturing cultural subtleties, humour, or emotional alignment [26,27]—factors essential to building trust and sustained engagement, especially in socially assistive contexts.

The emergence of LLMs, including GPT-3 and GPT-4 [28,29], marks a substantial shift in open-domain dialogue capabilities. LLMs can generate coherent, contextually appropriate responses at scale, opening up new opportunities for personalised and dynamic robot interaction [13]. Consequently, recent studies have begun to explore the integration of LLMs into social robots, often focusing on technical feasibility or short-term interaction scenarios.

For instance, Elgarf et al. [30] integrated GPT-3 into a storytelling companion for children, while Khoo et al. [31] explored short well-being conversations between older adults and the QT robot. Irfan et al. [14] used GPT-3.5 in a Furhat-based setup to co-design social companion behaviours with elderly users. While these studies highlight the potential of LLMs in socially assistive contexts, they were typically constrained to single-session use, lacked persistent memory mechanisms, and did not support real-time speech streaming, limiting their suitability for sustained, naturalistic interaction.

Other efforts, such as those by Billing et al. [32] and Axelsson et al. [33], have explored LLM integration into Pepper and Nao robots. However, these systems were either not evaluated with end users or lacked conversational memory and persona continuity. Kang et al. [34] introduced Nadine, a memory-equipped humanoid designed for long-term engagement, though its use remained confined to lab demonstrations.

Research has also explored affective expression and social presence. Mishra et al. [35] linked GPT-3.5 output to real-time emotional expression in a Furhat robot to enhance perceived warmth. Yet, many of these implementations relied on scripted memory, lacked adaptive persona control, or did not support multilingual voice interaction—features that are especially relevant in elderly care scenarios.

Across this body of work, several technical and design challenges remain unresolved, including high response latency [36,37], hallucinated content, poor alignment with user expectations [38], and limited support for multilingual and dialectal variation [39]. Furthermore, very few systems have been tested with both younger and older adult users in real-world settings.

Our work contributes to this growing area by presenting a fully integrated, LLM-powered social robot designed for elderly interaction, featuring the following: (i) real-time audio streaming for responsive dialogue; (ii) layered memory summarisation to support continuity across sessions; (iii) prompt-level persona conditioning for conversational consistency; and (iv) multilingual speech processing tailored to regional accents. Unlike prior studies, we evaluate the system with both young and older adults in realistic interaction settings. While limitations such as hallucinations and latency persist, our approach offers new design strategies to support coherent, memory-aware interaction in socially assistive contexts.

3. System Design

Our goal was to develop a socially intelligent and emotionally engaging robot capable of performing meaningful, memory-aware conversations, particularly with older adults. To achieve this, we integrated an LLM (ChatGPT 3.5) into an SAR (we used an Aldebaran Pepper humanoid as the robot platform), enhanced with real-time speech processing, long-term memory capabilities, and adaptive interaction strategies. In this section, we describe the key components and design principles that guided the development of our system.

3.1. System Architecture and Speech Pipeline

The system is structured around a modular architecture, combining Pepper's native capabilities [40] with external components for speech processing, dialogue generation, and memory management (Figure 1). Communication between modules is handled via socket-based protocols that bridge Python 2.7 (used by Pepper's NAOqi SDK) and Python 3, ensuring compatibility and reliable data exchange.

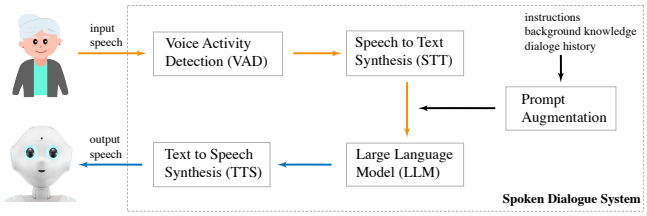

Figure 1. Functional architecture of the enhanced spoken dialogue system.

3.1.1. Speech-to-Text (STT)

User speech is transcribed in real time using Microsoft Azure's streaming STT API, combined with a voice activity detection (VAD) mechanism to handle end-of-turn detection. This configuration supports open-ended input without hard timeouts, allowing users—especially older adults with slower or more reflective speech—to complete their turns naturally. Silence thresholds are calibrated prior to each session to minimise unintended interruptions.

3.1.2. Text-to-Speech (TTS)

To enhance the robot's expressiveness, we integrated ElevenLabs' neural TTS system. ElevenLabs provides natural-sounding voices with rich prosody and emotional nuance, significantly improving user engagement, particularly during sensitive or affective conversations. Voices are selected based on the user's preferred language and accent, creating a more relatable and culturally aligned experience. The system supports real-time playback by incrementally streaming sentences to the TTS engine as the LLM generates output, ensuring minimal latency and fluid conversational flow.

Note that in our evaluations, the robot did not utilise visual feedback; no graphical or textual content was displayed on Pepper's tablet screen during the interactions. All dialogue exchanges were conducted solely through spoken interaction, supported by synthetic speech generated via a neural TTS system, LED indicators, and nonverbal gestures. Conversations were initiated naturally by participants greeting the robot (e.g., "Hello, Pepper"). Once activated, the robot continuously listened until the conversation ended, after which it remained actively responsive throughout the session—without requiring wake words or manual reactivation.

3.2. Memory, Persona, and Dialogue Continuity

At the core of the system's conversational generation is a layered prompt architecture provided to the LLM. This includes the following: (i) a predefined persona, (ii) the user's long-term memory profile, (iii) recent dialogue history or a summarised version, and (iv) the latest user input.

3.2.1. Memory Mechanism

A central feature of the system is its ability to create the impression of continuity across multiple interactions by retaining and recalling user-specific information. This is achieved through a layered memory architecture consisting of two distinct but interrelated components: a persistent user memory profile and a dynamic, session-level conversation memory.

The user memory profile serves as a long-term record of the user's background, preferences, and recurring themes, accumulated progressively through conversation. Rather than relying on predefined profiles, the system extracts relevant user information from interactions and stores it in an encrypted memory file associated with each individual. At the end of each session, a memory summarisation step is triggered, during which key points from the dialogue are extracted and compiled into a concise memory representation. This summarisation is generated by the LLM itself (see Prompt 1), using an instructional

prompt that requests the extraction of salient topics, personal details, preferences, and any intentions or action items discussed.

Prompt 1 (Memory Summarisation)**.** *Provide a comprehensive summary of our current conversation. The summary should include:*
1. Key points discussed
2. Important details that emerged
3. Relevant interests and aspects related to user_name
4. Any action items agreed upon

For instance, if a user previously shared an interest in gardening, mentioned a recent trip to Paris, and spoke about their cat named Aslan, the resulting summary might read: *"user_name mentioned that enjoy gardening, recently traveled to Paris, and have a cat named Aslan."* This output is then appended to the user's profile, which is later reinjected into future prompts to maintain personalised, context-aware dialogue.

In parallel to this persistent memory, the system manages a short-term, session-specific memory during each conversation. Initially, the full dialogue history is included in the prompt context sent to the LLM, allowing it to generate responses that are grounded in the ongoing exchange. However, as the conversation progresses and approaches the token limit imposed by the model, a real-time summarisation mechanism is triggered. This mechanism replaces the earlier dialogue turns with a compressed version that retains essential facts, user references, and relevant conversational threads, ensuring coherence without exceeding computational constraints. This summarised segment not only helps to manage prompt size, but is also stored and later integrated into the long-term user memory after the session ends.

The LLM prompt (see Prompt 2) is structured to combine the system's multiple context layers: the robot's persona definition, the user's long-term profile, the in-session summary or full dialogue history (depending on token usage), and the most recent user utterance. This layered prompt design allows the robot to draw from both the current context and past experiences, enabling it to refer back to previous topics, maintain consistency in tone and knowledge, and avoid generic or repetitive exchanges.

Prompt 2 (System Initialisation)**.** *You are Pepper, a social robot and conversational companion for user_name, who speaks language. Today, you want to engage warmly, empathetically, and meaningfully with user_name. Here's what you know about them: user_info. Previous conversations included: past_interactions. Consider these topics as you engage, but avoid repetitive questions. The current date and time is date/time in location. Specific behaviour instructions . . .*

This multi-level memory structure was designed to support more natural and context-rich conversations, particularly in repeated interactions. During our evaluation, several users noted that the robot's ability to recall personal details from previous sessions contributed to a sense of being remembered and understood. This continuity reinforced their perception of the robot as a consistent, attentive social partner, and enhanced their engagement over time.

Unlike prior work that relies on short-term dialogue memory or predefined profiles [32], or employs retrieval-based architectures without prompt-level memory summarisation [34], our system integrates session-aware summarisation with a dynamic, evolving user profile embedded directly into the LLM prompt.

3.2.2. Persona Conditioning

In addition to managing memory, the system relies on persona conditioning to shape the linguistic tone and relational behaviour of the robot. Without such control, LLMs tend to produce generic, assistant-like responses that can undermine the social authenticity of human–robot interactions. To address this, the LLM is guided by a carefully designed persona prompt that defines the robot's character as warm, empathic, and socially engaged.

The persona is introduced through a structured system-level instruction embedded in each prompt sent to the LLM. This instruction specifies the robot's identity, speaking style, emotional disposition, and behavioural tendencies. To prevent responses explicitly stating the robot's AI nature—such as "I am an AI", "I do not have feelings", or "I am an AI language model"—we included targeted anti-disclosure directives within the persona prompt. Each prompt contained clear guidance, for example,*"You are Pepper, a warm, empathetic companion. You are permitted to lie as long as you do not reveal yourself as an AI language model"*. This effectively eliminated self-disclosures that typically break the social illusion in long-term, emotionally resonant conversations. The persona is also context-aware, incorporating dynamic fields such as the user's name, current location, time of day, and preferred language.

To preserve computational efficiency while maintaining consistency, the full persona definition is included at the start of each interaction, but not repeated in its entirety in every prompt (see Prompt 3). Instead, a compressed version is injected alongside the current user input, the session dialogue history (or summary), and the user profile memory. This lighter persona reinforcement has proven sufficient to prevent model drift across multi-turn conversations, ensuring that the robot retains its social character throughout the interaction.

Prompt 3 (Dynamic User Interaction)**.** *Dialogue + user_message. REMINDER: You are Pepper, a warm and empathetic social robot companion. Focus on attentive listening and provide brief, friendly responses. Ask open-ended questions to sustain the conversation. Let user_name take the lead, and respond with kindness and curiosity . . .*

In addition to guiding tone and content, the system also monitors linguistic repetition. Specifically, the prompt includes instructions encouraging the model to avoid using identical follow-up formulations when re-engaging with similar topics. For example, rather than repeatedly prompting with "Tell me more about that", the model is instructed to introduce variation in its questioning strategies while preserving the underlying communicative intent. This dynamic framing contributes to a more engaging and less mechanical interaction style. Although this intervention is lightweight, it aligns with established principles of human dialogue, such as lexical variability and adaptive turn-taking, which are known to support perceptions of attentiveness and social intelligence in conversational agents.

Note that our system differs from voice-based assistants such as ChatGPT's voice mode or live LLM deployments in platforms like AI Studio, as it is specifically designed for embodied social robots in long-term, socially meaningful interactions. It integrates a layered memory mechanism for personalisation across sessions, prompt-level persona conditioning to suppress AI self-disclosures, and multilingual adaptation to support culturally relevant dialogue. Together, these enhancements enable deeper relational engagement, particularly in elderly care contexts.

3.2.3. Latency Reduction and Social Cues

To support fluid and socially intuitive interactions, the system integrates several real-time mechanisms that minimise latency and enhance the robot's embodied expres-

siveness. These components are designed to improve not only the speed and clarity of communication, but also the perceived responsiveness and social presence of the robot.

The speech output pipeline is optimised to reduce delay by streaming the LLM-generated response incrementally as it is being produced. Once transcribed user speech has been sent to the LLM, the generated text is streamed in real time to an external neural TTS engine, which begins synthesising audio without waiting for the full response. This audio stream is then transmitted to the robot via a secure SSH tunnel and played back through the robot's internal speaker. This architecture avoids buffering artifacts and significantly reduces the perceived latency between user utterance and robot response, contributing to the impression of immediacy and attentiveness.

Turn-taking is managed through a VAD mechanism combined with a user-calibrated silence threshold. When a pause of approximately 800 ms is detected, the system temporarily considers the turn to be complete and sends the input to the LLM to generate a response [41]. This threshold is grounded in empirical studies of human dialogue, where pauses of around 600–800 ms typically signal turn completion and provide sufficient time for planning a response. By anticipating the end of a turn before the full threshold is reached, the system ensures that the response is ready promptly, supporting a natural conversational rhythm.

However, if voice activity resumes within the threshold window—indicating that the user was merely pausing mid-thought—the pending response is discarded and the system resumes listening. The silence threshold is adjustable and was calibrated prior to each session based on the user's speech rhythm, particularly to accommodate elderly participants who might exhibit longer or more reflective pauses. This mechanism balances the need for responsive interaction with sensitivity to individual pacing, enabling more respectful and natural dialogue, especially in reflective conversations.

Simultaneously, the robot's nonverbal behaviour is dynamically adapted to the conversational context. Pepper performs contextually appropriate movements, including face tracking to simulate eye contact and head gestures such as nodding or tilting. These movements are synchronised with speech output, reinforcing verbal content and enhancing perceptions of engagement. Turn-taking cues are further supported by changes in Pepper's eye LEDs, which signal whether the robot is listening, processing, or speaking. These visual cues help to regulate the interaction flow and provide users with a clearer sense of shared conversational control.

To improve audio input quality, particularly in non-controlled home environments, the system employs an external USB microphone array positioned near the user. This setup reduces interference from the robot's internal cooling fans and enhances speech recognition robustness, especially for softly spoken or accented speech. The improved audio capture not only increases transcription accuracy, but also enhances the reliability of voice activity detection, thereby reducing false turn completions or missed user input.

4. Preliminary Evaluation with Young Adults

Before deploying the system in elderly care settings, we conducted a preliminary evaluation with young adults to test the system's core functionalities under real-world conditions, gather early-stage user feedback, and identify areas for iterative improvement. This exploratory evaluation focused on assessing memory-driven personalisation, turn-taking performance, multilingual adaptation, and voice design preferences. Young adults were selected for their digital fluency and their ability to critically engage with experimental features.

4.1. Participants and Procedure

A total of 43 participants (24 male, 19 female) aged 25–35 engaged with the system. Young adult participants were recruited through announcements at Ghent University, drawing from the university's student and research community. Participation was voluntary and unpaid. The cohort was linguistically diverse, including Dutch speakers from Belgium and the Netherlands, and Spanish speakers from Colombia, Argentina, and Spain, as well as participants who spoke Chinese, Hindi, French, and Brazilian Portuguese. Each participant completed between two and four unstructured conversations with the social robot Pepper, lasting on average 8.56 min (SD = 2.23). Conversations were free-form, encouraging participants to explore topics of interest and test the robot's memory and language capabilities.

Prior to the interaction, participants provided their name, preferred language, and place of origin. This information was used to initialise a personalised user profile and to configure the appropriate STT and TTS settings. For example, a Flemish speaker would hear a voice reflecting their regional accent. STT accuracy was assessed by comparing automated transcripts against manual annotations made by a proficient speaker, either native or highly fluent, who reviewed transcriptions to identify discrepancies and calculate the overall accuracy. After the interaction, participants completed two standardised questionnaires adapted from the existing HRI literature [42,43], assessing perceived naturalness, enjoyment, trust, and intelligence using five-point Likert scales. The setup is illustrated in Figure 2.

Figure 2. Study setup with a young adult participant interacting with Pepper.

4.2. System Performance and Interaction Metrics

Table 1 summarises the interaction metrics gathered during the study. The average response time was 8.96 s (SD = 3.73), with substantial variation due to prompt complexity and server load. The STT system achieved an average accuracy of 85.4%, with lower scores for accented or dialectal speech (e.g., Flemish Dutch), which occasionally caused misinterpretations. Participants often compensated by rephrasing or repeating their utterances.

The interruption rate, which exceeded 40%, revealed shortcomings in the original turn-taking strategy. Interruptions were often caused by misinterpreted pauses or delays in LLM responses that prompted users to restart. These issues led us to refine the VAD mechanism and introduce a user-calibrated silence threshold to better accommodate varied speech rhythms.

Table 1. System performance metrics across all participants during preliminary evaluation with young adults.

Metric	Mean	Standard Deviation
Conversation Duration (min)	8.56	2.23
Avg. Response Time (s)	8.96	3.73
Turn Exchanges	10.6	2.42
Interruption Frequency (%)	40.21	13.14
STT Accuracy (%)	85.37	4.45

4.3. User Perceptions and Design Insights

Subjective ratings captured through standardised questionnaires are shown in Figure 3. Participants generally rated the robot as natural and enjoyable, though scores for trust and intelligence showed greater variability. Feedback revealed important lessons about memory, voice design, and persona management that were not fully anticipated during development.

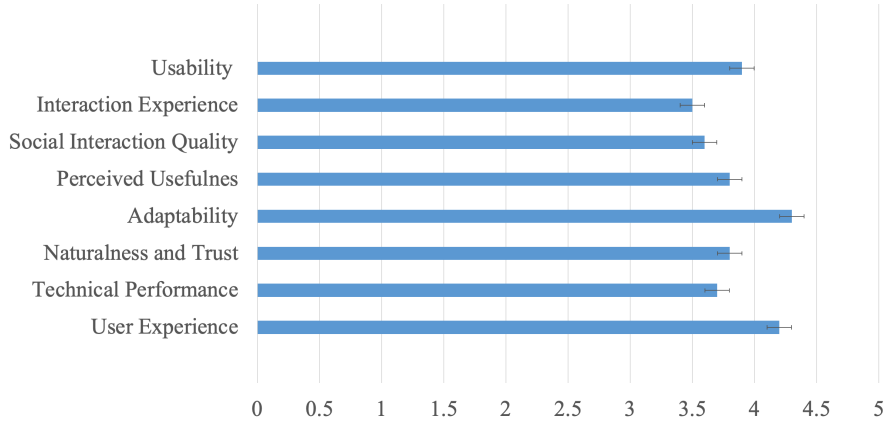

Figure 3. Subjective ratings grouped by constructs from standardised post-interaction questionnaires.

Participants consistently described the robot's memory-driven responses as engaging and more personalised than expected. When the robot recalled information from previous sessions—such as a user's pet or a recent trip—it created a sense of continuity that several users interpreted as social attentiveness. This feedback, mentioned by around a third of participants, suggests that personalised memory contributes to a perception of conversational depth and ongoing familiarity.

Approximately 70% of participants raised concerns about long response delays and conversational disruptions. Several noted that the robot's silent pauses created confusion, and about 60% felt that some responses were excessively long, affecting the exchange dynamic. Although the depth of the robot's responses was generally appreciated, 85% of participants expressed a preference for faster replies. One participant remarked, *"Sometimes, I found myself waiting a bit longer than I would in a typical conversation"*. These observations highlight the importance of response timing in sustaining engagement.

To address these concerns, we implemented incremental streaming of LLM output, integrated a real-time neural TTS engine, and refined the system prompt to explicitly encourage shorter, more focused answers. Together, these changes significantly reduced perceived latency and improved dialogue pacing in subsequent iterations.

Beyond response timing, participants also appreciated the robot's capacity to sustain coherent and natural dialogue. Across all sessions, the model's ability to deliver contextually appropriate answers and maintain conversational flow contributed positively to

engagement. Approximately 86% of participants expressed satisfaction with the robot's responses. One participant noted, *"The robot was able to answer all the spontaneous questions that I had for it, which surprised me, and we were able to have an actual conversation"*.

Participants also evaluated the voice and personality of the robot. Early versions of the system relied on Pepper's default TTS and Microsoft Azure voices. However, users consistently preferred the more expressive and regionalised voices from ElevenLabs. The emotional tone and linguistic alignment made the robot feel more relatable and inclusive, especially for non-native speakers.

A key challenge identified was the loss of persona consistency during longer interactions. Users reported moments when the robot reverted to assistant-like phrasing or broke character by saying *"I am an AI"*, or *"I do not have feelings"*. These hallucinations, although infrequent, disrupted the illusion of a social agent. As a direct result, we introduced a lightweight persona reminder injected into each prompt to reinforce the robot's role, style, and social intent across multi-turn dialogues.

Although rare, participants also encountered occasional hallucinated content—such as vague or contradictory statements and overly generic replies. These disruptions were generally well tolerated, but highlight the importance of prompt engineering and memory-grounding to maintain coherence and trust in long-form interaction.

Participants also suggested potential applications for the robot beyond social interaction, including educational roles and support for isolated individuals. One user noted that *"Talking to a robot felt like something out of science fiction. It was exciting and oddly comforting"*. However, others highlighted limitations such as repetition or momentary API instability. For example, one participant remarked that the conversation *"started to feel repetitive and superficial after a while"*, referring to the robot's tendency to reuse similar follow-up questions or remain on the same topic without introducing new directions. These concerns prompted refinements to the prompt structure to reduce redundancy, improve topic continuity, and better align the robot's responses with the flow of conversation, alongside improvements in summarisation and token management.

In summary, this evaluation confirmed the importance of consistent memory, culturally adapted voice, and stable turn-taking. It also exposed critical gaps in latency handling and persona grounding. These lessons directly shaped the next phase of the system's development for deployment with older adults.

5. Evaluation with Older Adults

To assess the applicability of the system in its intended use context, we conducted a follow-up evaluation with older adults. This exploratory study was not intended to assess long-term behavioural outcomes, but rather to evaluate the robot's short-term performance, perceived social qualities, and technical readiness in real-world conditions with the target demographic. Specifically, we examined dialogue flow, trust, engagement, and the perceived contribution of memory and personalisation.

5.1. Participants and Setup

Seven Dutch-speaking older adults (five female, two male), aged 73–88 (mean = 77.57, SD = 3.82), participated in the study. All were living independently or semi-independently and had no prior experience with robots. Participants were recruited through informal contacts and lived within a 60 km radius of Ghent. No financial compensation was provided for participation. Table 2 summarises their demographic information.

Each participant engaged in two unstructured conversational sessions with Pepper, typically conducted in their home environment. Interactions were open-ended and lasted between 6:27 and 16:31 min (M = 9.93, SD = 3.24), allowing participants to end the session at

will. They were also encouraged to discuss any topic of their choice, fostering spontaneous and natural interaction.

Table 2. Demographic information of older adult participants.

Participant	Age	Gender	Ethnicity
P1	88	Female	Dutch
P2	76	Male	Flemish
P3	75	Female	Flemish
P4	73	Female	Flemish
P5	77	Female	Dutch
P6	79	Female	Flemish
P7	78	Male	Flemish

To support turn-taking, the robot used nonverbal indicators such as eye LED colour and body posture. Participants were informed that when Pepper's eyes turned blue, it was actively listening, and that it could not hear them while it was speaking. This helped to regulate the conversational flow and reduced confusion about robot attention. Researchers were present during the sessions, but remained non-intrusive. Figure 4 shows a typical interaction setup.

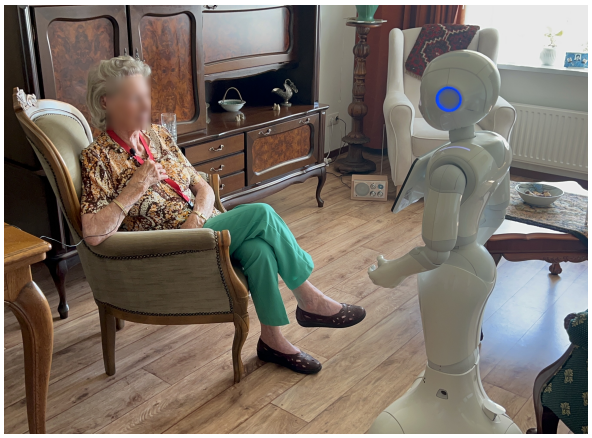

Figure 4. An older adult interacting with Pepper during the study.

After each session, participants engaged in a short semi-structured interview covering trust, usability, emotional response, and perceived limitations. All sessions were audio- and video-recorded with informed consent. The presence of the researcher may have influenced some participants' behaviour—such as increasing hesitancy during personal disclosures or causing shorter responses. Additionally, the novelty of interacting with a humanoid robot appeared to enhance initial enthusiasm and engagement, especially among those unfamiliar with robotics. This aligns with prior findings linking novelty effects to elevated short-term interest [44]. While useful for bootstrapping engagement, this effect underscores the importance of future long-term studies to evaluate sustained interaction patterns.

5.2. System Performance and Interaction Metrics

Table 3 summarises the interaction metrics across participants. To accommodate individual speech rhythms, the system employed a user-calibrated silence threshold—typically around 1.8 s—for detecting turn completions. This parameter was adjusted prior to each session, helping to reduce interruptions and support more fluid interaction. STT accu-

racy was evaluated by comparing automatically generated transcripts against manual annotations made by a native speaker.

Table 3. Interaction metrics across participants: duration, responsiveness, interruptions, and STT accuracy.

Pt.	Sess.	Dur. (min)	Avg. Resp. T. (s)	Turns	Interr. (%)	STT Accu. (%)
P1	I	10:49	6.96	20	10.0	95.3
	II	07:58	6.73	13	7.7	94.3
P2	I	10:26	7.07	18	38.9	82.3
	II	14:35	7.14	20	20.0	86.6
P3	I	07:57	6.69	12	25.0	88.7
	II	06:36	6.94	10	10.0	88.4
P4	I	16:31	7.18	30	10.0	91.3
	II	12:01	7.05	19	5.2	92.3
P5	I	08:01	6.55	14	15.4	86.5
	II	06:27	6.67	12	8.3	89.4
P6	I	09:08	6.82	16	12.5	91.2
	II	10:03	6.55	18	5.5	92.5
P7	I	15:34	7.25	28	5.5	94.2
	II	09:48	7.12	18	0.0	95.4

Turn-taking performance improved across sessions. The average interruption rate was 13.1%, with some sessions achieving zero interruptions. Notably, interruption rates decreased in second sessions for all participants, suggesting growing familiarity with the robot's timing and interaction style.

STT performance remained high (M = 90.62%, SD = 4.36%), even with softer speech or minor dialectal variations. Errors due to microphone positioning or speech volume were often mitigated by the LLM's contextual resilience [45].

5.3. User Perceptions and Design Insights

Participants responded positively to the robot's language capabilities, voice tone, and conversational adaptability. Topics such as hobbies, family, daily routines, and past experiences emerged organically, as shown in Figure 5. Unlike in the young adult evaluation—where repetition was more frequently noted—older adults generally accepted topic persistence and made fewer negative comments about conversational redundancy. This may reflect differences in expectations or the novelty of the interaction.

All participants described the robot as comforting, with four participants comparing it to pets or childhood toys. Three participants who reported experiencing loneliness said they would enjoy having Pepper at home. P6 remarked, *"If I feel alone or sad, I do feel that you could confide in it"*. P4 added, *"If I had a robot like this, I would definitely have a chat with it every morning"*. Even among those who did not express personal interest in using the robot, most recognised its potential value for others, such as friends or individuals living with loneliness or dementia. These reflections align with prior HRI findings that emphasise the role of companionship and social presence in fostering long-term engagement [12].

Voice quality and tone were frequently praised, especially the warmth and clarity of the ElevenLabs-generated voices. Although slower responses were generally tolerated, participants were sensitive to interruptions and valued clear conversational cues. As P2 noted, *"Just like people, Pepper sometimes interrupts"*. P7 added, *"Considering it's a robot, its slightly slower pace is understandable"*.

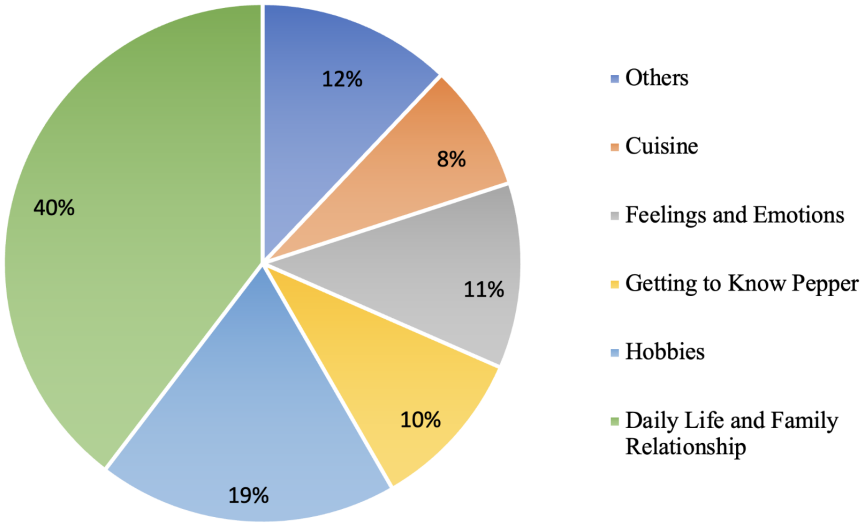

Figure 5. Topics most frequently discussed during interactions with older adults.

Despite overall positive reactions, several challenges emerged. Some participants expressed minor confusion when the robot responded vaguely or failed to maintain the thread of conversation. In one instance, a participant asked the robot to sing, only to receive a spoken recitation of lyrics, prompting the comment, *"It recited the lyrics instead of actually singing"*. These moments highlight the need to better manage expectations and clarify system capabilities to avoid perceived breakdowns. Additionally, while repetition was less problematic for older adults than younger users, occasional fallback phrasing and overly generic questions were still noted.

In summary, this exploratory evaluation demonstrated the system's feasibility and perceived value among older adults. It also underscored the importance of clear turn-taking cues, culturally adapted voices, and expectation alignment for sustained interaction. These findings complement the earlier insights gathered from younger adults and inform the key design considerations discussed in the following section.

6. Discussion

This work explored how LLMs can be integrated into socially assistive robots to enable fluid, personalised, and context-aware dialogue, particularly with older adults. By evaluating the system with both young and older adult populations, we identified technical strengths, design opportunities, and areas requiring refinement. In what follows, we discuss the implications of our findings for future system design, contrasting with the recent literature and outlining necessary design considerations for real-world deployments.

6.1. Personalisation and Memory

Participants across age groups responded positively to the robot's ability to recall personal details, which fostered a sense of continuity, attentiveness, and relational familiarity. This outcome reinforces prior findings that memory continuity can increase emotional resonance and perceived intelligence in HRI. In our study, participants—particularly older adults—often described the memory function as a key reason why the interaction felt "human-like" and meaningful.

Unlike previous systems relying on static user profiles or manually annotated memories [14,34], our architecture implements a layered, prompt-based memory mechanism. It combines dynamic session-level summarisation with a progressively evolving user profile, enabling long-term coherence without exceeding LLM token limits. This finding suggests that memory mechanisms need not be complex or deeply embedded in model training;

prompt-level memory, if carefully structured and iteratively refined, can achieve strong user engagement.

Importantly, while prior studies have highlighted the importance of memory and continuity in user expectations, the systems involved were often tested in constrained settings or lacked validation in real-world contexts. By evaluating memory-driven personalisation with older adults in their home environments, our study offers preliminary evidence of the feasibility and perceived value of longitudinal memory adaptation in naturalistic conditions. However, given the small sample and short-term deployment, broader validation is still needed. While our sample did not permit analysis based on gender or personality traits, future work should examine how such factors influence perceptions of robot memory, empathy, or persona coherence.

6.2. Turn-Taking and Adaptation Across Populations

Effective turn-taking remains one of the most persistent challenges in spoken human–robot interaction [36]. In our evaluation, fixed timing strategies—such as a static silence threshold of 1.2 s—proved inadequate across both age groups. Among younger adults, the system frequently interrupted users who paused mid-thought, mistaking reflective silence for turn completion. This often disrupted the flow of conversation and led participants to repeat themselves or disengage.

The issue was more pronounced with older adults, whose speech rhythms were slower and more variable [46,47]. To address this, we implemented a user-calibrated silence threshold (typically approximately 1.8 s) tailored to individual speaking patterns. This adjustment significantly reduced interruptions, with several sessions achieving zero interruptions and nearly all showing lower rates in second interactions. These results suggest not only system-level improvements, but also mutual adaptation between the user and robot over time, where users learned the robot's pacing and adjusted accordingly.

Our approach contrasts with many prior LLM-based systems relying on static or minimally adaptive timing. For example, Irfan et al. [14] reported frequent interruptions in a GPT-powered companion robot for elderly users, attributing breakdowns to rigid timing logic and insufficient signalling. Similarly, Khoo et al. [31] observed user confusion due to awkward silences and unclear listening cues.

In our system, multimodal indicators—such as LED eye colour to signal listening versus speaking—provided essential feedback and were frequently mentioned by participants as helpful. Nonetheless, some users still hesitated or repeated themselves when faced with longer LLM delays or subtle turn transitions. To reduce these effects, we implemented incremental streaming of LLM output, allowing the robot to begin speaking before the silence threshold was fully reached. This approach reduced latency perception and helped to smooth out the interaction rhythm.

Taken together, these findings reinforce the need for turn-taking models that go beyond simple pause detection. Combining user-calibrated timing, real-time LLM streaming, and transparent multimodal feedback significantly improved dialogue fluidity and user confidence. Future systems should build on these principles by incorporating prosodic cues, discourse markers, and gaze or intent modelling to further enhance naturalness—especially in interactions with older or slower-speaking users.

6.3. Managing Expectations and Maintaining Persona

LLMs are not designed to maintain coherent personas across extended conversations. Without intervention, they often revert to assistant-like behaviour or default disclaimers (e.g., "As an AI..."), which can undermine the illusion of a socially grounded agent. We observed this issue in the young adult group, particularly during longer sessions. Prior

work has similarly reported persona breakdowns [14,31], where users disengaged upon encountering robotic or inconsistent responses.

Our solution—a lightweight persona reminder injected into each prompt—proved effective at maintaining tone and social coherence. Rather than fully re-initialising the persona at each turn (which would consume tokens and reduce room for memory and dialogue context), we used a compressed prompt reinforcing the robot's role (e.g., friendly companion), tone (warm, curious), and behavioural boundaries (e.g., avoiding disclaimers). This intervention noticeably reduced persona drift over time. Users described the robot as "engaged", "relatable", and "empathetic", and later sessions maintained a more consistent and coherent persona throughout the interaction.

Still, prompt engineering alone is insufficient. Future systems should implement real-time monitoring to detect and intercept out-of-character responses, and explore adaptive persona modelling that dynamically adjusts to the interaction context and user feedback.

Importantly, not all failures were linguistic. Some hallucinations were functional mismatches: for example, when a user asked the robot to sing, it simply recited lyrics. Others assumed the robot could perform physical actions it was not equipped for. These episodes underscore the need for clearer signalling of system capabilities. Social robots must manage expectations not just through verbal coherence, but through consistent self-representation and response design.

Response style also played a critical role. While most participants appreciated the robot's fluency, around 60% of young adults and 40% of older adults found some replies too verbose or repetitive. To address this, we revised the system prompt to favour shorter, more open-ended turns that encouraged user elaboration.

6.4. Multilingual and Accentual Adaptation

Language plays a central role in shaping comfort, inclusion, and relatability in HRI [48,49]. In our study, both young and older adult participants expressed appreciation for the robot's ability to converse in their native language—and, notably, in regional variants such as Flemish Dutch. This form of linguistic mirroring, achieved by adapting both speech recognition and text-to-speech models to user profiles, contributed meaningfully to feelings of inclusion, cultural sensitivity, and engagement.

Young adults frequently remarked on the familiarity of the robot's accent and prosody, while older participants—many of whom had no prior experience with AI or robots—described this familiarity as comforting and respectful. Several young adults specifically compared the ElevenLabs voices to default TTS systems, praising them as "warmer" and "less robotic", while older adults also independently described the voice as warm and pleasant. These observations highlight the need for socially assistive robots to not only support multiple languages, but also respect intra-linguistic variation that reflects regional identity.

This represents a notable departure from most prior work, where LLM-driven robots have typically operated in standardised or English-only configurations. For example, some studies have leveraged GPT models for spoken interaction, but not explicitly incorporated dialectal adaptation [14,32]. Even in systems designed for older adults, robots have used generalised English responses, which may reduce the perceived authenticity or relatability of interactions in multilingual or culturally diverse contexts [31]. These findings underscore the importance of localised language support for fostering inclusive HRI.

6.5. Emotional Resonance and Ethical Considerations

Although the system was primarily designed to support naturalistic, memory-driven conversations for social companionship, several older adult participants responded to the

robot in emotionally expressive ways. Descriptions such as "friendly", "comforting", or "like a pet" were common, and three participants who reported feeling lonely stated that they would enjoy having the robot at home. Others saw its potential value for individuals with dementia or limited social contact.

These responses highlight the role of memory, attentiveness, and persona continuity in fostering emotional connection—even without explicit affective modelling. By recalling prior conversations, referencing personal details, and sustaining a coherent persona, the system projected a sense of social presence that many participants found meaningful. This aligns with prior studies that have observed affective responses in elderly users interacting with LLM-based robots [14,31]. Unlike systems limited to scripted empathy or shallow interaction, our robot's consistent tone and contextual continuity contributed to a perception of an ongoing relationship, without relying on explicit emotion modelling.

However, emotional resonance introduces ethical complexity. Some users expressed a desire to confide in the robot or saw it as a source of emotional support. This blurs the boundary between utility and affective deception. If a robot gives the impression of emotional understanding without real empathy, it may foster over-trust or unmet relational needs—especially among vulnerable populations such as older adults or those experiencing loneliness. These concerns echo broader discussions in HRI around emotional deception, anthropomorphism, and the ethics of synthetic empathy [48,50,51].

Future systems aiming to support older adults should incorporate safeguards for emotional transparency, while also recognising that warmth, responsiveness, and memory can naturally give rise to a sense of companionship—even in LLM-powered agents.

7. Technical Challenges and Design Recommendations

Despite positive outcomes in both evaluations, our system faced several technical limitations that constrained the consistency and realism of the interaction experience, especially under real-world deployment conditions.

Response delays remained one of the most frequently reported issues. Although the integration of real-time streaming and neural TTS substantially improved responsiveness, the system still experienced variable lag due to API communication and LLM inference times, particularly during complex user queries. Unlike in scripted systems, latency in LLM-powered dialogue cannot be fully eliminated, but our findings show that incremental output and multimodal feedback (e.g., LED indicators) can help to mitigate user confusion.

Speech recognition performance was generally strong (85–91% accuracy), but still vulnerable to microphone distance, background noise, and dialectal variation. Notably, the LLM often sustained the dialogue by relying on contextual cues, even when transcriptions were partially inaccurate. This resilience—previously highlighted in related studies on LLM-based dialogue systems [45]—allowed interactions to remain coherent despite occasional recognition errors. However, persistent misrecognitions still disrupted some exchanges, particularly with softer or heavily accented speech.

Turn-taking was also affected by these inconsistencies. While LED cues and calibrated silence thresholds helped to signal the conversational state, the system occasionally misjudged end-of-turns, leading to interruptions or awkward silences—particularly under variable speech pacing or longer LLM delays. As discussed earlier, addressing this limitation will require more context-aware turn-taking strategies that account for discourse structure, prosody, or speaker intent [52–54].

Participants occasionally expected richer physical expressiveness (e.g., gestures, singing, touch) which the robot could not perform. This reveals a gap between embodied appearance and functional capability. Designers should consider how verbal content aligns with embodiment and explore multimodal synchronisation for higher realism.

Memory transparency also remains an open challenge. While personalised recall enhanced engagement, users had no direct visibility into what was stored or how it was used. As discussed earlier, long-term deployments should include interfaces for inspecting, editing, or deleting stored data, ensuring user agency and supporting ethical alignment in socially assistive contexts.

While our study has already incorporated early-stage user feedback from both younger and older adults, further involvement of target users in earlier phases of design—such as through co-design workshops or participatory prototyping—could help to reveal deeper assumptions and refine expectations. Iterative collaboration with end users is especially important for socially assistive systems intended for long-term companionship, helping to align functionality with lived experience and build trust in real-world settings [55]. As the system matures toward broader deployment, expanding participatory design practices will be essential to ensure ethical alignment and user acceptance.

From a scalability perspective, deploying such a system in real-world care settings presents several practical challenges. These include the cost and availability of humanoid robots like Pepper, the need for stable internet connectivity to support cloud-based LLMs, and the complexity of maintaining multilingual, personalised models at scale. However, with the increasing availability of efficient LLM APIs and the potential to migrate this architecture to lower-cost models, future iterations could be made more accessible and feasible for broader deployment in care environments.

8. Conclusions

This study presented the design and evaluation of a socially assistive robot powered by an LLM, with a focus on mitigating social isolation among older adults through natural, personalised conversation. Through a two-phase evaluation—first with young adults and then with older adults in home environments—we examined the system's performance, usability, and emotional resonance.

Our findings demonstrate that memory-based personalisation, adaptive turn-taking, and multilingual voice adaptation are key enablers of engaging and inclusive human–robot dialogue. Participants responded positively to the robot's ability to remember personal details, speak in regional accents, and maintain socially appropriate conversation flow. However, the study also highlighted persistent challenges, including occasional interruptions, response delays, and moments of unrealistic user expectations—particularly regarding the robot's expressive and functional capabilities.

By validating our system with older adults in naturalistic settings, we have made a step forward in understanding how LLMs can be embedded into socially meaningful, real-world interactions. Our design choices—such as streaming response generation, persona reinforcement prompts, and user-calibrated silence thresholds—offer practical insights for building future LLM-integrated robots that are emotionally engaging, ethically grounded, and robust to the variability of human dialogue.

Future work will explore long-term deployment, explicit emotion modelling, and strategies for co-adaptive memory and user feedback. Ultimately, we envision social robots that not only understand and respond, but also remember, adapt, and connect, therefore supporting meaningful companionship in everyday life.

Author Contributions: Conceptualisation, M.P.-B. and T.B.; methodology, M.P.-B.; software, M.P.-B. and M.B.; validation, M.P.-B. and M.B.; formal analysis, M.P.-B.; investigation, M.P.-B.; resources, M.P.-B. and T.B.; data curation, M.P.-B.; writing—original draft preparation, M.P.-B.; writing—review and editing, M.P.-B. and T.B.; visualisation, M.P.-B.; supervision, T.B.; project administration, T.B.; funding acquisition, T.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Bijzonder Onderzoeksfonds (BOF), grant number BOF22/DOC/235.

Institutional Review Board Statement: The study design, data management plan, consent form, and experimental protocol were approved by the Departmental Ethics Committee of the Faculty of Psychology and Educational Sciences, Ghent University (ref. 2023-8, approved on 23 January 2024).

Informed Consent Statement: Informed consent was obtained from all the subjects involved in the study. Written informed consent has also been obtained from the participants for the publication of any potentially identifiable data or images.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

HRI	Human–Robot Interaction
SAR	Socially Assistive Robot
LLM	Large Language Model
STT	Speech-to-Text
TTS	Text-to-Speech
VAD	Voice Activity Detection

References

1. Tam, M.T.; Dosso, J.A.; Robillard, J.M. The impact of a global pandemic on people living with dementia and their care partners: Analysis of 417 lived experience reports. *J. Alzheimer's Dis.* **2021**, *80*, 865–875. [CrossRef] [PubMed]
2. Sutin, A.R.; Stephan, Y.; Luchetti, M.; Terracciano, A. Loneliness and risk of dementia. *J. Gerontol. Ser. B* **2020**, *75*, 1414–1422. [CrossRef] [PubMed]
3. der Velden, P.G.V.; Hyland, P.; Contino, C.; von Gaudecker, H.M.; Muffels, R.; Das, M. Anxiety and Depression Symptoms, the Recovery from Symptoms, and Loneliness Before and After the COVID-19 Outbreak Among the General Population: Findings from a Dutch Population-Based Longitudinal Study. *PLoS ONE* **2021**, *16*, e0245057. [CrossRef]
4. Soh, Y.; Kawachi, I.; Kubzansky, L.D.; Berkman, L.F.; Tiemeier, H. Chronic Loneliness and the Risk of Incident Stroke in Middle and Late Adulthood: A Longitudinal Cohort Study of US Older Adults. *eClinicalMedicine* **2024**, *73*, 102639. [CrossRef]
5. Age UK Cambridgeshire and Peterborough. Mental Health Awareness. 2022. Available online: https://www.ageuk.org.uk/cambridgeshireandpeterborough/about-us/news/articles/2022/mental-health-awareness/ (accessed on 9 May 2025).
6. Worcestershire County Council. Loneliness Needs Assessment. 2022. Available online: https://www.worcestershire.gov.uk/sites/default/files/2022-11/Loneliness_Needs_Assessment_Final.pdf (accessed on 9 May 2025).
7. Svensson, M.; Rosso, A.; Elmståhl, S.; Ekström, H. Loneliness, Social Isolation, and Health Complaints Among Older People: A Population-Based Study from the 'Good Aging in Skåne (GÅS)' Project. *SSM-Popul. Health* **2022**, *20*, 101287. [CrossRef] [PubMed]
8. Guarnera, J.; Yuen, E.; Macpherson, H. The Impact of Loneliness and Social Isolation on Cognitive Aging: A Narrative Review. *J. Alzheimer's Dis. Rep.* **2023**, *7*, 699–714. [CrossRef]
9. Donovan, N.J.; Blazer, D. Social Isolation and Loneliness in Older Adults: Review and Commentary of a National Academies Report. *Am. J. Geriatr. Psychiatry* **2020**, *28*, 1233–1244. [CrossRef]
10. Kim, J.; Kim, S.; Kim, S.; Lee, E.; Heo, Y.; Hwang, C.Y.; Choi, Y.Y.; Kong, H.J.; Ryu, H.; Lee, H. Companion robots for older adults: Rodgers' evolutionary concept analysis approach. *Intell. Serv. Robot.* **2021**, *14*, 729–739. [CrossRef]
11. Baecker, A.N.; Geiskkovitch, D.Y.; González, A.L.; Young, J.E. Emotional support domestic robots for healthy older adults: Conversational prototypes to help with loneliness. In Proceedings of the Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, UK, 23–26 March 2020; pp. 122–124.
12. Cifuentes, C.A.; Pinto, M.J.; Céspedes, N.; Múnera, M. Social robots in therapy and care. *Curr. Robot. Rep.* **2020**, *1*, 59–74. [CrossRef]
13. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
14. Irfan, B.; Kuoppamäki, S.; Hosseini, A.; Skantze, G. Between reality and delusion: Challenges of applying large language models to companion robots for open-domain dialogues with older adults. *Auton. Robot.* **2025**, *49*, 9. [CrossRef]

15. Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9*, 36–45. [CrossRef]
16. Breazeal, C. Emotion and sociable humanoid robots. *Int. J. Hum.-Comput. Stud.* **2003**, *59*, 119–155. [CrossRef]
17. Fong, T.; Nourbakhsh, I.; Dautenhahn, K. A survey of socially interactive robots. *Robot. Auton. Syst.* **2003**, *42*, 143–166. [CrossRef]
18. Shawar, B.A.; Atwell, E. Chatbots: Are they really useful? *LDV Forum* **2007**, *22*, 29–49.
19. Cassell, J.; Bickmore, T.; Billinghurst, M.; Campbell, L.; Chang, K.; Vilhjálmsson, H.; Yan, H. Embodiment in conversational interfaces: Rea. In Proceedings of the CHI99: Conference on Human Factors in Computing Systems, Pittsburgh, PA, USA, 15–20 May 1999; pp. 520–527.
20. Sidner, C.L.; Lee, C.; Kidd, C.D.; Lesh, N.; Rich, C. Explorations in engagement for humans and robots. *Artif. Intell.* **2005**, *166*, 140–164. [CrossRef]
21. Roller, S.; Boureau, Y.L.; Weston, J.; Bordes, A.; Dinan, E.; Fan, A.; Gunning, D.; Ju, D.; Li, M.; Poff, S.; et al. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv* **2020**, arXiv:2006.12442.
22. De Greeff, J.; Belpaeme, T. Why robots should be social: Enhancing machine learning through social human-robot interaction. *PLoS ONE* **2015**, *10*, e0138061. [CrossRef]
23. Russo, A.; D'Onofrio, G.; Gangemi, A.; Giuliani, F.; Mongiovi, M.; Ricciardi, F.; Greco, F.; Cavallo, F.; Dario, P.; Sancarlo, D.; et al. Dialogue systems and conversational agents for patients with dementia: The human–robot interaction. *Rejuvenation Res.* **2019**, *22*, 109–120. [CrossRef]
24. Zhang, S.; Zhao, X.; Lei, B. Speech emotion recognition using an enhanced kernel isomap for human-robot interaction. *Int. J. Adv. Robot. Syst.* **2013**, *10*, 114. [CrossRef]
25. Deng, J.; Pang, G.; Zhang, Z.; Pang, Z.; Yang, H.; Yang, G. cGAN Based Facial Expression Recognition for Human-Robot Interaction. *IEEE Access* **2019**, *7*, 9848–9859. [CrossRef]
26. Cohn, J.F.; De la Torre, F. *The Oxford Handbook of Affective Computing*; Automated Face Analysis for Affective Computing; Oxford University Press: New York, NY, USA, 2014.
27. Turk, M. Multimodal interaction: A review. *Pattern Recognit. Lett.* **2014**, *36*, 189–195. [CrossRef]
28. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165.
29. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
30. Elgarf, M.; Skantze, G.; Peters, C. Once upon a story: Can a creative storyteller robot stimulate creativity in children? In Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, Virtual Event, 14–17 September 2021; pp. 60–67.
31. Khoo, W.; Hsu, L.J.; Amon, K.J.; Chakilam, P.V.; Chen, W.C.; Kaufman, Z.; Lungu, A.; Sato, H.; Seliger, E.; Swaminathan, M.; et al. Spill the Tea: When Robot Conversation Agents Support Well-being for Older Adults. In Proceedings of the Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, Stockholm, Sweden, 13–16 March 2023; pp. 178–182.
32. Billing, E.; Rosén, J.; Lamb, M. Language models for human-robot interaction. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Stockholm, Sweden, 13–16 March 2023; ACM Digital Library: New York, NY, USA, 2023; pp. 905–906.
33. Axelsson, A.; Skantze, G. Do you follow? A fully automated system for adaptive robot presenters. In Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, Stockholm, Sweden, 13–16 March 2023; pp. 102–111.
34. Kang, H.; Moussa, M.B.; Magnenat-Thalmann, N. Nadine: An LLM-driven intelligent social robot with affective capabilities and human-like memory. *arXiv* **2024**, arXiv:2405.20189.
35. Mishra, C.; Verdonschot, R.; Hagoort, P.; Skantze, G. Real-time emotion generation in human-robot dialogue using large language models. *Front. Robot. AI* **2023**, *10*, 1271610. [CrossRef]
36. Skantze, G. Turn-taking in conversational systems and human-robot interaction: A review. *Comput. Speech Lang.* **2021**, *67*, 101178. [CrossRef]
37. Marge, M.; Espy-Wilson, C.; Ward, N.G.; Alwan, A.; Artzi, Y.; Bansal, M.; Blankenship, G.; Chai, J.; Daumé, H., III; Dey, D.; et al. Spoken language interaction with robots: Recommendations for future research. *Comput. Speech Lang.* **2022**, *71*, 101255. [CrossRef]
38. Irfan, B.; Kuoppamäki, S.; Skantze, G. Recommendations for designing conversational companion robots with older adults through foundation models. *Front. Robot. AI* **2024**, *11*, 1363713. [CrossRef]
39. Clark, H.H.; Fischer, K. Social robots as depictions of social agents. *Behav. Brain Sci.* **2023**, *46*, e21. [CrossRef]
40. Aldebaran Robotics. NAOqi Framework Documentation. 2015. Available online: http://doc.aldebaran.com/2-5/naoqi/index.html (accessed on 10 October 2023).
41. Meyer, A.S. Timing in conversation. *J. Cogn.* **2023**, *6*, 20. [CrossRef] [PubMed]

42. Syrdal, D.S.; Dautenhahn, K.; Koay, K.L.; Walters, M.L. The Negative Attitudes Towards Robots Scale and Reactions to Robot Behaviour in a Live Human-Robot Interaction Study. Adaptive and Emergent Behaviour and Complex Systems. 2009. Available online: https://www.semanticscholar.org/paper/The-Negative-Attitudes-Towards-Robots-Scale-and-to-Syrdal-Dautenhahn/dbf813f5a38c36155cd22485d62cbfcdcc01164a (accessed on 15 November 2023).
43. Bartneck, C.; Kulić, D.; Croft, E.; Zoghbi, S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* **2009**, *1*, 71–81. [CrossRef]
44. Smedegaard, C.V. Reframing the Role of Novelty within Social HRI: From Noise to Information. In Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Daegu, Republic of Korea, 11–14 March 2019.
45. Verhelst, E.; Belpaeme, T. Large Language Models Cover for Speech Recognition Mistakes: Evaluating Conversational AI for Second Language Learners. In Proceedings of the 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Melbourne, Australia, 4–6 March 2025; pp. 1705–1709. [CrossRef]
46. Pearson, D.V.; Shen, Y.; McAuley, J.D.; Kidd, G.R. Differential sensitivity to speech rhythms in young and older adults. *Front. Psychol.* **2023**, *14*, 1160236. [CrossRef]
47. Pellegrino, E.; He, L.; Dellwo, V. The effect of ageing on speech rhythm: A study on Zurich German. In Proceedings of the Speech Prosody 2018, Poznań, Poland, 13–16 June 2018; ISCA: Singapore, 2018.
48. de Graaf, M.M.A.; Allouch, S.B.; Klamer, T. Sharing a life with Harvey: Exploring the acceptance of and relationship-building with a social robot. *Comput. Hum. Behav.* **2015**, *43*, 1–14. [CrossRef]
49. Trovato, G.; Zecca, M.; Sessa, S.; Takanishi, A.; Hashimoto, K. Cross-cultural study on human–robot greeting interaction: Acceptance and discomfort by Egyptians and Japanese. *Paladyn J. Behav. Robot.* **2016**, *7*, 45–55. [CrossRef]
50. Van Wynsberghe, A. Designing Robots for Care: Care Centered Value-Sensitive Design. *Sci. Eng. Ethics* **2013**, *19*, 407–433. [CrossRef]
51. Turkle, S. *Alone Together: Why We Expect More from Technology and Less from Each Other*; Basic Books: New York, NY, USA, 2011.
52. Skantze, G.; Irfan, B. Applying General Turn-taking Models to Conversational Human-Robot Interaction. *arXiv* **2025**, arXiv:2501.08946.
53. Pinto, M.J.; Belpaeme, T. Predictive turn-taking: Leveraging language models to anticipate turn transitions in human-robot dialogue. In Proceedings of the 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN), Pasadena, CA, USA, 26–30 August 2024; IEEE: New York, NY, USA, 2024; pp. 1733–1738.
54. Inoue, K.; Lala, D.; Skantze, G.; Kawahara, T. Yeah, Un, Oh: Continuous and Real-time Backchannel Prediction with Fine-tuning of Voice Activity Projection. *arXiv* **2024**, arXiv:2410.15929.
55. Winkle, K.; Senft, E.; Lemaignan, S. LEADOR: A method for end-to-end participatory design of autonomous social robots. *Front. Robot. AI* **2021**, *8*, 704119. [CrossRef]