Autonomous Robots (2025) 49:9
https://doi.org/10.1007/s10514-025-10190-y

Between reality and delusion: challenges of applying large language models to companion robots for open-domain dialogues with older adults

Bahar Irfan[1] · Sanna Kuoppamäki[2] · Aida Hosseini[2] · Gabriel Skantze[1]

Received: 2 May 2023 / Accepted: 12 January 2025 / Published online: 10 March 2025

Abstract

Throughout our lives, we interact daily in conversations with our friends and family, covering a wide range of topics, known as open-domain dialogue. As we age, these interactions may diminish due to changes in social and personal relationships, leading to loneliness in older adults. Conversational companion robots can alleviate this issue by providing daily social support. Large language models (LLMs) offer flexibility for enabling open-domain dialogue in these robots. However, LLMs are typically trained and evaluated on textual data, while robots introduce additional complexity through multi-modal interactions, which has not been explored in prior studies. Moreover, it is crucial to involve older adults in the development of robots to ensure alignment with their needs and expectations. Correspondingly, using iterative participatory design approaches, this paper exposes the challenges of integrating LLMs into conversational robots, deriving from 34 Swedish-speaking older adults' (one-to-one) interactions with a personalized companion robot, built on Furhat robot with GPT−3.5. These challenges encompass disruptions in conversations, including frequent interruptions, slow, repetitive, superficial, incoherent, and disengaging responses, language barriers, hallucinations, and outdated information, leading to frustration, confusion, and worry among older adults. Drawing on insights from these challenges, we offer recommendations to enhance the integration of LLMs into conversational robots, encompassing both general suggestions and those tailored to companion robots for older adults.

Keywords Large language models · Companion robot · Elderly care · Open-domain dialogue · Socially assistive robot · Participatory design

1 Introduction

With more than 1 billion people over 60 worldwide,[1] there is a growing need for innovative solutions that can improve the quality of life of older adults. In particular, loneliness in older adults is a risk factor that negatively influences mental and physical health, leading to depression, lower quality of life, decline in health, and mortality rates (Cacioppo et al., 2006; Luo et al., 2012). Companion robots are targeted to enhance the well-being, quality of life, and independence of older adults, by providing service and companionship and assisting (e.g., carrying out a variety of tasks) in everyday life (Kim et al., 2021; Dautenhahn, 2007). Their designed functionalities may include cognitive and social support, support for mobility, health monitoring, and care. Several studies have shown their benefits in reducing social isolation and loneliness, thereby significantly contributing to improving the quality of life or well-being of older adults (Kim et al., 2021).

'Participatory design' (or 'co-design') (Šabanović, 2010) approaches have been recognized as a powerful tool for developing technologies that meet the needs and preferences of end-users. These approaches emphasize collaboration between designers and end-users, as well as iterative proto-

[1] https://www.who.int/health-topics/ageing.

Bahar Irfan
birfan@kth.se

Sanna Kuoppamäki
sannaku@kth.se

Aida Hosseini
idaho@kth.se

Gabriel Skantze
skantze@kth.se

[1] Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm 100 44, Sweden

[2] Division of Health Informatics and Logistics, KTH Royal Institute of Technology, Stockholm 100 44, Sweden

typing, interviews, and testing, to ensure that the final product is usable, useful, and desirable for the target population. Participatory design approaches can help ensure that companion robots are tailored to the needs and preferences of older adults (Lee et al., 2017; Stegner et al., 2023; Kuoppamäki et al., 2023), such that they can be effective in facilitating social interaction and engagement in the user.

In order to integrate companion robots into the daily lives of older adults, they need to be fully autonomous, and provide a natural way to interact without the older adults having to learn how to use them. Hence, spoken dialogue in companion robots is integral, and needs to be flexible enough to adapt to unforeseen circumstances during the conversation (Fernández-Rodicio et al., 2020). Large language models (LLMs) provide that flexibility with their open-domain dialogue capabilities (Zhao et al., 2023) (i.e., conversing on any topic) that was previously lacking in rule-based and probabilistic dialogue architectures, commonly employed in human-robot interaction (HRI) studies (Reimann et al., 2023). However, they are typically trained and evaluated on textual data, whereas robots carry the additional complexity of multi-modal interactions, such as speech and visual cues. The limitations caused by these modalities, such as errors in turn-taking, speech recognition, and generation, should be identified to ensure that the robot's non-verbal cues align with the verbal content. To enable richer and more contextually aware conversations, these limitations should be considered for developing LLM architectures that are suitable for conversational robots. Furthermore, this multi-modal complexity introduces issues related to user expectations, user experience, and system performance (Tatarian et al., 2022). Addressing these challenges is essential for harnessing the full potential of LLMs in conversational robots. Moreover, LLMs are still far from perfect, with significant limitations in their ability to generate coherent and factual responses in open-domain dialogue (Ji et al., 2023). That is why, research in open-domain dialogue systems, such as for chatbots (Shuster et al., 2022) and the Alexa Prize SocialBot Grand Challenge,[2] focuses on modular architectures that overcome these failures with hand-crafting, fine-tuning, or filtering solutions. However, these components bring about confounds, such as the accuracy of the additional solutions, as well as the high computational power required, which is not typically available on robots. LLMs can instead be used in a zero-shot fashion as 'scarecrows' (Williams et al., 2023), i.e., black-box modules that quickly enable full-pipeline solutions for HRI, similar to using humans to operate robots in a Wizard of Oz (Kelley, 1984) approach, fostering the development of better algorithms for companion robots. However, no prior research evaluated their performance and the corresponding challenges when deployed on conversational robots, interacting with older adults.

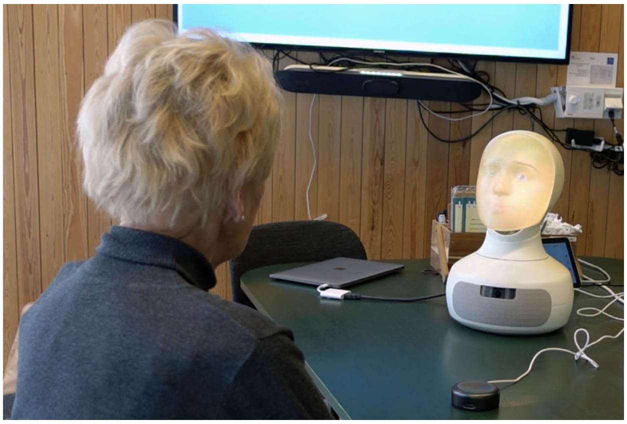

Fig. 1 Older adult interacting with the autonomous Furhat robot with GPT−3.5 during the participatory design workshop

Balancing between technical advancements and user-centered design principles, this article contributes a novel analysis of the challenges of applying LLMs to conversational robots for multi-modal open-domain dialogue, in the context of older adults' interactions with a personalized companion robot (Fig. 1), using iterative participatory design with bottom-up development, based on a widely used LLM, GPT−3.5. Our approach derives insights into the technical development of companion robots through an initial study with 6 Swedish-speaking older adults (65 and older) and a subsequent participatory design workshop with 28 Swedish-speaking older adults. Through a combination of quantitative (for speech recognition performance, LLM latency, questionnaire results, conversation topic initiation and duration) and qualitative analysis (for dialogue disruptions, topic detection, and interviews), encompassing one-to-one conversations, questionnaires, and semi-structured interviews, we uncover challenges deriving from the multi-modal aspects of the robot, spanning turn-taking, repetitive responses, superficial conversations, language barriers, hallucinations, obsolete information, and disengagements. In light of the insights gained from these challenges and the feedback from participants, the paper advances broad recommendations for each theme to expedite the integration of LLMs into conversational robots, in addition to underlining recommendations specific to companion robots for older adults.

2 Background

2.1 Conversational companion robots for older adults

Conversational user interfaces have been shown to reduce technology adoption barriers that older adults typically experience with computing devices, indicating that conversational

[2] https://www.amazon.science/alexa-prize/socialbot-grand-challenge.

companion robots are suitable to be integrated into the daily lives of older adults (Pradhan et al., 2020). However, older adults also perceive unique challenges in interacting with conversational agents, from choosing an appropriate linguistic style and content with the agent (Sayago et al., 2019) to improving the accessibility with hearing impairments (Blair & Abdullah, 2019). Therefore, speech-based interaction should consider questions of synthesis choices and conversation content in responding to the experiences of this age group (Sayago et al., 2019).[3] The linguistic content and default voices that conversational agents provide when interacting with older adults should be designed appropriately from the conversational user experience, and to support a more inclusive interaction.

The design of conversational agents aimed at improving the daily lives of older adults encompasses various aspects, such as promoting daily reflection, strengthening family bonds, and introducing new experiences and volunteer opportunities for their involvement (Randall et al., 2022). Correspondingly, companion robots can evoke feelings of independence and empowerment for older adults (Abdolrahmani et al., 2018). In addition, older adults tend to anthropomorphize the agent by using polite greetings when communicating with the agent, while younger adults tend to consider it as a tool by placing more importance on its convenience (Chung et al., 2019). Older adults also use voice assistants for specific purposes, such as seeking online information. However, there are age-specific challenges when interacting with conversational agents, such as concerns about the reliability and trust towards the agent, and unpredictability, unclarity, and inconsistency of voice commands (Pradhan et al., 2020).

Gollasch & Weber (2021) identified age-specific strategies in dialogue systems and speech recognition accuracy. To respond to the needs of older adults, conversational agents should be able to correctly recognize even unusual formulations; complex dialogues comprising multiple pieces of information should be presented as simple or guided dialogues; agent should ask only one question per dialogue with a limited set of possible answers; it should be able to keep information about the conversation context.

2.2 Participatory design with older adults

Participatory design (also known as co-design) (Šabanović, 2010) builds on participants' self-identified issues and concerns, which are taken as a starting point for developing robotic applications. Participants are given the possibility to interpret the capabilities of robotic systems and discuss the potential social consequences and meanings of robots in daily life contexts (Šabanović et al., 2015). This aims to promote end-users as designers, rather than only users of robotic technologies (Lee et al., 2017), which is particularly important for older adults such that we can understand their expectations and create robots that are suitable to their needs (Frennert & Östlund, 2014; Søraa et al., 2022). Participatory design can involve focus groups (Jenkins & Draper, 2015; Lee et al., 2017; Winkle et al., 2018; Søraa et al., 2022), surveys and interviews (Caleb-Solly et al., 2014; Winkle et al., 2018; Ostrowski et al., 2019; Ostrowski et al., 2021; Søraa et al., 2022; Gasteiger et al., 2022), concept generation and design activities (e.g., storyboards (Bedaf et al., 2019; Björling & Rose, 2019), card sorting (Ostrowski et al., 2019), sketching (Rehm et al., 2016; Lee et al., 2017; Alves-Oliveira et al., 2021), role-playing (Björling & Rose, 2019)), prototyping (Azenkot et al., 2016; Björling & Rose, 2019; Lee et al., 2017)), and interactions with designed robots (Rehm et al., 2016; Ostrowski et al., 2021; Gasteiger et al., 2022; Stegner et al., 2023). Participatory design workshops refer to using a combination of some of these methods, bringing together end-users and researchers. However, participants may not have experience with robotic technologies, and they may not see themselves as designers of any technology (Randall et al., 2018). Therefore, participatory design requires mutual trust and understanding of the everyday life conditions among older adults, and a reflective approach towards designing companion robots with older adults (Lee et al., 2017).

Participatory design has been offered as a solution to overcome challenges in the design process of companion robots with older adults (Jenkins & Draper, 2015; Lee et al., 2017; Ostrowski et al., 2021; Gasteiger et al., 2022; Stegner et al., 2023; Rogers et al., 2022). However, only one study (Ostrowski et al., 2021) targeted co-designing autonomous conversational robots for older adults, but they did not incorporate LLMs. Ours is the first study that uses participatory design with older adults to explore the integration of LLMs into conversational companion robots.

2.3 Large language models for open-domain dialogue

The term 'open-domain' is often used to refer to dialogue systems that are more unrestricted when it comes to the topic of the conversation, compared to systems that are targeted towards more specific domains, such as restaurant booking or language learning (Deriu et al., 2020; Adiwardana et al., 2020; Roller et al., 2020). Whereas task-oriented dialog systems have traditionally been designed using a modular architecture (Natural Language Understanding, Dialogue Management, Natural Language Generation), current open-domain chatbots are typically implemented in an end-to-end fashion using LLMs (Zhao et al., 2023). These LLMs are trained to do next-token prediction on large amounts of text

[3] Furhat Robotics: https://furhatrobotics.com/.

data and then used to predict the system response word-by-word, autoregressively (Vinyals & Le, 2015). Earlier LLMs for chatbots, such as Meena (Adiwardana et al., 2020) from Google, used social media conversations as training data. Similarly, BlenderBot from Meta (Roller et al., 2020) was initially based on training data collected from Reddit, with later versions fine-tuned on other publicly available datasets for performing specific functionalities, such as browsing the internet, personalizing conversations, and maintaining coherent responses with a long-term memory (Shuster et al., 2022). Other recent chatbots, like Bard[4] (from Google) powered by LaMDA (Thoppilan et al., 2022), have been trained using larger, more general datasets (including both dialogue and other public web documents). Similarly, GPT−3.5 (Brown et al., 2020) and ChatGPT[5] from OpenAI are general-purpose language models that can be used as chatbots in a 'zero-shot' fashion, using the last few turns of the dialogue, together with a description of how the agent should behave, as a 'prompt' that the model then makes its word-by-word predictions from.

While general language models can be used directly as chatbots, their responses will reflect ordinary language use, which might not always align with the desired output in terms of, for example, truthfulness and toxicity (the so-called 'alignment problem'). To address this, LaMDA was fine-tuned to optimize human ratings of safety and other qualitative metrics (Thoppilan et al., 2022). A more sophisticated approach was taken for InstructGPT (Ouyang et al., 2022), which uses 'reinforcement learning from human feedback' (RLHF), where a model of human raters is used during reinforcement learning to optimize the model towards the desired criteria. The RLHF approach was also used when training the chatbot ChatGPT[6] (OpenAI).

The term 'open-domain' has been questioned by Doğruöz & Skantze (2021), since it is not clear what the boundaries of this 'openness' are, and whether systems that are described in that way are truly open to all the different forms of dialogue that humans engage in, such as persuasion, asking for favors, small talk, recapping events, and making plans (Goldsmith & Baxter, 1996). However, when analyzing the Google Meena chatlogs, most interactions were found to be simply small talk, not exhibiting this variety of topics (Doğruöz & Skantze, 2021). One explanation for this is the lack of common ground (Clark, 1996) in human–machine interaction, which naturally restricts the user's expectations of what is appropriate and meaningful to talk about. Often, when open-domain chatbots are evaluated, the user is simply instructed to "chat with the system", without any further context (Adiwardana et al., 2020; Thoppilan et al., 2022). Thus, the setting of the interaction, and the user's expectations of that setting, are very important for how the dialogue will unfold and ultimately whether it will be perceived as meaningful by the user.

2.4 Large language models in social robots

LLMs have so far demonstrated their utility across a range of applications in social robotics, such as simulating zero-shot human models (Zhang & Soh, 2023), enabling empathetic non-verbal cues (Lee et al., 2023), acting as a receptionist (Cherakara et al., 2023; Yamazaki et al., 2023), presenting adaptively (Axelsson & Skantze, 2023), supporting multi-party interactions (Murali et al., 2023), storytelling (Elgarf et al., 2021), collecting self-reported user data (Wei et al., 2023), and promoting the well-being of older adults (Khoo et al., 2023). The most similar work to ours is Khoo et al. (2023), which used a fine-tuned GPT-3 with the personalized QT robot for open-domain dialogue with (7) older adults. While the majority of study participants reported a positive and enjoyable interaction with the robot, expressing feelings of comfort and perceiving it as friendly, 3 out of 7 participants did not want to use the robot in their homes. Some participants found the responses slow, and one participant indicated that it might be better suited for older adults living alone and dealing with dementia rather than healthy older adults. In contrast to that study, our work provides perspectives of older adults in addition to the technical evaluation of LLMs applied to companion robots, particularly through an iterative co-design approach. In addition, this study was published after both of our studies had been conducted.

Multi-modal interaction in the context of conversational robots involves the simultaneous use of multiple communication channels, such as text, speech, gestures, and visual cues. This challenge is further amplified by the need to combine the linguistic capabilities of LLMs with the physical and sensory capabilities of robots, which is not considered in chatbots. However, due to the tight coupling between the modalities in the robot, failures in one component (e.g., speech recognition) can affect the performance of the other components (e.g., dialogue manager) (Funakoshi et al., 2007; Lala et al., 2017; Inoue et al., 2020; Quinderé et al., 2013; Reimann et al., 2023). Thus, assessing the various components of the system helps identify potential points of failure and determine the relative significance of each component (Shervedani et al., 2022). Hence, our work provides an overview of the performance of each component in the robot and how they affect or are affected by the LLMs to develop solutions for LLMs with their limitations in mind.

In addition, modalities should be designed with user attributes in mind, such as age, gender, and prior experience, and the target population should be able to easily understand and operate the system (Lazaro et al., 2021). In fact, there is

[4] https://bard.google.com/.

[5] https://chat.openai.com/.

[6] https://openai.com/blog/chatgpt/.

a lack of user-centered research on multi-modal aspects in robots (Lazaro et al., 2021), which this work aims to address for conversational robots incorporated with LLMs.

3 Integrating a large language model with a social robot

As described in the previous sections, LLMs enable open-domain dialogues, but have been rarely explored outside of single-session interactions with short durations, nor with older adults. In addition, prior to our work, LLMs were not evaluated on companion robots for open-domain dialogue. Hence, it is challenging to design a personalized companion robot with LLMs without knowing their limitations in prolonged conversations and the requirements for the target population. Correspondingly, we kept the integration incremental (i.e., with minor changes), with bottom-up methodology using an LLM in a zero-shot fashion and basic robot functionality to identify their drawbacks without any confounds that additional methods such as turn-taking algorithms, fact-checking strategies, summarization methods, and external long-term memory can introduce. The resulting design choices for a fully autonomous personalized companion robot for adults are described in this section, and shown in Fig. 2. The first interaction with the robot can be seen in the video excerpt from the first study.[7]

3.1 Robot

In conversations, gaze is an important element that signals the addressee of the attention, helps coordinate turn-taking in conversation, helps disambiguate references to objects, and establishes joint attention (Kendon, 1967; Skantze et al., 2015). Another important aspect is the appearance of the agent. Since large language models can sound human-like in conversation, it is critical to project that aspect to the agent's appearance to maintain the naturalness in conversation and create believable agents. However, it is necessary to avoid the Uncanny Valley (Mori et al., 2012) effect with human-likeness, which refers to the theory that the likeability will increase with anthropomorphism (human-likeness of the agent) until a point where there is imperfect resemblance (a mismatch of appearance or capabilities), which will cause a sudden drop until the agent looks (and acts) exactly like a human for it to rise again. Furhat robot (Furhat Robotics)[1] (Al Moubayed et al., 2012) satisfies these conditions, especially over repeated interactions (Paetzel et al., 2020), and is perceived to be more human-like than other common robot platforms such as NAO and Pepper (Aldebaran Robotics)[8] (Phillips et al., 2018). Thus, the Furhat robot was chosen as the companion robot in this work, as shown in Fig. 1.

Furhat is a social robot, with a back-projected face, which allows displaying a range of facial expressions (e.g., smiles, frowns) and movements (e.g., blinks, eyebrow raises, gaze), lip syncing to speech, and performing head movements (e.g., nods, head shakes). It has two in-built microphones and dual speakers. It offers the possibility to use Google Cloud Speech-to-Text[9] or Microsoft Azure Speech-to-Text[10] engines for speech recognition. It has a 1080p RGB 120°diagonal field-of-view camera. It has face detection and tracking that allows both gaze and head orientation to follow the user.

3.1.1 Face

In this work, FaceCore engine and the latest SDK (2.4.0 for the first study, 2.5.0 for the second study) in Kotlin are used. A neutral-looking face is chosen ('Alex').

The face engine of the robot creates random smiles and eyebrow raises while talking and listening to the user. This behavior is intended to improve the naturalness of the conversation and give a non-verbal backchannelling to the user. However, this happens without context, since speech is not analyzed until after the user stops speaking. Hence, it may be inappropriate in some contexts. In addition, to improve the naturalness of the interaction, the robot blinks, shifts its eyes, and looks away (gaze aversion) briefly while talking based on the silence in the input speech to avoid staring at the user, randomly alternating between looking down or up to the left or right. Face detection and tracking (default engines) are used to give the illusion of agency and awareness.

3.1.2 Speech

English was used as the language for speech recognition and synthesis, since large language models have more training data in English and, thus, are expected to be more capable of conversing in English. It is worth noting that Swedish (young and older) adults generally exhibit a very high level of proficiency in English, with Sweden ranking in 7th place globally for English proficiency in non-native English speakers.[11] The 'Matthew-Neural' voice in Amazon Polly within Furhat was used for text-to-speech (TTS), as a natural-sounding male voice.

A USB microphone array (Seeed Studio)[12] is used to obtain higher-quality audio data for speech recognition, as

[7] Excerpt showing the interaction flow with the robot: https://youtu.be/rkuoOfFuvRU

[8] https://www.aldebaran.com/en.

[9] https://cloud.google.com/speech-to-text/.

[10] https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text/.

[11] https://www.ef.com/wwen/epi/regions/europe/sweden/.

[12] https://wiki.seeedstudio.com/ReSpeaker-USB-Mic-Array/.

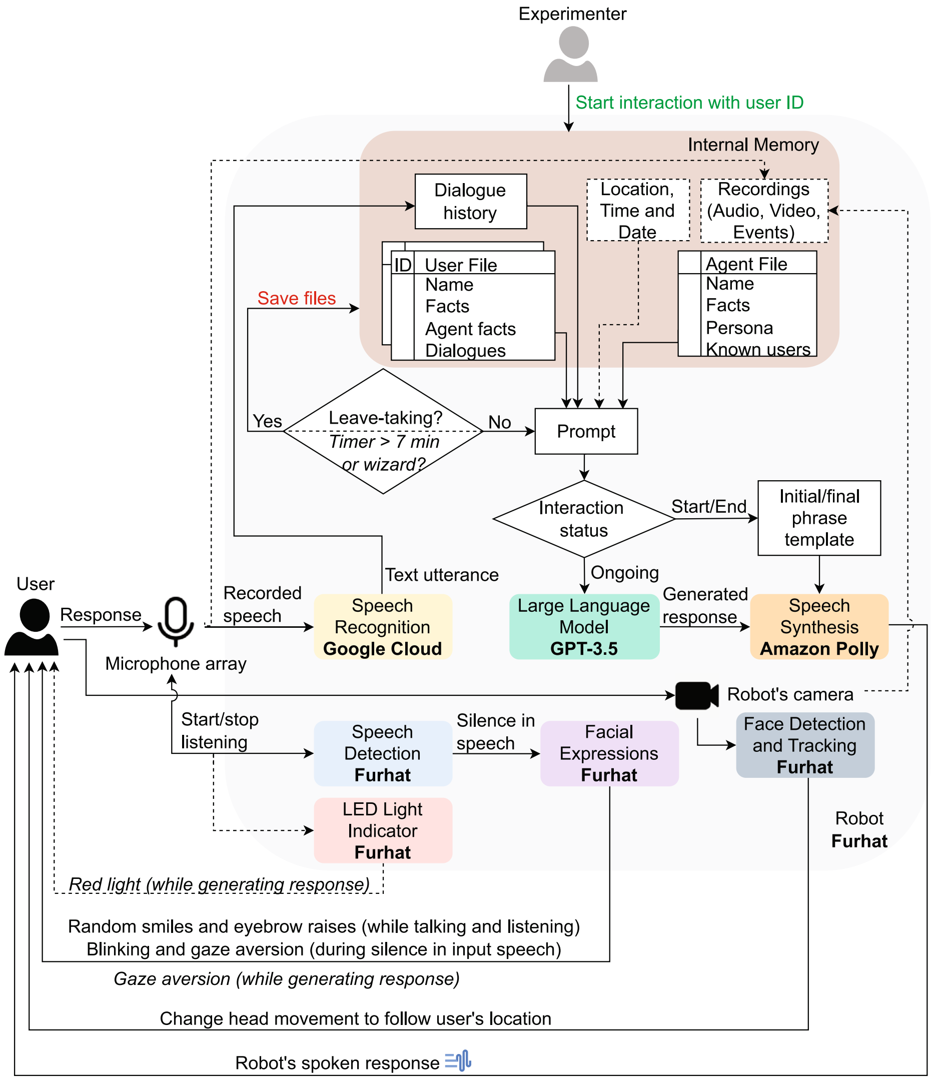

Fig. 2 Interaction flow (Sect. 3.3) diagram for the personalized companion robot. Dashed lines and italic text denote the added features in the technical improvements after the first study (Sect. 4.5). Interaction starts by the experimenter (in green) and ends after the files are saved (red) (Color figure online)

the robot's fans can interfere with the audio recording. Google Cloud Speech-to-Text is used for speech recognition from the audio obtained from the microphone during 'listening', which is determined by the following parameters of the Furhat robot:

1. **No speech timeout**: the duration of silence before the robot deems that the user did not respond. This is increased to 8 s (as suggested in the documentation[13]), instead of the default 5 s. The robot restarts listening after this period has passed.
2. **Silence timeout**: the maximum duration after the user stops speaking before speech recognition is triggered. This was set to 1.2 s. The default value in Furhat is 0.8 s, however, this was increased empirically to account for the pauses in older adult speech.
3. **Maximum speech timeout**: the maximum length of the user's utterance before an interruption from the robot. This is increased to 30 s (the default is 15 s, but 30 s is suggested in the documentation) to prevent frequently interrupting the user when speaking, as older adults may not be accustomed to interacting with spoken dialogue systems that typically require short and clear sentences.

3.2 Large language model

The development of the personalized companion robot started in July 2022, prior to the release of ChatGPT[14] (OpenAI) and the surge of the new LLMs that followed in 2023 (see Zhao et al. (2023) for the most up-to-date survey on LLMs). When the design choices were made, the only publicly available models suitable for open-domain dialogue

[13] https://docs.furhat.io/listening/#listening

[14] ChatGPT was released on November 30, 2022: https://openai.com/blog/chatgpt.

were GPT−3.5 Mar 15 (OpenAI, text-davinci-002 model) (Brown et al. 2020), BLOOM (BigScience Workshop 2023), and BlenderBot 2.0 (Meta) (Komeili et al. 2021; Xu et al. 2022). BlenderBot and BLOOM both require high computing power and graphics for achieving fast inference, thus, taking away from the portability of the robot (e.g., for using it in elderly houses or senior care centers). Hence, GPT−3.5 was used as the LLM, with text-davinci-002[15] for the initial study[16] and text-davinci-003[17] for the second study.[18] Both davinci models have training data up to June 2021, and allow up to 4097 tokens in a prompt. The hyperparameters of the models used in the study are provided in Table 4 in the Appendix. GPT-4 became publicly available after both studies.[19]

While ChatGPT (or gpt−3.5-turbo models[20] in the API) is faster in inference than davinci models and includes slightly more recent training data (up to September 2021), the instructional safety filters (guardrails) from OpenAI to decrease the anthropomorphism of the model resulted in responses typically starting with "As an artificial intelligence, I do not have preferences/I do not have feelings", which would take away from the desired naturalness of the interaction with a companion robot. Note that our intention is not to deceive the users that the robot is a human. The prompt used clearly states that it is a robot (in Sect. 3.2.1), and responses are generated accordingly. However, anthropomorphism and small talk can lead to higher acceptance and trust in the robot (Bickmore & Cassell, 1999; Paradeda et al., 2016; Babel et al., 2021). Hence, text-davinci-003 was used for the second study, which does not have these safety filters in place. To differentiate gpt−3.5-turbo models from davinci models, this article refers to the former as ChatGPT and the latter as GPT−3.5.

GPT−3.5 integration to Kotlin for the Furhat robot was adapted from the OpenAIChat[21] skill by Furhat Robotics, which uses the OpenAI-Java library.[22]

3.2.1 Agent model

The persona of the personalized companion robot (Prompt 1) was based on the work by Zhang et al. (2018) and the follow-up work by Xu et al. (2022). In Zhang et al. (2018), the crowdworkers were asked to chitchat with another worker for a single session based on given personas by instructing them to "chat with the other person naturally and try to get to know each other", and "both ask questions and answer questions of your chat partner". In the follow-up work (Xu et al., 2022), the focus was on personalizing dialogues in multiple sessions, which was achieved by giving crowdworkers the personas and the interaction from the first session from the former work and instructing them to "chitchat with another worker for 6 turns, as if you were catching up since last time you two spoke." and "when you expand the topic, make sure it makes sense with the personal details already mentioned.". We adapted these instructions for a personalized companion robot. Since a male voice and face were used for the robot, 'he' was used in the persona prompt to avoid a mismatch in generated responses with the robot features. The robot's name was kept as Furhat.

> Furhat is a personalized companion robot. Furhat tries to get to know more about his conversation partner, their interests, and activities. When he expands on a topic of conversation, he uses personal details already mentioned to personalize the conversation.
> **Prompt 1:**GPT-3.5 prompt for agent persona.

Agent's and user's utterances in the interaction are stored in the *DIALOGUE_HISTORY* variable. To develop and maintain a consistent persona of the agent through time with multiple users, the facts that the agent says about itself during the conversation are extracted to be stored for future interactions with users. The facts learned from each conversation are concatenated, and stored in the agent file, in addition to the number of known users and persona of the agent (Prompt 1). The facts learned at the end of an interaction for the agent and the user were extracted from *DIALOGUE_HISTORY* using a prompt: Summarize what we know about *NAME*.

3.2.2 User model

In order to create a personalized companion robot that remembers its previous conversations with users to personalize subsequent interactions, the information about the users obtained from the conversations is extracted through GPT−3.5 and stored as JSON files on the robot. Each user is assigned a manual ID (by the experimenter), which is used as their identifier (and in the file name) to start the interaction manually. An index file stores all known user IDs for easier access to the files.

The name of the user is extracted at the end of the interaction from the dialogue history with a prompt. Per interaction, the extracted information from the dialogue (i.e., the user's name and facts learned about the user and the agent), the date and time of the interaction, the dialogue duration, the num-

[15] text-davinci-002 was released on March 15, 2022.

[16] Our initial study was conducted between September 27 and October 11, 2022.

[17] text-davinci-003 was released on November 28, 2022.

[18] Our second study was conducted on March 6 and March 8, 2023.

[19] GPT-4 became publicly available on July 6, 2023.

[20] gpt−3.5-turbo models were released on March 7, 2023.

[21] https://github.com/FurhatRobotics/example-skills/tree/master/OpenAIChat.

[22] https://github.com/TheoKanning/openai-java

ber of turns, and gender and emotions detected by Furhat (for performance analysis) are stored in the user file.

3.2.3 Response generation

In order to ensure a consistent interaction with the robot between users, the initial and final phrases of the robot were manually written (instead of generated). However, the rest of the agent utterances were generated by GPT−3.5. If it is the first interaction with the user, the robot says "Hello! I am Furhat, the personalized companion robot. What is your name?" to obtain the name of the user. If the user is known (i.e., has been interacted with before, thus has a user model file), the robot says "Hi *USER_NAME*! Nice to see you again. What have you been up to since the last time?" to 'catch up' with the activities the user has done in between the interactions with the robot.

The prompt to generate the robot utterances using GPT−3.5 was correspondingly:

> **Prompt 1**
> Furhat is *AGENT_FACTS*
> Furhat knows that *USER_NAME* is *USER_FACTS*
> *USER_NAME* knows that Furhat is *AGENT_FACTS*
> The following is a/second/third/.. conversation between the person/*USER_NAME* and Furhat.
> *DIALOGUE_HISTORY*
> Furhat:
> **Prompt 2:**GPT-3.5 prompt for response generation.

The *AGENT_FACTS* is used if the robot had at least one interaction with a user. Prompts with *USER_FACTS* and *USER_NAME* are used if the user had at least one interaction with the robot.

If GPT−3.5 generates an empty response or a service failure occurs, a clarification request is made to obtain the participant's response again, by choosing a random phrase from: "I am sorry. I didn't understand you.", "Could you repeat that please?", "Not sure if I understood you.", "Could you rephrase that please?", "Sorry, I didn't hear you clearly.".

In order to ensure that the robot expresses leave-taking at the end of the interaction (triggered when the user expresses leave-taking) instead of trying to continue the conversation, a random response was chosen from the list of leave-taking expressions (i.e., "Bye! Hope to see you again soon.", "See you soon!", "Take care until next time.", "Looking forward to seeing you again soon!") as the final response of the robot.

3.3 Interaction flow

The interaction manually starts (by the experimenter) by entering the user ID. The user and agent files are parsed to activate the current user and agent models and set the prompts. When the user is in the engagement zone, the robot says the greeting (either generic or personalized), and it starts listening to the user. After receiving the response from the user, the audio is transcribed by speech recognition. Afterward, the prompts and the dialogue history (including the newly transcribed utterance) are sent to GPT−3.5, as described above. The generated response is said by the robot, after which the robot starts listening to the user again. The robot does not listen to the user while speaking or generating a response. The interaction ends when the user expresses leave-taking (e.g., "Goodbye!", "See you later."), and the robot responds with a random response from its list of leave-taking expressions, given in Sect. 3.2.3. The user and agent models are saved to files, and the robot goes to the idle stage. Interaction flow diagram is shown in Fig. 2, depicting the initial system outline and the changes made after the first study (Sect. 4.5).

4 Preliminary interviews

We investigated the challenges of applying LLMs to conversational companion robots for open-domain dialogue with older adults in two separate phases. In the first phase, preliminary interviews were conducted with 6 older adults, lasting an hour each, based on their 10–15 min of open-domain dialogue with the robot. The second phase consisted of participatory design workshops (Sect. 5) based on the developments made (Sect. 4.5) to address the challenges observed in this study (Sect. 4.4), incorporating the feedback from these preliminary interviews. The interviews took place in September and October 2022 at KTH Digital Futures premises. Each participant was interviewed and interacted with the robot individually. The interactions with the robot were in English, but the interviews were made in Swedish.

4.1 Procedure

Each participant was first asked about their expectations towards companion robots based on the scales from the literature (Heerink et al., 2010; Graaf et al., 2019), as provided in Table 7 in the Appendix. In addition, their prior experience with robots and their living conditions (i.e., alone, with a partner or family member, or in senior housing) were obtained. This was followed by a 5-minute demonstration of the robot's capabilities for autonomous dialogue by the experimenter.

After the demonstration, the participants were instructed that they can **talk about anything they want** with the robot, that the robot would start the interaction, and they can end the conversation **whenever they want**, by saying "Goodbye". The participants were also told that the robot would not hear them when it is speaking. The overall interaction with the robot lasted 4 to 13 min ($M = 9.24$, $SD = 3.45$). The experimenters were present throughout the robot interactions, but did not interfere with the participant's interaction with the robot, unless the participant asked for help. The interaction dynamic with the robot can be seen in the video excerpt from a participant's conversation.[23]

After the interaction with the robot, a semi-structured interview was conducted with the participants, in which they were asked about their experiences of having a social conversation with the robot, adapted from several constructs in the common questionnaires in HRI (Heerink et al., 2010; Weiss et al., 2009; Nomura et al., 2006; Graaf et al., 2019). A range of open-ended questions was covered, focusing on the comprehension, perceived usefulness, ease of use, enjoyment, trust, sociability, social presence, and adaptiveness of the robot, combined with questions about ethical concerns of using robots for social and emotional support, personalization aspects for future applications, and how their expectations compared to their interaction. The questions are provided in Table 8 in the Appendix.

All individual interactions with the robot were video-recorded through an external camera, and interviews with the participants were audio-recorded. All participants gave informed consent for the study, which included options for consenting to anonymized (i.e., blurred face and without full name released) image and/or video sharing in publications.

4.2 Participants

For the preliminary interviews, 6 Swedish-speaking older adults (3 female, 3 male) aged 65 and over were recruited. Participant demographics are provided in detail in Table 1. The invitation to the study was distributed on social media, through KTH communication channels, and sending an email invitation to a group of older adults who had previously attended another experiment at KTH with a social robot (Kuoppamäki et al., 2021). Consequently, 2 of the 6 participants had prior experience interacting with another robot from that study, but 4 had no prior experience. It is important to note that those two participants may have had higher expectations towards the robot than those without prior experience. The participants in the study were able to converse in English fluently, with minor grammatical errors in only two of the participants. The participants were not offered any compensation.

4.3 Data analysis

A total of 55 min of video data for the participants' interactions with the robot were recorded. The pre- and post-interaction recorded audio interviews were in total 3 h and 9 min long. Both the audio and video data were transcribed. The data were qualitatively analyzed, using a combination of conversation analysis (Sidnell & Stivers, 2012) to detect disruptions in the conversation due to poor task performance and content analysis (Krippendorff, 2019) for categorizing the topics in the conversation with the robot and the feedback in the interviews. The outcomes of the data analysis that combine the findings from the robot interactions with the interviews are reported as technical challenges in Sect. 4.4, which formed the basis of the technical improvements described in Sect. 4.5.

4.3.1 Transcriptions

Whisper[24] (OpenAI) Radford et al. (2022) was used on the videos from external cameras for English transcriptions. Large-v2 model (with default parameters) was used, since it achieves the best overall performance for English (Radford et al., 2022). The transcriptions and their timings were corrected manually.

Interviews were manually transcribed using audio recordings by a native Swedish speaker to obtain more accurate responses in a multi-speaker setting (three experimenters and the participant).

4.3.2 Qualitative analysis

The video and audio data were manually coded. Deductive coding (i.e., coding the data based on a predefined set of common failures in HRI (Honig & Oron-Gilad, 2018)) was applied to detect the disruptions in the dialogue, based on conversation analysis to identify turn-taking errors, speech recognition failures, speech detection errors, malfunctioning, and experimenter interference. Inductive coding (i.e., coding derived from the data) was applied to identify additional causes for dialogue disruptions. The resulting dialogue disruption codes for the analysis are provided in Table 5 in the Appendix.

Content analysis was used to derive conversation topics from the robot interactions and categorize feedback in the interviews of the first study. A hybrid coding strategy was taken to use frequency coding (i.e., counting the number of times a code occurs) for dialogue disruptions, descriptive coding (i.e., single word/phrase coding) for identifying topics discussed and detecting topic initiator/closer (participant/robot) in the dialogues, process coding (i.e., noting actions) for out of ordinary responses and reactions in the context of the conversation, and in vivo coding (i.e., quoting the participants) for highlighting the participants' opinions in the interviews or utterances in the dialogue in particular cases. To evaluate the open-domain nature of the conversations with the robot, the identified topics were categorized

[23] Excerpt showing the interaction flow with the robot: https://youtu.be/rkuoOfFuvRU

[24] https://github.com/openai/whisper.

Table 1 Participant demographics in the studies

	Study 1	Study 2
Participants	6	28
Gender	3 female, 3 male	15 female, 13 male
Age	78.3 (8.3)	74.5 (5.6)
Age range	66–86	66–86
Household type		
Living alone	2	5
Living with a partner	4	22
Living in a senior care house	–	–
Other	–	1
Prior experience with robots		
Prior experience with robot(s)	2	5
Talked to robot(s)	2	1
Owns robot(s)	–	3
No prior experience	4	23

in the speech events using the taxonomy by Goldsmith and Baxter (1996) and analyzed under the categories defined by Doğruöz & Skantze (2021) (see Doğruöz & Skantze (2021) for a detailed description of the speech events and categories). The resulting topic codes are provided in Table 6 in the Appendix. Additionally, *anthropomorphism* (i.e., participant attributing human aspects to a robot, e.g., "Are you healthy?") and *mirroring* (i.e., robot mirrors the user's preferences and likes as its own, e.g., "I like... too") are coded in the interactions.

In this article, the feedback from the preliminary interviews is only used for supporting the conclusions from the coded interactions, anecdotally (i.e., using comments from participants) or thematically (i.e., based on recurrent themes in the interviews using thematic analysis), in regards to the technical challenges of applying large language models to conversational companion robots in open-domain dialogue. The detailed analysis of the interviews is outside the scope of this article.

4.4 Technical challenges

Based on the analysis of the robot interactions, four main categories of challenges of integrating LLMs into conversational robots were identified that caused disruptions in the interaction and negative user experiences: frequent interruptions and slow responses, repetitive responses, superficial conversations, and language barrier.

4.4.1 Turn-taking: frequent interruptions and slow responses

When the robot interrupts the user or the person speaks over the robot, the recorded user input can be incomplete or incorrect (e.g., interruption during double negatives or long sentences), which, in turn, affects the LLM performance and the user experience. During the interactions, the participants were interrupted quite frequently by the robot ($M = 14.00$, $SD = 6.26$), as can be observed in the video,[25] which resulted in negative reactions from the user, such as frowns, irritated responses, foreign language anxiety, and the need to rush in speaking. There were three main reasons behind the interruptions that impaired the LLM performance:

1. The participants were not aware when the robot stopped listening, since there were no indicators for turn-taking in the robot. There were pauses in the participants' utterances that were longer than the 'silence timeout', which triggered the robot to start generating a response, while the person continued speaking, with the reasons being (a) the long thought process within speech for an unexpected question with spread-out lexical fillers (e.g., "My favorite movie is... hmm..."), (b) speaking in a foreign language (e.g., trying to come up with the correct translation), or (c) pauses and hesitancy in older adults' speech (Benjamin, 1997):

   - Robot: What kind of books do you like?
   - P1: I... am... *(robot stops listening)* literature, literature I like, *(robot starts talking)* fiction, fiction.
   - Robot: I'm sorry, I don't understand what you mean by locks. Could you explain it to me? *(referring to an earlier part of the dialogue –shown in Sect. 4.4.4– due to lack of a meaningful recorded response)*

[25] Frequent interruptions in conversation by the robot: https://youtu.be/2H5ufXVXKl4.

- P4: And... you asked what I like? Hmm... *(robot stops listening)* I think, well *(robot starts talking)*, sports.
- Robot: Yes, I am just trying to get to know you better.
- P4: *(backchannelling)* Yeah.
- Robot: What do you do for fun?
- P4: I think... back to nature. *(referring to a previous part of the dialogue, not mentioning sports)*

2 The participants were not aware that the robot was generating a response. Due to the long response generation time (2–3 s) from GPT−3.5, there were long silences in robot speech, which made them unsure of whether to continue speaking and urged them to maintain the flow of conversation by continuing (Bernstein, 1962; McLaughlin & Cody, 1982). Correspondingly, their response was not recorded, and the LLM response was out of sync with the user and caused the user to repeat their responses:

- P4: So I have been working with communicating with sign language through video communication. *(waits for the robot's response for 3.5 seconds)* And sometimes with *(robot starts speaking)* written.
- Robot: That is really interesting. Can you tell me more about it?
- P4: Yeah, you need to be able to combine, sometimes use vision, sometimes use written communication. For example, so you... *(robot starts speaking)*

Two participants commented in the interviews that "Response time was too long, so it felt strange" (P4, male). This resulted in participants experiencing the need to learn how to communicate with the robot in a different manner in comparison to human-human conversations:

> You have to learn it (response time). It takes time, you have to learn, you have to wait more. It needs to have a relatively short response time. (P3, male)

> Response time was long, so it felt strange. There was also a lot of rolling back and forth with the eyes and thoughtfulness before he dared to say anything, so you didn't really feel the response. (P4, male)

3 The participants gave longer responses than the 'maximum speech timeout', either due to word repetition or the slow speaking rate of older adults (Benjamin, 1997), which resulted in missing parts of the speech, thus, leading the LLM to generate incorrect, irrelevant, or empty responses. One of the participants noted that he might have said too long phrases when interacting with the robot:

> I probably said too long phrases. So that's why he didn't get what I meant. It should be improved, that he calms down so that I say long sentences, it takes a while before I get to any interaction. That he could wait it out and try to understand what is the question or the source of my long sentences. (P4, male)

In rare occasions ($M = 1.83$, $SD = 1.94$), overlaps happened when the participants talked over the robot (1) by backchannelling (e.g., "hmm hmm", "yes"), (2) when they tried to finish or repair their utterance after the robot interrupted them, or (3) when they started speaking again while the robot was generating a response. In the third case, the participants mostly stopped speaking when the robot started speaking.

The errors in turn-taking also resulted in speech detection errors and speech recognition failures since only part of the speech was processed, thus resulting in incorrect or out-of-context generated responses from GPT−3.5.

4.4.2 Repetitive responses

There were a lot of repetitive utterances from the robot that were either due to (1) an empty response or a server connection problem with GPT−3.5 ($M = 9.50$, $SD = 6.50$) up to 18 times in an interaction (for P6, female), which triggered clarification requests, as described in Sect. 3.2.3 and evident in the video excerpt[26] or (2) the same exact response being generated from GPT−3.5 ($M = 12.00$, $SD = 17.92$) which occurred up to 47 times in an interaction (*P5*), such as repeating a previous response consecutively (e.g., "I love nature too" in the video[27]), asking the user to talk more about the topic (e.g., "That's really interesting. Can you tell me more about it?"), or thanking the user (e.g., "Thanks for sharing that with me."). It is important to note that some of the empty response errors were due to speech recognition failures ($M = 4.67$, $SD = 3.78$) and interruptions ($M = 2.00$, $SD = 3.52$). However, these repeated responses were not due to a bug in the code, it was all generated by GPT−3.5.

In one of the interactions (with *P5*), the same response ("That is really great. I love swimming in the sea too. What is your favorite thing about swimming in the sea?") was generated by GPT−3.5 consecutively 12 times, after which the experimenter interfered to restart the robot. After restarting, the same set of questions was asked to the participant in the same order and exact phrasing twice (i.e., 6 questions on different topics with the same follow-ups from the robot, and repeated in the same order at the end of 6 questions). The

[26] Repetitive clarification requests due to empty response from GPT−3.5: https://youtu.be/nPknDSj2GBs.

[27] Excerpt where the robot repeats the same phrase consecutively: https://youtu.be/2hirZc0kTEI

participant either rephrased their response or changed them when responding on all occasions, and did not ask for help.

In addition to the repeated phrases, the template of the utterance was quite often the same as well (e.g., "I love to... too", "I like to... too", "That sounds like a..."), often mirroring the user's preferences and likes as its own, rarely expressing a different preference or idea, like a 'stochastic parrot' (Bender et al., 2021).

Participants often perceived these utterances as frustrating, which was expressed through gestures, facial expressions, and responses. Repetitive responses caused participants to change their response or topic just to get the conversation flowing, and led two of the participants to ask for help from the experimenters, with one of them asking to stop the interaction after 4 min (*P2*) and requesting the experimenter to continue talking for them. The same generated responses also led to amusement on one occasion. Three participants noted that the responses were repetitive and should be varied:

> It needs to become more varied. (P3, male)

> Try not to repeat the same, exact same phrase. If it says "Yes, it was interesting" or something like that or too often. People tend to avoid such things instead. Get around the same thing in some other way. (P4, male)

> Maybe it was a bit one-sided. It repeated its views quite a bit. (P5, male)

4.4.3 Superficial conversations

In the majority of cases, participants had an informal conversation with the robot, which is visible in Fig. 3. These informal conversation topics were mainly (50% of topics) focused on hobbies and interests, such as literature, movies, music, sports activities, and outdoor activities. The robot took an active role in the conversations, and initiated most of the topics by asking questions about the participant's favorite book, movie, or hobby. Considering that this was the participants' first interaction with the robot (Sect. 3.2.1), the main purpose of the robot persona was to get to know more about the conversation partner, their interests, and activities (Prompt 1). However, this led 4 out of 6 participants to perceive the conversation with the robot as superficial. As a result, conversation topics and responses received from the robot were considered to be lacking depth and meaning. The robot was perceived to be simple, single-minded, and self-reserved about its own hobbies and interests:

> Since the conversation didn't have any depth, it's hard to know how far (the robot) can go when having an important conversation. (P1, female)

> (The robot) doesn't want to talk about his own book, "I don't have a favorite book", "I don't have a favorite movie". And then the conversation gets weird. (P1, female)

> The robot needs to answer more of different things and questions. (P3, male)

As a conversational partner, participants expected the robot to have opinions on its own or to have mutual interests in order to facilitate a mutual conversation, as noted in a conversation by a participant.[28] Robot merely asking questions from the user without giving any feedback or answers was not considered to be conversational:

> There wasn't much feedback, the robot didn't talk about what jazz music it liked. (P5, male)

> I would be very annoyed to have such an idiot robot at home. I mean if you're going to have an interesting conversation about a Nobel Prize winner in literature, it has to be someone who has read a few thousand books and has something interesting to say. (P6, female)

Few participants mentioned that the robot should have talked more about feelings and emotions, by asking the user polite questions such as "How do you feel today?". When the robot straightforwardly started the interaction about queries considering personal interests, it was perceived as impolite.

> "How do you feel today?", "Are you happy today?", "Did you sleep well?", "Do you have pain anywhere?" Like such things would be very good if it could ask. (P1, female)

In many cases, the robot abruptly changed the topics, without any clear connection between the conversation topics. On one occasion (P5), the robot changed the conversation topic five times with only one or two follow-up questions. This could have led the robot to be perceived as interrogative, rather than a pleasant conversational partner. Nonetheless, 4 out of 6 participants anthropomorphized the robot by asking questions about itself, such as "Are you healthy?" (P1, female) or "Where do you go when you go out in nature?" (P4, male). The tendency to anthropomorphize the robot was pronounced with users framing their questions and responses in a manner reminiscent of human-human conversations, expecting the robot to have its own opinions, memories, needs, and desires in a similar way to humans:

- Robot: I will definitely try to visit Africa one day. I think it will be a very interesting place to see.
- P3: Yes. Because you have the desserts, you have, erm, what you call, rainforests. Erm, you have all

[28] Excerpt for superficial conversation: https://youtu.be/7y-nu0YbzVo.

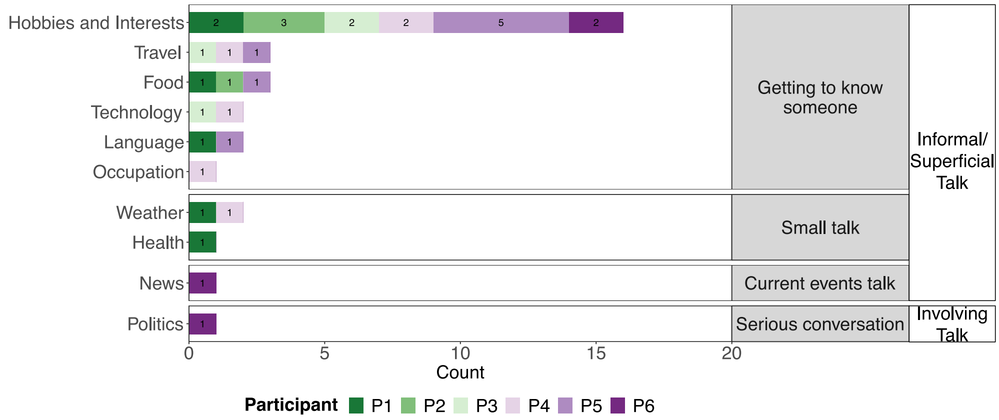

Fig. 3 Topics participants discussed with the robot during the preliminary interviews

> different climates and different things, so it's nice. So I'd really advise or suggest that if you have the opportunity, please do so. But what is your last travel?

- Robot: Well, maybe we have something in common. I also like to read books by Nobel prize winners.
- P6: And what kind of books do you read? Do you have a favorite?

Despite the similarity of the topics discussed, user experiences with the robot varied depending on the interaction disfluencies, the generated text from GPT−3.5 based on the responses of the participants, and the adaptation of the robot persona over time, based on its interaction with users. In addition, the interaction style of the robot (e.g., interrogative, interested, helpful) changed between users, which caused inconsistencies in interactions between users. The robot asked questions based on the topics discussed with previous users, based on the learned facts in the robot persona, which might prove to be a security and privacy issue, as 3 participants were concerned about.

> You talk to a robot in secret. It depends on how the robot, that is, how the entire mechanics, how it handles what I give to it. So, it can be a little... If you think that you lived in China, I would be very afraid to say anything. What I like and think or something like that. Just taking an example, there are certainly other countries also which register everyone like this... What you say and what you think and stuff like this. (–) the security, where it lets go, must be hugely high. (P2, female)

> I'm thinking about testing him and asking, do you remember when you talked to (another participant), but I guess he wouldn't say anything about such situations, but the question is, is there any way to check it out. I think I would be pretty sure he wouldn't gossip. (P4, male)

Nonetheless, all participants wanted the robot to learn from their interactions and be personalized, such that it can provide deeper and richer conversations (2 participants), be "more like a person" (P3, male) and "friend" (P1, female), and refer back to the previous conversations to help users remember in old age.

> *(When asked whether the robot should be personalized)* I think so, because then it won't be so stereotypical. Then there will be a little more conversation, we know each other, and that's clear, then you get closer to each other. I'm a bit torn, because it can be a bit like that... you have to remember it's a robot. But I can imagine if you're older and start to forget things (–) that you can then be reminded of different things, what you've talked about before and stuff like that. It can repeat them. (P2, female)

> Because when you and I sit and talk, if we have several different discussions, you will remember the earlier ones and then it becomes a little more like a person then, if Furhat remembers them too. He can say "You said you were going to see this movie" for example, "What did you think of it?". (P3, male)

4.4.4 Language Barrier

Disruptions in the interaction due to speech recognition failures (i.e., misunderstanding the user) were not very common ($M = 3.67$, $SD = 3.27$), and most were due to Swedish names or phrases (e.g., participant's name, a Swedish book or author) in the dialogue. For instance, the robot recognized Swedish speech as "like going to circus" (shown in the video[29]), which caused the user to be confused but amused. Nevertheless, combined with other disfluencies described above, speech recognition failures resulted in negative user experiences for all participants, since they caused the LLM to generate incoherent responses that were not possible to recover from, with users not understanding that the robot misunderstood them or puzzled on how to respond:

- Robot: Do you have any hobbies?
- P1: I ~~read a lot~~ *(speech recognition failure)* see the locks and I also like to pick berries.
- Robot: That sounds like fun! I'm sorry, I don't know what locks are. Could you tell me more about them?
- P1: About ~~berries~~ Buddy's? ~~Berries~~ That they are in the forest.
- Robot: That's very interesting. I'm sorry, I don't know what a lock is. Could you tell me more about it?
- P1: More about ~~what~~ thought?
- Robot: I'm sorry, I don't understand what you mean by thought. Could you explain it to me?
- P1: I don't understand what you want me to explain. Do you want me to explain how to pick berries, or?

Correspondingly, 4 out of 6 participants emphasized the need to have conversations in Swedish with the robot in the interviews. One participant also noted that the robot spoke very fast. This language barrier resulted in participants being uncomfortable in talking with the robot:

> The (English) language inhibits me, so I can't speak the way I normally do, and therefore, I don't know how it would have gone if we had had the conversation in Swedish. Then I might have started talking about feelings. I tried to get him to say what he thought of the weather, but it didn't work, I got nowhere with that. It didn't seem to work. But maybe I didn't ask the right questions. (P1, female)

Similarly, due to the speech recognition failures, a participant had to stop the interaction after 4 min, and asked the researcher to translate their responses to the robot, which lasted an additional 7 min. In the interviews, she stated that she regretted not being able to talk to the robot more, and her ability to speak their second language was inhibited due to lack of practice in daily life (Bonfieni et al., 2019):

> It's a shame that I couldn't talk to it, so I almost never use English in, like, normal cases so that. But it's like another thing, this is big just to come here and talk to it. And then, well.. there will be several different things *(to talk about at home)*. I don't really know how to talk to it then. (P2, female)

The language barrier was also manifested as the robot's inability to generate dialects and different intonations. If the robot pronounced the person's name in a wrong way, the robot was considered as impersonal:

> That's the obvious one, more language of course. Richer understanding of words, (the robot) can't handle (participant name), but has (–) trouble with dialects. I mean, maybe if (the robot) learns Swedish, it will be difficult with Skåne then *(a Swedish dialect)*. It seems that a big investment in language would probably be needed. (P4, male)

However, a particular use case for having the robot speak English with non-English speakers is to practice English with the robot. This came up spontaneously in one of the interactions, as can be seen in the video.[30] Participants also mentioned their interest in learning a foreign language with the help of the robot:

> I would love to do that *(learn English with the robot)*. Just tell me and I'll come and get him, haha. Yes, but it would be great fun actually. I think that's a great thing to use. Because there are many older people who are really into it *(learning a new language)*. (P2, female)

When the robot is used for learning a foreign language, the robot's ability to make the user feel comfortable and allow the user to make mistakes become critical:

> (The robot) was very friendly, and he was very friendly when he kind of said "If you and I speak a little more English, you will learn a little better". So that there was like that, a sort of.. something that would be very good in a social context, because my English wasn't that good. (P1, female)

4.5 Technical improvements

Technical improvements were made to the robot based on the challenges identified. However, to prevent confounds in the evaluation of LLMs in companion robots, no additional library (e.g., for turn-taking, speech recognition, or LLM) was used.

4.5.1 Turn-taking

The 'silence timeout' was increased by 50% (to 1.8 s) to account for the long pauses in speech. Since the response

[29] Speech recognition failure due to participant talking in Swedish: https://youtu.be/rjDipaS0bis

[30] Robot suggests practicing English with the participant: https://youtu.be/sWL8uGK4u0Q.

time of GPT−3.5 is very long, the added duration for silence would create further disruption in interaction ('awkward silence'), especially if the users restart speaking when the robot starts generating a response, hence, this value was not increased substantially.

Deriving from spoken dialogue systems that use LED lights to indicate whether the device is 'listening' (e.g., Amazon Echo), which older adults may be familiar with, and previous studies in HRI (e.g., Perera et al. (2017); McMillan et al. (2019); Senaratna et al. (2020); Pollmann & Ziegler (2021); Maniscalco et al. (2022); Lekova et al. (2023)), the LED underneath the Furhat robot was used to indicate when the robot takes the turn. The LED turned red when the robot stopped 'listening' (i.e., right before the recorded audio is sent to speech recognition), and turned off when it was (after the robot finished saying the utterance generated by the GPT−3.5 response).

Gazing away is a powerful indicator to improve turn-taking in HRI (Skantze et al., 2015), hence, gaze aversion (i.e., looking either top/bottom left/right) during response generation was implemented on the robot to demonstrate that the robot is 'thinking'. The robot returned its gaze to the user when the response was generated to talk and maintain eye contact with the user.

4.5.2 Repetitive responses

In order to decrease the number of empty responses or connection failures due to GPT−3.5, the request was changed to be sent five times. In the case that the response was still not filled, instead of only asking clarification requests, which were frustrating to the participants, backchannelling responses (i.e., "I see!", "Hmm hmm.", "Right.") and invitation for elaboration ("Could you tell me more about that?") were added to push the conversation forward.

'Frequency penalty' was not changed to evaluate whether improvement of the other factors in the robot would overcome the repeated responses.

4.5.3 Superficial conversations

In order to establish deeper conversations with participants, the persona of the robot was changed based on the feedback in the interviews, such as "The robot needs to ask questions about feelings and health", and "The robot needs to talk more about itself". In addition, an anthropologist, who had ongoing ethnographic research for integrating robots in senior care centers, was consulted (e.g., "Senior care staff often talk to senior citizens about their families, memories, and emotions", "The robot needs to be empathetic"). The resulting 'empathetic' persona is given in Prompt 3). Instead of Furhat, a Swedish name (Linda) was used to improve both speech recognition (Furhat is often transcribed as "for hat") and the believability of the robot persona. 'Jane' was used as the Furhat's face, as shown in Fig. 1, based on the preferences of older adults in the participatory design workshops.

> Linda is a personalized empathetic friendly companion robot for older adults. She talks about people's lives, interests, experiences, emotions, relationships with others, and reflects on them. She values people's opinions, recognizes their feelings, and provides social and emotional support. She also talks about her own experiences to reflect on situations as a friend. She is an active listener, and understanding. She asks open questions. She wants to talk about people's memories and family members. She tries to create positive emotions in the person. When she expands on a topic of conversation, she uses personal details already mentioned to personalize the conversation.
> **Prompt 3:**GPT-3.5 prompt for empathetic robot persona.

The location, date, and time of the interaction were added to the prompt to provide more accurate responses. For personalized interactions, the previous date and time of the interaction were also provided. However, learning of persona over time was removed to ensure consistent interaction between users and prevent any privacy issues.

> **Prompt 3 :**
> Time at the start of this conversation is *DATE_TIME*.
> *USER_NAME* and Linda are located in Stockholm, Sweden.
> Last time Linda and *USER_NAME* spoke was *LAST_DATE_TIME*.
> The following is a/second/third conversation between the person/*USER_NAME* and Linda.
> *DIALOGUE_HISTORY*
> Linda:
> **Prompt 4:**Updated GPT-3.5 prompt for response generation.

LLM in the robot was updated to GPT−3.5 text-davinci-003 model, based on its "higher quality writing with clearer, more engaging, and more compelling content".[31]

4.5.4 Language Barrier

While the older adults in the study were able to converse fluently in English, with minor grammatical errors in only two of the participants, turn-taking interruptions and speech recognition failures, followed by several clarification requests, led them to hesitate in their answers and English level, thus,

[31] https://help.openai.com/en/articles/6779149-how-do-text-davinci-002-and-text-davinci-003-differ.

refrain from exploring a wider range of topics or even talking altogether, as noted by the participants in their interviews, with 4 of them highlighting that the robot needs to speak in Swedish. Thus, the prompts were translated to Swedish for GPT−3.5, which in response, generates Swedish utterances. The language for Google Cloud Speech-to-Text was set to Swedish as well. In addition, a Swedish TTS (Amazon Polly, 'Astrid') was used, based on the preferences of older adults in the participatory design workshop. Based on the feedback from one of the participants, the speaking rate was decreased to 80% of the TTS, since older adults speak 20 to 25% slower than young adult speakers (Benjamin, 1997).

4.5.5 Other

With the aim of conducting participatory design workshops with older adults, a wizard interface was created to ensure that experimenters can simultaneously start multiple robots to interact with users without changing any code. To ensure that each participant had sufficient and equivalent time to interact with the robot, without being concerned about when they should end it, a (7-minute) timer was added. When the robot stopped listening, before a response was requested from GPT−3.5, the timer was checked. If the dialogue duration is equal or greater, the robot would end the interaction with a pre-scripted phrase: "I would love to talk more another time, but for the sake of time, I need to say goodbye. Thank you for talking with me. Take care!" (in Swedish). Hence, the overall conversation duration may be slightly longer than the set timer. The robot would then listen to the response from the user, before saving the interaction files. In case the participant wanted to end the interaction prior to the timer, an option was added in the wizard interface to end the interaction with the same phrase to ensure consistency and prevent the robot from continuing the conversation. For malfunctioning of the robot, another option was added in the wizard interface to abort the interaction, which saved the current interaction files, and notified the user about the error in the system: "There seems to be an error in my system. Give me a minute please.". When the interaction is restarted with the same user from the wizard interface, the saved files are used to continue the conversation from where it was left off: "Sorry for the wait! Could you tell me what you said right before?".

A recorder for the robot's camera and microphone, and a logger for events of the robot were added for performance and data analysis.

5 Participatory design workshops

After the technical improvements in the robot's abilities to tackle the challenges faced in the preliminary interviews, we moved on to the second phase in the co-design process of a personalized companion robot, through participatory design workshops. 28 older adults aged 65 and over participated in the workshops. A total of 4 workshops were conducted, which lasted two hours each. They took place on March 6 and March 8, 2023, at KTH Digital Futures premises. The workshops and the robot interactions were conducted in Swedish.

5.1 Procedure

The participatory design workshops consisted of three stages (see Irfan et al. (2024a) for the qualitative analysis of the focus group discussions during the workshops):

1. **Design scenarios:** Focus group discussions (Fig. 4) were conducted using design scenarios from everyday life situations of older adults to understand their expectations towards companion robots. Design scenarios were videos without audio that show an older adult (1) visiting family, (2) feeling lonely, (3) receiving bad and (4) good news, (5) waking up, and (6) having friends over. The participants were asked about their perceptions of the robot, and what kind of conversation they would like to have with the robot in these scenarios. Prior to the discussions, the experimenter talked to the robot for 2 min to demonstrate its capabilities. Focus group discussions lasted an hour.
2. **Robot interactions:** Each participant had an individual interaction with the robot (Fig. 1) in open-domain dialogue for 7 min, followed by a robot acceptability questionnaire provided in Tables 9 and 10.
3. **Interviews:** After the robot interaction, small groups interviews were conducted with 3–4 participants (8 groups) to better understand their user experience and underlying reasons behind their ratings in the questionnaires. The interviews were based on the questions from the preliminary study (Table 8), and lasted 30 to 40 min.

In this article, our emphasis is on the technical aspects of the conversation, aiming to pinpoint the limitations of LLMs and understand how the multi-modal features of the robot impact LLM performance, based on iterative improvements. Hence, the examination of the design scenarios and interviews is omitted in this study, with questionnaire results serving as the basis for supporting our findings from the videos.

Similar to the initial study, each participant talked to the robot in open-domain dialogue ("**talk about anything you want**"). Contrary to the first study, the duration was pre-set for 7 min, which was initiated and monitored by an experimenter through the wizard interface. The experimenter did not interfere with the interaction (i.e., the robot was fully autonomous during the conversation), unless there was a malfunction in the system or if the participant wanted to end the interaction before the 7-minute duration. The interaction ended with the robot saying a pre-scripted phrase, as

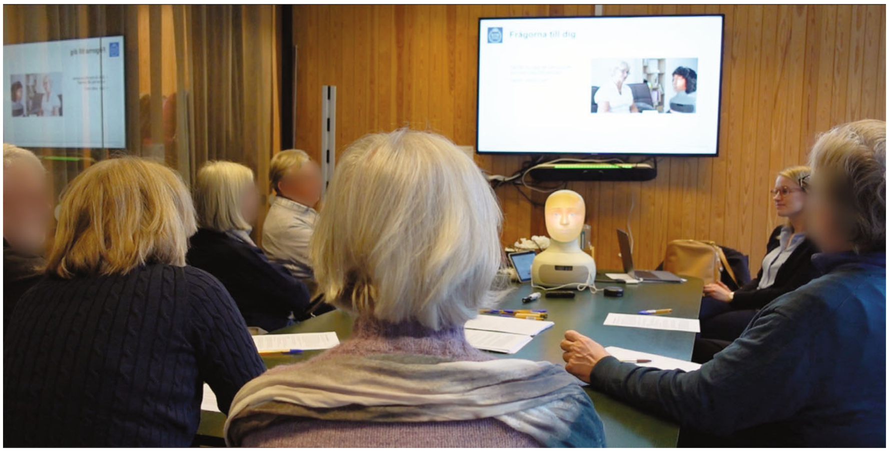

Fig. 4 Participatory design workshop with older adults (6–8 participants per workshop)

described in Sect. 4.5.5. The participants were instructed not to speak when the red light underneath the robot was on, as it meant the robot was either generating a response or talking, and the robot would not hear them if they talked. The gaze aversion was not mentioned to maintain the naturalness of the interaction.

After the interaction, all participants filled out a questionnaire that was adapted from the interview questions from the first study (Table 8 in the Appendix), which was extended with additional adapted questions from commonly conducted questionnaires (Likert scales from 1 to 5) in HRI (Nomura et al., 2006; Lee et al., 2006; Bartneck et al., 2009; Syrdal et al., 2009; Heerink et al., 2010; Weiss et al., 2009; Malhotra et al., 2004; Graaf et al., 2019) and open-domain dialogue (Zhang et al., 2018; Shuster et al., 2022; Borsci et al., 2022). The questionnaire was divided under the factors: (A) Experience with Linda, (B) Satisfaction with Linda, (C) Safety with Linda, (D) Linda's Personality, (E) Prior Experience with Robots, (F) Background (i.e., gender, age, marital status, children/grandchildren, and household type), including a question on whether the participant would like to come to KTH (study venue) to talk to the robot again (Yes/Maybe/No), and additional feedback through an open question. Questions and the corresponding results from the study are provided in detail in Tables 9 and 10 in the Appendix.

All interactions with the robot were video-recorded both through an external camera and the camera on the robot.

5.2 Participants

The participants were recruited by distributing the invitation to KTH communication channels, social media, and platforms for gathering senior citizens. In total, 28 participants from the age group 65 and over registered as volunteers. All participants were Swedish speakers, and 15 of 28 were females. Most participants (23) did not have prior experience with robots. Further participant demographics are provided in Table 1 (Sect. 4.2). The participants from the preliminary interviews and the previous kitchen robot study (Kuoppamäki et al., 2021) were not part of this study. The participants were offered a small compensation (100 SEK gift card) at the end of the study. The distributed study invitation mentioned that a gift card will be given as compensation, but did not specify the amount. All participants gave an informed consent for the study.

5.3 Data analysis

A total of 3 h and 24 min of video data for the participants' interactions with the robot were recorded. The audio recordings for the design scenarios were 3 h and 40 min long, and the post-interaction interviews were 3 h and 52 min long in total. Both the audio and video data were transcribed. The video data were qualitatively analyzed in the same structure as described in Sect. 4.3. In addition, quantitative analysis was conducted for response generation time, speech recognition performance, conversation topic duration, and the questionnaire data. As explained in Sect. 5.1, the questionnaire results are only used to support the conclusions for the technical challenges. The detailed analysis of that data, the design scenarios (analyzed in Irfan et al. (2024a)), and the feedback from the interviews are outside of the scope of this article.

5.3.1 Transcriptions

Whisper (OpenAI) was used on the videos from external cameras for transcriptions of robot interactions. Large-v2 model (with default parameters) was used, since it achieved the best overall performance for Swedish (Radford et al., 2022). Google Translate was used to translate transcripts into English. All videos were manually checked to correct the text and timing of transcriptions, and to remove transcribed text for background voices.

Table 2 Questionnaire results per construct from the participatory design workshop, where Likert scale is from 1 (strongly disagree) to 5 (strongly agree).

Construct	Median	IQR	Consistency (α)
Comprehension	4	1	–
Clarity	4	1	–
Turn-taking	4	1	0.6
Engagingness	3	1.25	–
Consistency	4	1	–
Fluency	3	1	–
Credibility	3.5	1	–
Use of knowledge	3	1	–
Contextual memory	3	1	0.84
Usefulness	3	1	–
Ease of use	4	1	0.73
Enjoyment	4	1	0.84
Emotion(al influence)	1	2	0.53
Sociability	2	2	0.77
Social presence	2	1	–
Personality	2	2	–
Adaptiveness	3	1	0.61
Trust	3	0.5	–
Security	3	0.25	–
Privacy concern	3	3	0.78
Anxiety toward robots	2	2	0.7
Attitude towards technology	4	1.25	–
Intention to use	2	2	–
Anthropomorphism	2	1.25	0.88
Animacy	3	1	0.88
Likeability	4	1	0.86
Intelligence	3	1	0.82

Detailed results are available in Table 9 and 10. Constructs with a single item do not have Cronbach's α (consistency between questions)

5.3.2 Qualitative analysis

The qualitative analysis was conducted for dialogue disruptions and topic detection on the interactions with the robot in the same structure as in Sect. 4.3. This work omits the qualitative analysis of the design workshops and interviews in order to prioritize the examination of the technical challenges associated with LLM integration into conversational robots.

Topics and dialogue disruptions are coded by two coders (expert and junior), which had moderate agreement ($\iota = 0.5$[32]) for topics and low agreement ($\iota = 0.17$) for disruptions. The reason for low agreement could be the high number of disruption codes (14) with multivariate aspects, and the high number of topics (26) with occasional overlaps (e.g., memories about family should be categorized under both 'memory and 'family' topics), which create a high cognitive load on the annotators, thus, possible to miss or misinterpret. The expert annotator defined the topics and disruptions based on qualitative analysis, as described in Sect. 4.3, hence, was more familiar with the data and the corresponding coding categories. Thus, the expert coder's annotations are reported.

5.3.3 Quantitative analysis

All the events that are triggered in the robot during an interaction were logged in milliseconds for accurate data analysis. From these logs, speech recognition duration and GPT−3.5 response time are obtained. Since the full dialogue history is added to the prompt for GPT−3.5 at each turn of the interaction, thus, increasing the prompt length, the response time is expected to increase throughout the conversation. Hence, the means and standard deviations of the *initial* (to generate the first response), *average* (response time on average throughout the conversation), and *final* (to generate the last response) response times are provided, along with the prompt length (in tokens[33]) and speech recognition duration.

Word error rate (WER) is calculated for speech recognition using manually-corrected transcriptions of the videos and the conversation log of the robot (i.e., text obtained from speech recognition). Note that the disruptions in turn-taking have affected the speech recognition performance, that is, if the participant(s) talked over the robot when it was not listening, WER would increase.

Topic duration is annotated using the video interactions for this study. This was not analyzed for the first study due to the several disruptions in turn-taking resulting in the participants repeating what they said multiple times, and sudden topic changes due to speech recognition failures.

Median (Md) and interquartile range (IQR) are reported for the questionnaire results (Tables 9 and 10), because the data is ordinal (Jamieson, 2004). In addition, the percentiles of users who either agreed (rated 4 or 5) or disagreed (rated 1 or 2) with the questions are also provided for further insight into the responses. Cronbach's alpha (α) is measured per construct for internal consistency between items. The results of the questionnaire are used to underpin our conclusions drawn from the videos. The detailed analysis of the questionnaire is beyond the scope of this work.

[32] ι is calculated using iota method in irr package in R based on Janson and Olsson (2001), which is an expansion of Cohen's κ for multi-variate ratings.

[33] Calculated using OpenAI's tiktoken library: https://github.com/openai/tiktoken

Table 3 GPT−3.5 response generation time, prompt length, and speech recognition duration for initial and final dialogue turn, and averaged throughout the dialogue

Category	GPT−3.5 Response time (s)	Prompt length (tokens)	Speech recognition duration[1] (s)
Initial	$M = 2.45$, $SD = 0.55$	$M = 391.75$, $SD = 9.76$	$M = 0$, $SD = 0.0008$
Average	$M = 2.38$, $SD = 0.22$	$M = 1115.32$, $SD = 177.51$	$M = 0$, $SD = 0.0005$
Final	$M = 2.42$, $SD = 0.59$	$M = 1658.92$, $SD = 260.05$	$M = 0$, $SD = 0$

[1] Timestamps of events were recorded with milliseconds. 0 refers to a duration of less than a millisecond

5.4 Technical challenges

In addition to the previously identified categories of challenges in the first study (i.e., turn-taking, repetitive responses, superficial conversations, and language barrier), hallucinations and obsolete information, disengagement cues, and premature closures were identified as challenges of LLMs in conversational companion robots through this study. Note that the duration of interaction with the robot was less in this study ($M = 7.27$, $SD = 1.52$) than in the first study ($M = 9.24$, $SD = 3.45$). Table 2 shows the overall user experience based on the questionnaire in terms of constructs, and Table 9 (in the Appendix) shows the results per question.

5.4.1 Turn-taking

Robot interruptions decreased drastically ($M = 3.18$, $SD = 2.57$) compared to the initial study ($M = 14.00$, $SD = 6.26$), which shows that turn-taking indicators (the red light and gaze aversion) were beneficial. Correspondingly, 81% of the words spoken were transcribed, such that the LLM could respond without a high number of clarification requests, leading to moderately fluent discussions (B.6: $Md = 3.0$, $IQR = 1.0$, 46% agree). However, interruptions still occurred in 24 (out of 28) interactions up to 9 times in an interaction. Nonetheless, the participants perceived to a large extent that the robot did not interrupt them (A.16: $Md = 4.0$, $IQR = 2.0$, 71% agreed), the robot could understand when they wanted to take a turn (A.15: $Md = 4.0$, $IQR = 1.0$, 57%), and it was easy to start and continue the conversation without any help (A.10: $Md = 4.0$, $IQR = 1.0$, 86%). Participants gazed away when they were thinking, even when they had the turn, hence, it was possible to miss the light indicator, which a participant (P6, male) noted after the interactions.

Similar to the initial study, the participants' long responses depending on the context (e.g., talking about daily activities prior to the interaction) resulted in partial responses to be recorded, however, none of these led to disruptions in the interaction, as the LLM was able to generate a response that kept the conversation flow.

While the speech recognition duration was very fast (less than a millisecond), the response time was very long due to slow GPT−3.5 response generation (Table 3). Participants had conflicting perceptions when asked if the robot was slow in its responses (B.9: $Md = 3.0$, $IQR = 2.0$, 46% agreed). While the prompt length increased with each dialogue turn, as the dialogue history is appended to the prompt, no correlation was found ($\tau = 0.04$, $z = 1.47$, $p = 0.14$) between LLM latency and prompt length,[34] which is evident from both Table 3 and the wide variance in Fig. 5. This suggests that the majority of the response time is due to the server response time rather than the LLM generation time. Note that the maximum prompt in an interaction (2243 tokens) was shorter than the maximum prompt length of GPT−3.5 (4097).

5.4.2 Repetitive responses

The empty responses from GPT−3.5 ($M = 0.54$, $SD = 0.74$) were substantially fewer ($M = 9.50$, $SD = 6.50$ in the first study), which happened to 11 participants, at most 2 times in an interaction. Due to the (only) bug in the code, there was no utterance made by the robot in those cases. Hence, instead of making a clarification request or backchannelling, it started listening to the user again (the red light turned on and off). This behavior confused the participants on whether they should wait for the robot to answer or they should say something. When the participants asked for help from the experimenters, they were told the robot did not hear them, and to repeat their response. Otherwise, the participants commented on the robot's silence and the robot responded accordingly (e.g., "I was just trying to find information about the weather." as shown in the excerpt[35] or "I was lost in thought about your beautiful boat. What did you say again?"), which were good repairs from the LLM.

Repetitive phrases from the LLM were also considerably lower ($M = 0.75$, $SD = 2.76$) than the first study ($M = 12.00$, $SD = 17.92$), and only occurred in 3 interactions. But

[34] Since both variables are not normally distributed ($p < 0.01$ in Shapiro test), Kendall's τ is measured.

[35] Empty response from the robot with a recovery phrase: https://youtu.be/mXc_Xa_vkCc.

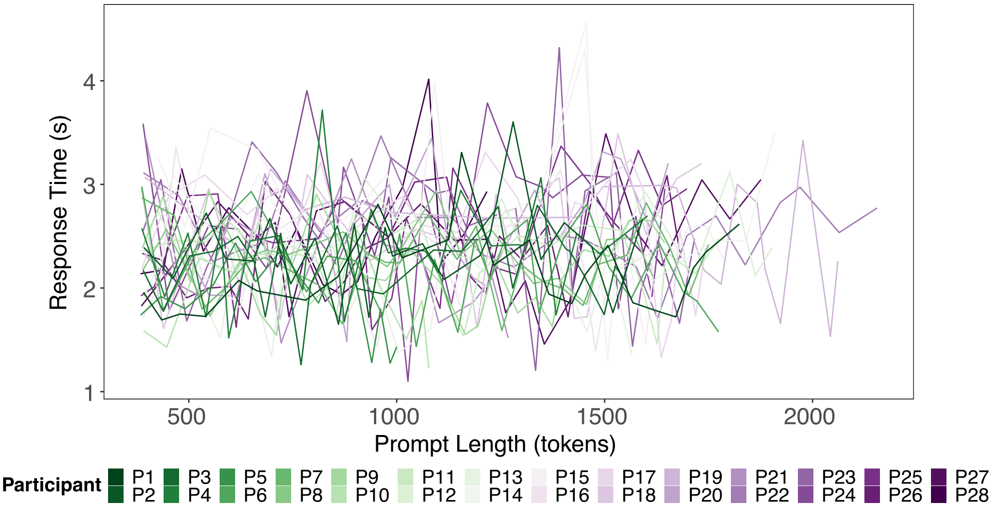

Fig. 5 GPT−3.5 response time variance with prompt length per participant

it occurred up to 14 times in an interaction, in which the robot might have been perceived as interrogative and offensive, since it kept on questioning why the user was thinking in a certain way (about UFOs) with the exact same phrase. The participant turned to the experimenter to ask how they could change the topic, but this triggered a topic change from the robot.

5.4.3 Superficial conversations

The topics discussed with the robot were more diverse than the preliminary interviews, as shown in Fig. 6. Albeit being much less frequent than the initial study, hobbies and interests formed most of the conversations (23.07%), but in itself had a wide range of topics, from science, space, and UFOs to literature, music, exercise, and outdoor activities. The underlying reasons for the higher diversity of the speech events or topics could be due to (1) the modified LLM persona that prioritized social and emotional support over small talk, (2) the lower number of disruptions in the conversation flow, which encouraged users to engage more in the conversation, (3) the design scenarios prior to the interaction in which possible conversation topics were discussed between the participants, and the robot's capabilities were clarified by the researchers whenever requested, and (4) the variability in user preferences and a higher number of participants.

The majority of the conversations (69.87%) were 'informal/superficial', due to it being their first interaction with the robot, as well as the robot trying to instantiate conversations around interests, memories, activities, and feelings due to its persona. However, the participants also explored the task-oriented capabilities of the robot (or rather the LLM) in 21.15% of the conversations. These 'goal-directed talks' consisted of (1) 'information request' about a restaurant, transportation, TV program, and recipes, (2) 'action request', such as the user requesting the robot to book a restaurant, ask it to sing or say something, (3) 'advice seeking' in which the user asks the robot to provide opinions on a subject and discusses with it on that topic (e.g., a recommendation for a rug, gardening, activities for the weekend, travel), and (4) 'plans with robot', in which the user invites the robot over or makes future plans with it. Note that these topics are separate from 'complaining' about the robot's behaviors (e.g., gaze), 'apology' from the person or robot about their action or utterance in the conversation, 'disagreement' with the information provided by the robot, and 'technology' in which the user either asked the robot about its capabilities to get to know it more or discussed broader topics about technology and AI with the robot, rather than requesting information, action, or advice. Only a small part of the conversations (8.97%) went into deeper topics (i.e., 'involving talk').

In comparison to the previous persona, the robot was more submissive, i.e., the participants took the major lead in initiating (70.9% of topics) and closing (61.4%) the topics rather than the robot, which could be due to the 'active listener' part of its new personality prompt or due to the personality of the participants. This is likely why, less than half of the participants perceived the interaction as interesting (B.4: $Md = 3.0$, $IQR = 1.25$, 46% agree). 4 participants were also worried about what to say or talk about with the robot (C.8: $Md = 2$, $IQR = 2$). The user experience was also more varied due to the wide range of topics covered, thus, the user satisfaction with the conversation varied (B.8: $Md = 3.0$, $IQR = 1.25$, 21% were not satisfied). Consequently, the participants had mixed opinions about having the robot at home (A.7: $Md = 2.0$, $IQR = 2.0$), with the majority (57%) not considering it ready yet.

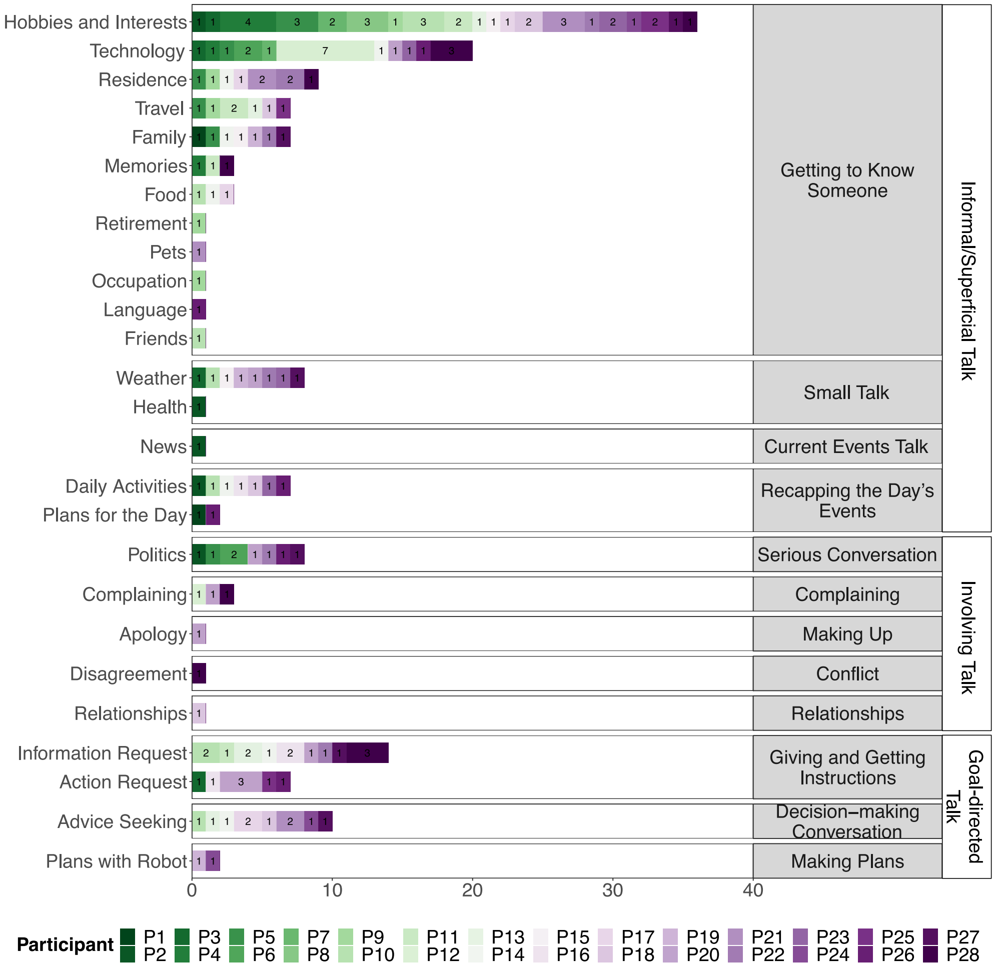

Fig. 6 Topics participants discussed with the robot during the participatory design workshops

While in some cases (in 10% of the interactions), the LLM mirrored the user's responses similar to the initial study, in most cases, it expressed opinions and offered suggestions, which might have improved its agency and human-likeness for the participants. 9 participants anthropomorphized the robot in their interactions (e.g., "Do you like modern music?", "Have you watched TV?"). However, Godspeed showed low anthropomorphism ($Md = 2.0$, $IQR = 1.25$, Cronbach's $alpha = 0.885$) the robot did not "feel like a real person" for the majority of (79%) participants (A.17: $Md = 2.0$, $IQR = 1.0$) or have a personality (B.10: $Md = 2.0$, $IQR = 2.0$, 57% disagree). This could be due to the lack of a 'background' or memories, leading to shallow responses. Nonetheless, most participants (71%) considered the robot's responses as consistent (B.3: $Md = 4.0$, $IQR = 1.0$). However, the contextual memory (i.e., remembering prior user responses and responding accordingly) of the LLM was not perceived highly ($Md = 3.0$, $IQR = 1.0$, $\alpha = 0.84$), showing the limitations of LLMs for memory in long conversations. In fact, the majority (71%) of the participants wanted the robot to remember their conversations (C.9: $Md = 2.0$, $IQR = 2.0$), but they were also worried about privacy and data collection with the robot (C.2: $Md = 4.0$, $IQR = 2.25$, 54% agree).

5.4.4 Language Barrier

WER ($M = 0.348$, $SD = 0.099$) for speech recognition (of Google Cloud Speech-to-Text) was lower than previous findings of the architecture on Swedish conversations (0.412) by Cumbal et al. (2021), which indicates that the robot's architecture or the microphone array did not affect the performance of speech recognition. Speech recognition failures (i.e., misunderstanding user response) were also less frequent ($M = 0.22$, $SD = 0.64$) than in the initial study ($M = 3.67$, $SD = 3.27$), occurring in 4 interactions (up to 3 times per interaction), often due to Swedish names ($M = 0.21$, $SD = 0.69$). Hence, the interaction was smoother, as the LLM responses were coherent, showing the importance of language choice for speech recognition and

LLM. Nonetheless, when occurred, these failures resulted in either an out-of-context response from the LLM, which led to confusion and topic change, or triggered a correction by the participant. 61% of the participants agreed that the robot understood them (A.14: $Md = 4.0$, $IQR = 1.0$), while 57% of the participants worried that the robot wouldn't understand or hear them (C.6: $Md = 2$, $IQR = 1.25$).

LLM responses are also affected by the speech generation of the robot, since LLM assumes that the task is done according to what is said. For instance, as shown in this video,[36] when the participant requested the robot to sing, the robot read the lyrics of an English song with a Swedish accent to which the participant responded with "You didn't sing, you talked". However, LLM had no context of what the participant referred to, hence, the robot replied "I sang the song, but if you want to talk about it, I am happy to do that". In addition, while LLMs can generate multi-lingual outputs based on prompts, the fact that only a single language (Swedish) was used as the input language in speech recognition led to a failure when one of the participants asked the robot whether it could speak German in German.

5.4.5 Hallucinations and obsolete information

Due to 'goal-directed talk', which did not take place in the initial study, 'hallucinations' (e.g., misinformation generated by the LLM about a restaurant, bus schedule, TV series, opera, and exaggerating robot capabilities) and obsolete information (e.g., saying it is sunny weather during snowstorm,[37] or an old movie recommendation) came into sight ($M = 0.57$, $SD = 0.84$) in 11 of the interactions, occurring up to 3 times. This led to neutral trust (61% neither agreed, nor disagreed) in the robot (A.11: $Md = 3.0$, $IQR = 0.25$), caused varied perceptions of the correctness of the robot's responses (B.17: $Md = 3.5$, $IQR = 1.0$, 50% thought that the robot answered correctly) and the competency of the robot (B.5: $Md = 3.0$, $IQR = 1.0$, 46% did not find the conversation insightful). But it also led one participant to doubt themselves about the information they knew (about the production date of a TV series).[38]

5.4.6 Disengagement cues and premature closures

Disengagement cues generated by the LLM, such as "I understand" and "That is good to know", brought the conversation to a halt, with participants not knowing how to respond in 5 of the interactions, up to 4 times in an interaction ($M = 0.39$, $SD = 0.99$), as can be seen in this video excerpt.[39] Moreover, the robot tried to end the conversation prematurely (earlier than 7 min) in 7 interactions, up to 2 times ($M = 0.32$, $SD = 0.61$). The participant either changed the topic to continue the interaction without intervention, or the experimenter told them they could continue if they wanted to. This behavior might have led to mostly positive (46%), but mixed opinions for the engagingness of the conversation (B.2: $Md = 3.0$, $IQR = 1.25$).

Topics were less frequently changed in comparison to the initial study, with several levels of follow-up questions in most interactions, with the participant changing the topic 61.4% of the time. On average, topics were discussed for less than a minute ($M = 50.1$ seconds, $SD = 39.3$), depending on the choice of topic, as shown in Fig. 7. Politics was discussed for the longest period ($M = 82.9$, $SD = 44.5$) as expected, since it is involving talk with several follow-up questions from the robot and the participant. While the conversations around each hobby or interest (separately timed) were also discussed for a long time ($M = 70.1$, $SD = 51.5$) with follow-up questions, some of them lasted as short as a single turn or led to other topics, such as family. The longest conversation (212 s) was on an action request, where the participant (P25, male) requested the robot to tell him a story, with follow-up questions. The topic was changed when a speech recognition failure occurred (without a turn-taking error):

- P25: I want you to tell me a story, one folk tale if I say "Once upon a time".
- Robot: Certainly. One of my favorite fairy tales is the one about a magical dragon that lives in the forest.
- P25: Continue.
- Robot: The dragon had magical powers that gave him the ability to talk to animals and trees.
- P25: A little more.
- Robot: The animals and trees taught the dragon about secrets of life and important lessons.
- P25: How did the dragon talk to the tree?
- Robot: The dragon could talk to the tree by cutting holes in it.
- P25: Aha. Exciting.
- Robot: Yes. He felt the tree's energy and warm vibration causing it to open up to him.
- P25: ~~How did the dragon talk to the animals? (Hur pratade draken med djuren?)~~ Have you talked to the dragon the great? (Hardu prata då draken med jättebra?)

[36] Robot reads the lyrics of an English song with a Swedish accent instead of singing: https://youtu.be/Nv_jadbnAcI.

[37] Incorrect weather prediction (obsolete information) by the robot: https://youtu.be/GCtDqLcrco0.

[38] Incorrect information about the TV series (hallucination) confused the user: https://youtu.be/8sHknc1mbhM

[39] Disengagement cue leaves the participant puzzled: https://youtu.be/xOzwxdiuA94.

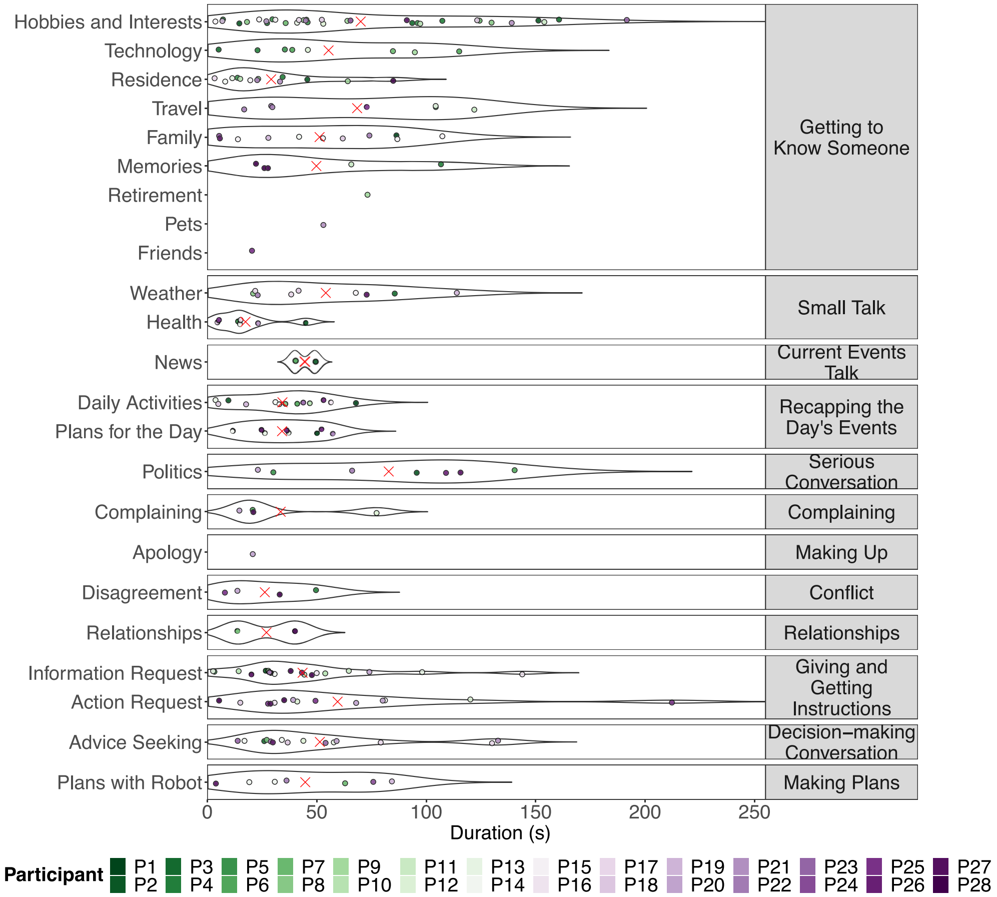

Fig. 7 Duration of topics discussed with the robot during the participatory design workshops

- Robot: I have actually read about the dragon and his life stories, but unfortunately I have never had the chance to meet him.
- P25: No. Do you know Sigurd Fafnäs, the child who killed the dragon?

No particular pattern was observed for disengagement cues in relation to the topics discussed or their duration.

6 Recommendations for integrating large language models to conversational robots

Minor technical improvements changed the user experience drastically. However, these developments were not sufficient to completely remove the presence of any of the challenges. To help address the identified challenges further, building up from the use of a widely accessible robot platform (Furhat) and frequently used components for speech recognition (Google Cloud Speech-to-Text), language generation (GPT−3.5), and speech synthesis (Amazon Polly), and drawing from the challenges encountered in the robot's conversations with older adults, we offer generalized recommendations for integrating LLMs into conversational robots, also highlighting aspects that are specific to companion robots for older adults. These recommendations aim to (1) diminish speech disruptions with advanced turn-taking approaches, low-latency LLMs, inference strategies, and speech fillers, (2) diversify response patterns by alternative LLMs or repetitiveness mitigation strategies, (3) create richer and personalized interactions with follow-up questions and lifelong learning from users, (4) bridge the linguistic gap with speech recognition algorithms trained for older adults and multi-lingual approaches, (5) augment the veracity of the robot by targeting hallucinations, updating facts, and browsing the internet, and (6) keep the conversation flow through detecting user engagement and proactively adapting the conversation.

6.1 Diminish speech disruptions

Both studies contrastively show the importance of turn-taking in HRI for user experiences, and that the disruptions in speech can have negative effects on the performance of LLMs. While gaze aversion and light indication helped decrease the interruptions drastically, they still occurred in most of the conversations. Instead of using speech recognition timeouts that need to be adjusted based on user age, turn-taking behavior can be improved by using dialogue context (e.g., Ekstedt & Skantze (2020)), speech prosody (e.g., Ekstedt & Skantze (2022)), estimating the user's gaze, gestures, and facial expressions, such as eyebrow movement and mouth opening (Danner et al., 2021), or a combination of these features (e.g., Skantze & Irfan, 2025; Johansson & Skantze, 2015; Shahverdi et al. 2022; Yang et al., 2022) (Skantze (2021) provides an in-depth review of turn-taking in HRI and conversational systems) to prevent interrupting the user, both for improving user experience and speech recognition (Marge et al., 2022).

The long response generation time, which caused unnatural interactions, can be decreased by using a smaller LLM (with the trade-off of task performance or factuality (Lee et al., 2022)), an LLM with faster inference (e.g., Baby LLaMa[40]), low-rank adaptation[41] (Hu et al., 2022), quantization techniques (see survey by Yao et al. (2023)), through streaming the LLM response (e.g., for GPT−3.5 or a streaming framework[42] (Xiao et al., 2023) on LLMs), making progressive responses (Wang et al., 2023; Reddy et al., 2023; Johnston et al., 2023), generating likely responses while the user responds (Chi et al., 2023), or by using conversation summaries rather than the full dialogue history (Zhang et al., 2018) to decrease the prompt length. However, summaries may result in further hallucinations, which could be detrimental to the interaction as well as to the personalization (of learned facts).

To overcome awkward silences during response generation, as noted by the participants, fillers, such as "Hmm" and "Let me think", can be used (Lala et al., 2019). However, overuse of them might also lead to the robot being perceived as repetitive. In addition, to decrease the superficialness of turns (sequential nature), backchannelling can be added using both verbal (e.g., "mm hm", "uh huh", "yeah") and non-verbal cues (e.g., nods, head shakes, facial expressions) (Moujahid et al., 2022), but it is important to do this in context with the dialogue, otherwise, it may also further deteriorate the user experience. Random gaze aversion should not be made during robot speech or while listening, as noted by two participants, because keeping eye contact with the user helps maintain user engagement, augment perceived sociability, and improve the quality of interaction (Kompatsiari et al., 2021).

6.2 Diversify response patterns

The increasing use of ChatGPT and other GPT−3.5 models, in addition to OpenAI prioritizing Plus subscribers in access to the servers, resulted in frequent response generation failures, as experienced in both studies, which can be mostly overcome through multiple response requests. However, this increases the already long response time. Thus, using an onboard open-source LLM,[43] such as LLaMA (2)[44] (Touvron et al., 2023) or its variants (e.g., Alpaca,[45] Vicuna,[46] Stable Beluga,[47] RedPajama INCITE,[48] Open LLaMa[49]), Falcon[50] (Penedo et al., 2023), Pythia[51] (Biderman et al., 2023), MPT,[52] Mistral,[53] Platypus[54] (Lee et al., 2023) or other methods available in HuggingFace,[55] might be more suitable to achieve more reliable response generation. However, most of these LLMs require high computational power due to their sizes, which are typically not available on robots, thus, requiring cloud-based solutions, additional hardware, or faster inference methods, described in Sect. 6.1. Nonetheless, in the case of response generation failures, the conversation can be pushed forward by invitation for elaboration (e.g., "Could you tell me more about that?"), in addition to clarification requests (e.g., "I don't think I understood you clearly") and backchannelling (e.g., "Hmm hmm"), but they should be varied and used to a minimum to avoid backlash from the user. Backchannelling should also be done according to context, since it might not be appropriate, e.g., when the user asked a question to the robot.

Large language models tend to be repetitive and can get stuck in a loop (Zhang et al., 2022), as evidenced in both studies, resulting in user frustration. The 'frequency penalty' parameter in GPT−3.5 can be increased to enforce the model to produce more varied responses (words). This can be further accompanied by sampling (Holtzman et al., 2020; Lee et al., 2022), unlikelihood training (Welleck et al., 2019), best-first decoding (Meister et al., 2020), or reinforcement learning from demonstration (Shi et al., 2022) to reduce repetitive

[40] https://github.com/karpathy/llama2.c

[41] https://github.com/microsoft/LoRA.

[42] https://github.com/mit-han-lab/streaming-llm.

[43] Note that these open-source LLMs became available after the experiments described in this work.

[44] https://ai.meta.com/llama/.

[45] https://github.com/tatsu-lab/stanford_alpaca

[46] https://lmsys.org/blog/2023-03-30-vicuna/.

[47] https://huggingface.co/stabilityai/StableBeluga2.

[48] https://together.ai/blog/redpajama-models-v1.

[49] https://github.com/openlm-research/open_llama

[50] https://falconllm.tii.ae/.

[51] https://github.com/EleutherAI/pythia.

[52] https://www.mosaicml.com/blog/mpt-7b.

[53] https://mistral.ai/.

[54] https://platypus-llm.github.io/.

[55] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

behavior. To reduce the LLM from getting conditioned on itself to produce the same response again, prompt initializers can be used (e.g., "Rephrase the sentence") if the previously generated response is repeated outside of user clarification requests.

6.3 Create Richer and personalized interactions

To decrease the superficialness of conversations and provide a more natural and engaging interaction, the range of topics discussed can be varied using the 'presence penalty' and 'temperature' in GPT−3.5. However, this may lead to more abrupt changes in topics, instead of smooth transitions with follow-ups (i.e., exploration vs exploitation trade-off). Instead, fine-tuning can be applied on datasets with follow-up questions (Khoo et al., 2023; Reddy et al., 2023) or based on human feedback (Ouyang et al., 2022; Wang et al., 2023) to address this issue, however, that might also limit the model's goal-directed capabilities. Combining the output of two or more LLMs for specific purposes can enable this (e.g., (Kobza et al., 2023; Shen et al., 2023; Reddy et al., 2023)), however, this could increase the response latency and require high computational power, which is typically not available on robots.

In addition, the persona prompt can be modified to use more follow-up questions and give a deeper personality (e.g., pre-defined preferences, dislikes, memories) to the robot to improve the believability of the character, instead of starting from 'scratch'. Moreover, the generated responses can be combined with hand-crafted dialogues to maintain a coherent persona (Konrád et al., 2021). The believability can further be improved by generating appropriate emotions for the agent through large language models (Mishra et al., 2023), in addition to detecting the user's emotions to adapt the conversation accordingly (Irfan et al., 2020). Emotions can also be used as unsupervised reinforcement learning from human feedback (RLHF) signals for task performance (Lin et al., 2020), such that frustration in a conversation topic or phrase can be avoided in future interactions.

To generate deeper conversations that rely on shared history with the user, their preferences, dislikes, and daily activities, the robot should learn from the user, as noted by the participants in the first study. While in our current work, we stored the learned facts in a knowledge-base, over long-term interactions, this is not scalable, as prompts based on an ever-growing database would be too long. In addition, using a retrieval-based method and a database to update the results of LLMs with new data can result in hallucinations (Zhang & Choi, 2021). One option is to retrain models with new information, however, this would be computationally inefficient with vast amounts of data, in addition to resulting in 'catastrophic forgetting' of previously learned information (McCloskey & Cohen, 1989; Irfan et al., 2021). 'Lifelong (or continual) learning' (Parisi et al., 2019) aims to address this problem by continually learning (the parameters of the model) over time to accommodate new knowledge (i.e., plasticity) while retaining previously learned information (i.e., stability). Within these methods, combining LLMs with 'parameter-expansion' methods offers the best balance of stability and plasticity (Jang et al., 2022). In comparison to '(reinforcement) learning from human feedback' approaches for LLMs (Ziegler et al., 2020; Hancock et al., 2019; Xu et al., 2022; Wang et al., 2023; Casper et al., 2023), lifelong learning does not require explicit feedback, which can be used to acquire new knowledge from conversations and update previously learned information. Lifelong learning can help personalize the interactions, which, in turn, can help mitigate the negative user experiences that may arise due to poor performance in dialogue (Irfan et al., 2020), in addition to improving user engagement (Oertel et al., 2020), acceptability, and trust (Whelan et al., 2018). Despite its benefits in enabling learning and adaptation, only a few studies have incorporated lifelong learning into LLMs (Jin et al., 2022; Qin et al., 2022; Jang et al., 2022; Scialom et al., 2022; Cahyawijaya et al., 2023; Wang et al., 2023; Xiang et al., 2023; Ke et al., 2023; Yue et al., 2023; Luo et al., 2023; Razdaibiedina et al., 2023; Gupta et al., 2023), with open-domain dialogue yet to be explored.

Over long-term deployments in real-world environments (e.g., homes, or senior care centers), it is important to have full autonomy, such that the user can start the interaction without the need for a wizarded system. For a personalized companion robot, this requires detecting new users and incrementally learning them for user recognition to prevent sharing the information learned about the user with others, hence, an 'open world' user recognition algorithm, such as the Multi-modal Incremental Bayesian Network (Irfan et al., 2021), can be used.

For personalizing daily interactions through recommendations based on dialogue history, retrieval and memory-augmented models (e.g., (Xu et al., 2022)), 'in-context learning' (Dong et al., 2023) and 'chain of thought' (Wei et al., 2023; Wang et al., 2023) (i.e., dividing the prompt into steps) reasoning or planning (Wang & Lim, 2023) can be used. Chen et al. (2023) provide a comprehensive survey on LLMs for personalization.

With personalization, the privacy of the interaction should be considered with utmost importance, as indicated by the older adults in our studies. Since current offline LLMs require vast computing power, it is challenging to deploy them on robots, but online models, as we did in our studies, risk the privacy of the individual. While the first interaction with a robot, as evidenced in our results, is mostly superficial, deeper conversations can be achieved over long-term interactions, in which users can talk about sensitive topics (Irfan & Skantze, 2025). The risks, trade-offs, and possible solutions to achieve privacy are discussed in depth in the works by Bommasani et al. (2022), Weidinger et al. (2021), and Zhang et al. (2023), which should be considered in parallel to the challenges highlighted in this work, while developing conversational robots with LLMs, especially for older adults who may be unaware of these risks.

6.4 Bridge the linguistic gap

While speech recognition failures were low in both studies, their rare presence led to abrupt disruptions in the conversation flow. Since older adult speech contains pauses and hesitancy, and is slower than younger adult speech, it is challenging to obtain good performance in speech recognition. One way to address this is to evaluate different speech recognition algorithms with older adult speech to choose the algorithm that works the best, such as Microsoft Azure or Whisper JAX,[56] which optimizes Whisper for speed that can enable real-time transcription of the user utterances. Werner et al. (2019) compare speech recognition algorithms trained on older adults' speech. Moreover, parsing can be used on speech recognition outputs (e.g., from Google Cloud) to add punctuation marks, to overcome disambiguities in speech for the LLM, especially for long responses.

In addition, foreign language exposure in daily life affects the proficiency of a second language (Bonfieni et al., 2019), which may be lower in later stages of life, thus, increasing the level of complexity in interacting with the robot, which was evident in the feedback from the older adults in the first study. Thus, it is important to adapt to the native language of the user. However, for multi-lingual households, learning a second language, or for foreign visitors, a multi-lingual LLM, in combination with multi-lingual speech recognition and generation, would allow changing the conversation language when the user response (prompt) is in a different language, thus, improving the capabilities of the robot and the user perceptions. However, allowing multiple languages to be transcribed may result in further interaction failures, especially due to overlapping words in different languages that may mean different things.

6.5 Augment the veracity

Disinformation can be risky for everyone, but it is especially true for older adults who trust robots in health-related issues (Giorgi et al., 2023). For instance, if the user asks for advice on medicine, and the LLM produces a wrong answer, the consequences can be critical. Thus, it is important to ensure that correct information is provided by LLMs. One approach is to prevent response generation in sensitive or unsafe topics with filtering (Weidinger et al., 2021; Xu et al., 2021; Reddy et al., 2023; Kobza et al., 2023). The hallucinations in the model can also be mitigated by evaluating uncertainty in responses (e.g., Xiao et al. (2021)), reasoning for knowledge Adolphs et al. (2021), augmenting LLMs with a memory (e.g., Madaan et al. (2023)), or applying attention mechanism (e.g., Wu et al. (2021); Leng et al. (2023)), regularization (e.g., Lee et al. (2019)), or retrieval-based methods (e.g., Dziri et al. (2021)). Ji et al. (2023) provide a detailed survey on hallucinations in LLMs with further suggestions. It is also important to recover from safety failures in conversation when they happen, such as through fine-tuning on dialogues with such recoveries (Ung et al., 2022).

In addition, due to their training data cut-off dates, LLMs contain obsolete information, which was present in the interactions in the second study, that can be addressed by fact-checking in knowledge bases (Gupta et al., 2022). As suggested in Sect. 6.3, lifelong learning can be applied to update the information over time. However, this is a slippery slope, as adversarial users can teach the robot incorrect facts or behavior styles, as was the case of Microsoft's Tay (Davis, 2016), which learned from users and became offensive within 16 h of its launch and had to be shut down. Hence, it is important to include strategies that can overcome these to maintain factual accuracy and the intended persona, which Ju et al. (2022) compare in detail in their work. Facts can be also extracted from the internet (such as weather) with LLMs that have browsing capabilities (e.g., AutoGPT,[57] Google Bard,[58] GPT-4,[59] BlenderBot3 (Shuster et al., 2022)) or through tool use (e.g., LangChain[60]) in combination with accuracy checking. Fine-tuning can be used to decide when to search the internet, generate a corresponding query, and incorporate the response into the dialogue (Shuster et al., 2022).

6.6 Keep the conversation flow

Our work aims to design a personal companion robot that decreases loneliness in older adults and supports them in their daily lives. If a robot tries to end a conversation in a short duration, this could further deepen their feelings of loneliness. Similarly, for chatbots or other conversational robots, it is important to keep the user engaged in the conversation to provide effective social support. One possible solution is to detect engagement or enjoyment in the conversation in both the user's and agent's responses (Irfan et al., 2024b; Janssens et al., 2025; Pereira et al., 2024), as well as through gaze and other non-verbal or contextual information. Oertel et al. (2020) provide an extensive survey on user engagement in HRI, and Johnston et al. (2023) refer to engagement detection methods for LLMs. Upon detection of disengagement,

[56] https://huggingface.co/spaces/sanchit-gandhi/whisper-jax.

[57] https://github.com/Significant-Gravitas/Auto-GPT.

[58] https://bard.google.com/.

[59] https://chat.openai.com/.

[60] https://www.langchain.com/

proactive prompts can be used to change the topic (Wang et al., 2023), personalize the conversation topic according to user preferences and shared history (Shen et al., 2023; Wang et al., 2023; Konrád et al., 2021; Kobza et al., 2023), ask the user what they would like to talk about, or make the current topic more engaging (Irvine et al., 2023; Fan et al., 2023) by avoiding generic small talk (Wang et al., 2023) to encourage continuation in the conversation.

7 Limitations and future work

This work identified the challenges of applying LLMs to conversational companion robots for open-domain dialogue with older adults, as frequent interruptions and slow responses, repetitive responses, superficial conversations, language barriers, hallucinations and obsolete information, and disengagement cues. These challenges are by no means an exhaustive list, but aim to provide initial insights for developing companion robots with LLMs that can be part of the everyday lives of older adults. They also show the complexity of multi-modal interactions, in comparison to the text-based interactions that LLMs are typically evaluated on, especially for a population that is not familiar with the current state of the technology.

The LLM (GPT−3.5) and the robot (Furhat) used in this work, in addition to the technical design decisions made, have led to the identification of dialogue disruptions that negatively affected older adults' experiences with companion robots. However, the identified challenges may or may not replicate in other LLMs or robots, depending on their open-domain dialogue capabilities, modalities used, and the additional libraries that can address these challenges.

Moreover, as previously described, in both studies, only a few participants (2 in the first study, 5 in the second) had previously interacted with a robot, with 2 participants talking to a wizarded robot prior to the first study, and only 1 participant talking to an (unknown) robot prior to the second study. Hence, this was the participants' first interaction with a (social) robot in the majority of the cases, leading to the 'novelty effect' (Kahn et al., 2004; Smedegaard , 2019) (i.e., users' perceptions and behaviors are affected by the novelty of the technology). Nonetheless, their perceptions might have also been affected by their prior experience with spoken dialogue systems (e.g., Alexa, Siri, Google Assistant).

Talking to a robot as part of an experiment rather than having the robot 'in the wild' (e.g., at home or a senior care center) may lead to changes in the behavior due to being observed (Irfan et al., 2018) and create an artificial context within conversations, preventing to unlock the full range of topics in open-domain conversation and corresponding challenges that can arise in conversations with a companion robot over long-term interactions (Skantze & Doğruöz, 2023). The topics discussed and user perceptions might have further been influenced by the 'priming effect' (Segaert, 2020) based on the discussions around the design scenarios prior to the robot interaction in the second study. Swedish culture might have also influenced these factors (Marion, 2017; Haring et al., 2014). In addition, the participants in both studies had volunteered (with only a small compensation offered in the second study), which indicates an interest in robots (i.e., 'participation bias') (Irfan et al., 2018) and may not reflect the views of the general population of older adults. Moreover, our findings were derived from the user experiences of healthy older adults aged 66 to 86 years, which may not necessarily extend to more senior adults or individuals with cognitive impairments. Nonetheless, these studies aimed to identify the primary obstacles that may arise in conversations and overcome them, in addition to understanding the needs, preferences, and expectations of older adults through iterative participatory design, prior to exposing them to new technology in their homes or senior care centers to avoid backlash with fear, annoyance, or reluctance to use robots.

Our findings showed that LLMs when applied in a zero-shot fashion are not yet ready to be deployed on conversational robots, and require additional components, as highlighted in Sect. 6, to provide an acceptable user experience for older adults. When these developments are applied and confirmed to be sufficient with minimal or zero errors in interactions, iterative participatory design with older adults can be continued. In that case, comparing the robot to a human and Wizard of Oz with restricted perception (e.g., robot camera and microphone feed) as baselines would allow evaluating whether enhanced LLMs can provide a user experience as good as humans, and what are the additional challenges in achieving that, which we intend to do in future work. However, it is important to consider that humans interact differently with robots than with humans (Fischer , 2011; Reimann et al., 2023), and humans will understand the situation to respond correctly even with restricted perception (Ambady & Rosenthal, 1992), which is not possible to achieve with robots in the current state of technology (Riek, 2012). Nonetheless, when the robot is sufficiently well-perceived in comparison to humans, we intend to deploy it for a real-world long-term study in older adults' homes, which can overcome the unnaturalness of conversations and perhaps show deeper challenges to address.

While the challenges and the corresponding recommendations presented here for applying LLMs to conversational robots are based on the interactions with older adults, they may also hold for socially assistive robots or companion robots with other populations, as well as general-purpose robots, chatbots, and other spoken dialogue systems, in other words, wherever an open-domain conversation may take place.

Appendix A: Hyperparameters of GPT−3.5

Table 4 Hyperparameters of GPT−3.5 used in both studies.

Hyperparameter	Value	Range
Temperature	0.9	[0,1]
Maximum length	50[1]	[1,4000]
Top P	1.0	[0,1]
Frequency Penalty	0.5[2]	[0,2]
Presence penalty	0.6	[0,2]
Stop words	Person:, Furhat:	
Best of	1	[1,20]

[1] The maximum length for tokens was decreased (default value 150) to generate shorter responses to keep the conversation flowing and engage the user

[2] The frequency penalty was increased (default value 0) to decrease the model's likelihood to repeat the same responses. Because the users may ask the robot to repeat its responses (e.g., when they don't hear it well), this value was only increased to 0.5

The default values from the OpenAI Playground for Chat preset were used, except for the maximum length and frequency penalty (see footnotes)

Appendix B: Qualitative codebook

Table 5 Dialogue disruption codes for the qualitative analysis of robot interactions

Method	Code	Explanation
Deductive	Speech detection error	Robot did not hear the participant
	Speech recognition failure	Robot misunderstands what the person said
	Empty response	Robot does not reply (empty response from GPT−3.5). The red light goes on and off, but the robot does not say anything
	Repetitive response	Robot repeats itself several times
	Turn-taking Error	Robot interrupts the participant or vice versa
	Malfunction	Robot gets stuck in a response or has a system failure
	Asking for help	Participant turns to the experimenter to ask for help
	Experimenter interference	Experimenter either talks with the participant to respond to their help request, or restarts the robot
Inductive	Hallucinations and Obsolete Information	Robot says incorrect (e.g., suggesting a non-existing restaurant) or out-of-date facts (e.g., incorrectly predicted current weather)
	Disengagement Cue	Robot replies in a way that brings the conversation to a halt, such as "I understand", "That is good to know"
	Prematurely ending conversation	Robot tries to end the conversation early
	Other	Unclassified events, such as robot mistaking the date, using made-up words (combination of English/Swedish), or avoiding answering a question and repeating it

Table 6 Topic codes for the qualitative analysis of robot interactions

Category	Code	Explanation (Talking about...)
Informal/Superficial Talk	*Getting to Know Someone*	
	Hobbies and interests	Hobbies (e.g., sports, music, literature, outdoor activities) or topics that interest the user
	Food	Recipes, likes or dislikes about food, cooking
	Travel	Locations participant visited, or would like to visit
	Residence	Where participant lives, or born in
	Technology	Asking about robot capabilities to get to know it or discusses broader topics about technology and AI
	Language	Asking or talking about languages
	Occupation	Job-related skills and activities
	Retirement	Retirement-related activities or perceptions
	Memories	Childhood memories, or memories from the past
	Family	Family members and activities with them
	Friends	Friends and activities with them
	Pets	Pets and activities with them
	Small Talk	
	Weather	Weather of the day or weather of Sweden/other countries in general
	Health	Health-related conversation (e.g., health status, doctor appointment)
	Current Events Talk	
	News	Local or international news
	Recapping the Day's Events	
	Daily Activities	Activities on the day before talking to the robot
	Plans for the Day	Activities planned after the robot interaction
Involving Talk	*Serious Conversation*	
	Politics	Local or international politics, either current or in the past
	Complaining	
	Complaining	Complaining about the robot's behavior or feature
	Making Up	
	Apology	Apology from the person or robot about their action or utterance in the conversation
	Conflict	
	Disagreement	Disagreement with the information provided by the robot
	Relationships	
	Relationships	Relationship with the robot or others (except for family and friends)
Goal-directed Talk	*Giving and Getting Instructions*	
	Information Request	Asking for information (e.g., restaurant, cinema, transportation, TV program, recipe request, weather)
	Action Request	Asking to do an action (e.g., booking a restaurant, tell a story)
	Decision-making Conversation	
	Advice Seeking	Asking to provide opinions on a subject (e.g., recommendation for a decision, activities for the weekend, travel)
	Making Plans	
	Plans with Robot	Invitation to the robot or make future plans with it

Appendix C: Study materials

Table 7 Questions used in the pre-interaction interview in the first study. Questions were adapted based on the scales in Heerink et al. (2010); Graaf et al. (2019). Questions were formulated in a semi-structured fashion with "Why/How/What?". Original questions were in Swedish

Construct	Question
Usefulness	Do you think the robot could help you reduce the experience of loneliness?
Enjoyment	Do you think the robot can make social conversation and support easier/interesting?
Sociability	Consider having a conversation with a robot about everyday things. Do you think a robot could be a nice conversation partner?
Privacy concern	Do you feel worried about your privacy with the robot?
	Would you like the robot to remember your conversation?

Table 8 Questions used in the post-interaction interview in the first study.

Construct	Question
Comprehension	Did you feel that Furhat could understand and hear you?$^{\diamond}$
Contextual Memory	Would you like Furhat to remember your conversation?
	What aspects would you like Furhat to remember?
	If Furhat could remember your conversation, do you think it would be easier to have a conversation with Furhat?
	What would you like to talk about with Furhat every day?
	What kind of conversation would you not like to have?
	Would you like to talk to Furhat about just one or two topics, or more?
	What social conversation would you most like to have without the help of Furhat?
Usefulness	Do you think Furhat could help you reduce the experience of loneliness?$^{\diamond}$
	In what kind of social conversation do you think Furhat can be helpful?
	What was the most important benefit for you from talking to Furhat?
Ease of Use	Was it easy to understand how to talk to Furhat?$^{\diamond}$
	Do you think you can use Furhat without any help?$^{\diamond}$
Enjoyment	What was your first impression of Furhat in social conversation and support?
	Do you think the robot can make social conversation and support easier/interesting?
	What did you like most about your conversation with Furhat?
	Was there anything you did not like about the conversation?
	How did it feel to have a conversation with Furhat?
Sociability	Do you consider Furhat a nice conversation partner?$^{\diamond}$
	Do you think Furhat can give you similar social conversation and support as your friends?$^{\diamond}$
	Do you feel that it would be possible to form a relationship with Furhat?
Social Presence	Did you feel that Furhat was a real person talking?$^{\diamond}$
	Did you feel that Furhat was looking at you?
	Do you consider Furhat as a living being?$^{\diamond}$
Adaptiveness	Do you think that Furhat can help you when you consider it necessary? Or do you think it will help even when it is not requested?$^{\diamond}$
	Did you feel that Furhat could adapt to your interests and needs in the conversation?
	Did you feel that Furhat was a personal robot?
	What kind of social conversation would give you a personal robot experience?
	If Furhat could adapt to your interests, what kind of new conversations would you like to have with it?
Social Influence	Do you think your family and friends would want you to use Furhat?
	Do you think it would give a good impression of you to your family and friends if you were to use Furhat?
	Do you think your family would find Furhat fascinating or boring?

Table 8 continued

Construct	Question
	Do you think you, or any other friend or family member, would find Furhat pleasant to hang out with?
	Would you like to use Furhat alone or together with friend(s) or family member(s)?
Trust	Do you think you can trust Furhat?$^{\diamond}$
	What kind of conversation with Furhat can give you a sense of security?
	Would you follow the advice that Furhat gives you?
Security	Do you feel that Furhat is safe to use in your own home?
Privacy Concern	Are you worried about using Furhat in your own home?
	Do you feel worried about your privacy when you talk to Furhat?$^{\diamond}$
Anxiety toward Robots	Can Furhat contribute to isolation and less human contact?$^{\diamond}$
	Were you worried about what to say to Furhat or how to talk to it?$^{\diamond}$
	Do you find Furhat scary?$^{\diamond}$
	Would you be afraid of making mistakes with Furhat?
	Would you be afraid of accidentally breaking Furhat?
	Is there something in Furhat that you are worried about?
Attitude towards Technology	Do you think it is a good idea to use Furhat in social conversation?$^{\diamond}$
	Do you think it would be a good idea to use Furhat in senior housing?
	What differences would there be between having Furhat in the home or in a senior care housing?

Questions were adapted based on the common questionnaires in HRI (Heerink et al., 2010; Weiss et al., 2009; Nomura et al., 2006; Graaf et al., 2019). Questions were formulated in a semi-structured fashion with "Why/How/What?". Items marked with $^{\diamond}$ were used for the questionnaire in Table 9. Original questions were in Swedish

Table 9 Post-interaction questionnaire (A-C factors) in the participatory design workshop.

Construct	Order	Question	Md	IQR
Comprehension	A.14^{+}	Linda could understand and hear me	4	1
Clarity	A.8^{+}	It was easy to understand Linda	4	1
Turn-taking (Borsci et al., 2022)	A.15^{+}	Linda could understand when I wanted to start the conversation	4	1
	A.16^{+}	Linda did not interrupt me	4	2
	B.9*§	Linda was slow to respond	3	2
Engagingness (Zhang et al., 2018; Borsci et al., 2022; Shuster et al., 2022)	B.2	Linda was engaging in the conversation	3	1.25
Consistency (Zhang et al., 2018; Shuster et al., 2022)	B.3	Linda responds in a consistent manner	4	1
Fluency (Zhang et al., 2018)	B.6	The conversation with Linda was fluent	3	1
Credibility (Borsci et al., 2022; Shuster et al., 2022)	B.17	Linda answered correctly	3.5	1
Use of Knowledge (Shuster et al., 2022)	B.5*	The conversation with Linda was insightful	3	1
Contextual Memory (Borsci et al., 2022)	B.15*	Linda could remember what I told him earlier	3	1
	B.16*	Linda could respond to what I told him earlier	3	1
Usefulness (Heerink et al., 2010)	A.3*	Linda can help me reduce the experience of loneliness	3	2
Ease of Use (Heerink et al., 2010)	A.9*	It was easy to understand how to talk to Linda	4	1
	A.10*	It was easy to start and continue conversations with Linda without any help	4	1
Enjoyment (Lee et al., 2006; Heerink et al., 2010; Iio et al., 2020)	B.1	It was fun talking to Linda	4	1
	B.4	The conversation with Linda was interesting	3	1.25
	B.7*§	It felt strange talking to Linda	3	2
	B.8	I was satisfied with my conversation with Linda	3	1
Emotion(al Influence) (Weiss et al. 2009)	B.11*	Linda made me feel happy	1	2

Table 9 continued

Construct	Order	Question	Md	IQR
	B.12^{+}	Linda made me feel sad	1	1.25
Sociability (Heerink et al., 2010)	A.1	Linda was a pleasant conversation partner	3	1.25
	A.4*	Conversations with Linda were similar to conversations with my friends or family members	2	1
	A.5*	Linda can give me similar social conversation and support as my friends	2	1.25
	B.13*	Linda could understand what I need	2	2
	B.14*	Linda could understand my feelings	2	2
Social Presence (Heerink et al., 2010)	A.17	Linda felt like a real person	3.5	1
Personality (Zhang et al., 2018; Borsci et al., 2022)	B.10	Linda had a personality	2	2
Adaptiveness (Heerink et al., 2010)	A.12	Linda can help me when I consider it necessary	3	1.5
	A.13*§	Linda will give me unnecessary advice	3	1
Trust (Heerink et al., 2010)	A.11*	I can trust Linda	3	0.5
Security (Weiss et al., 2009)	A.6*	It would be safe to use Linda in my own home	3	1.25
Privacy Concern	C.1^{+}	I am concerned about integrity with Linda	3	2.5
(Malhotra et al., 2004; Graaf et al., 2019)	C.2*	I am worried about the data collection with Linda	4	2.25
	C.3^{+}	I am worried that Linda might be recording me when I am not aware	3	3
	C.4*	I feel safer if I know how Linda uses and collects the data	4	1
	C.7	I feel anxious about having to speak out about private information with Linda	2	1.25
	C.9^{+}	I would not want Linda to remember my conversation	2	2
Anxiety toward	C.5^{+}	I am worried that Linda will lead to less human contact	2	1.25
Robots (Nomura et al., 2006; Syrdal et al., 2009; Heerink et al., 2010)	C.6	I am worried that Linda won't understand or hear what I am saying	2	1.25
	C.8	I am worried about what to say and/or how to talk to Linda	2	2
	C.10*	I think Linda was terrifying	1	1
	C.11*	I am worried about becoming too dependent on Linda	1	1
	C.12*	I would feel uncomfortable if Linda shows emotions	3	2
	C.13	I am worried about not being able to understand Linda	2	1
	C.14	I am worried that Linda may talk about unnecessary things	2	2
Attitude towards	A.2*	It is a good idea to use Linda in social conversation	4	1.25
Technology (Heerink et al., 2010)				
Intention to Use (Heerink et al. 2010)	A.7*	I would like to use Linda in my own home	2	2

Questions are presented per construct for readability. The original order is given in the second column. $^{+}$ items are created and * items are adapted for the study based on the scale. § items are reverse-coded for constructs. Likert scale is 1 to 5 (Strongly Disagree to Strongly Agree). Original questionnaire was in Swedish

Table 10 Post-interaction questionnaire (D factor) in the participatory design workshop based on Godspeed (Bartneck et al., 2009).

Construct	Question	Md	IQR
Anthropomorphism	Fake - Natural	2	1.5
	Machinelike - Humanlike	2	2.5
	Unconscious - Conscious	3	1
	Artificial - Lifelike	2	2
	Moving rigidly - Moving elegantly	2	1
Animacy	Dead - Alive	2	1
	Stagnant - Lively	2	1
	Mechanical - Organic	2	1
	Inert - Interactive	2.5	1

Table 10 continued

Construct	Question	Md	IQR
	Apathetic - Responsive	3	1
Likeability	Dislike - Like	3	1
	Unfriendly - Friendly	4	1
	Unkind - Kind	4	1
	Unpleasant - Pleasant	4	1
	Awful - Nice	4	1
Perceived Intelligence	Incompetent - Competent	3	1
	Ignorant - Knowledgeable	3	1
	Irresponsible - Responsible	3	0
	Unintelligent - Intelligent	3	1.75
	Foolish - Sensible	3	1

Likert scale is 1 to 5. Items were in Swedish (Thunberg et al., 2017)

Acknowledgements The authors would like to thank Youssef Mohamed for helping conduct the robot interactions in two of the design workshops, and Mikaela Hellstrand for helping shape the empathetic robot persona based on her research in senior care centers. The authors would like to thank the participants at the initial study and participatory design workshops for their time and feedback.

Funding Open access funding provided by Royal Institute of Technology. This work is funded by KTH Digital Futures.

Declarations

Conflict of interest G.S. is co-affiliated with Furhat Robotics, as its Co-founder and Chief Scientist. The remaining authors have no relevant financial or non-financial interests to disclose.

Compliance with Ethical Standards The study was conducted according to the standards of the Ethical Review Authority in Sweden. All participants gave an informed consent to participate in the studies. Participants, whose images and/or videos appear in this article, gave informed consent for anonymized (i.e., blurred face, and without full name released) image and video sharing in publications (conferences and journals). The study did not collect any sensitive or health-related information from the participants.

Usage of a Large Language Model in the Article B.I. used ChatGPT (OpenAI) to generate ideas for the paper title, section headers, terminology, and synonyms for words, and rephrasing sentences. The information generated by the model was never used directly: it was iterated, modified, fact-checked, based on and combined with the author's own ideas and text.

References

Abdolrahmani, A., Kuber, R., Branham, S.M.(2018). "Siri talks at you": An empirical investigation of voice-activated personal assistant (VAPA) usage by individuals who are blind. In: Proceedings of the 20th international ACM SIGACCESS conference on computers and accessibility, pp. 249– 258. ACM, Galway Ireland . https://doi.org/10.1145/3234695.3236344

Adiwardana, D., Luong, M.-T., So, D.R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., Le, Q.V. (2020). Towards a human-like open-domain chatbot. Preprint at arXiv: org/abs/2001.09977

Adolphs, L., Shuster, K., Urbanek, J., Szlam, A., Weston, J. (2021). Reason first, then respond: Modular generation for knowledge-infused dialogue. Preprint at arXiv: org/abs/2111.05204

Al Moubayed, S., Beskow, J., Skantze, G., Granström, B.(2012).Furhat: A back-projected human-like robot head for multiparty human-machine interaction. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) Cognitive behavioural systems, pp. 114– 130. Springer, Berlin, Heidelberg https://doi.org/10.1007/978-3-642-34584-5_9

Alves-Oliveira, P., Arriaga, P., Paiva, A., Hoffman, G.(2021). Children as robot designers. In: Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction. HRI '21, pp. 399– 408. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3434073.3444650

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*(2), 256.

Axelsson, A., Skantze, G.(2023). Do you follow? a fully automated system for adaptive robot presenters. In: Proceedings of the 2023 ACM/IEEE international conference on human–robot interaction. HRI '23, pp. 102– 111. Association for computing machinery, New York, NY, USA . https://doi.org/10.1145/3568162.3576958

Azenkot, S., Feng, C., Cakmak, M.(2016). Enabling building service robots to guide blind people a participatory design approach. In: 2016 11th ACM/IEEE international conference on human–robot interaction (HRI), pp. 3– 10 . https://doi.org/10.1109/HRI.2016.7451727

Babel, F., Kraus, J., Miller, L., Kraus, M., Wagner, N., Minker, W., & Baumann, M. (2021). Small talk with a robot? The impact of dialog content, talk initiative, and gaze behavior of a social robot on trust, acceptance, and proximity. *International Journal of Social Robotics, 13*(6), 1485–1498. https://doi.org/10.1007/s12369-020-00730-0

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Bedaf, S., Marti, P., & De Witte, L. (2019). What are the preferred characteristics of a service robot for the elderly? A multi-country focus group study with older adults and caregivers. *Assistive Technology, 31*(3), 147–157. https://doi.org/10.1080/10400435.2017.1402390

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.(2021). On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. FAccT '21, pp. 610– 623. Association for computing machinery, New York, NY, USA . https://doi.org/10.1145/3442188.3445922

Benjamin, B. J. (1997). Speech production of normally aging adults. *Seminars in Speech and Language, 18* (pp. 135–141). New York, NY, USA: Thieme Medical Publishers Inc.

Bernstein, B. (1962). Social class, linguistic codes and grammatical elements. *Language and Speech, 5*(4), 221–240.

Bickmore, T. W., & Cassell, J. (1999). Small talk and conversational storytelling in embodied conversational interface agents. *AAAI fall symposium on narrative intelligence* (pp. 87–92). FL, USA: Orlando.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M.A., Purohit, S., Prashanth, U.S., Raff, E., Skowron, A., Sutawika, L., Wal, O (2023). Pythia: A suite for analyzing large language models across training and scaling. Preprint at arXiv: org/abs/2304.01373

BigScience Workshop (2023): Le Scao, Teven, et al.: BLOOM: A 176B-parameter open-access multilingual language model. Preprint at arXiv: org/abs/2211.05100

Björling, E. A., & Rose, E. (2019). Participatory research principles in human-centered design: Engaging teens in the co-design of a social robot. *Multimodal Technologies and Interaction, 3*(1), 8. https://doi.org/10.3390/mti3010008

Blair, J., Abdullah, S(2019). Understanding the Needs and challenges of using conversational agents for deaf older adults. In: Conference companion publication of the 2019 on computer supported cooperative work and social computing, pp. 161– 165. ACM, Austin, TX, USA . https://doi.org/10.1145/3311957.3359487

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P. (2022). On the opportunities and risks of foundation models. Preprint at arXiv: org/abs/2108.07258

Bonfieni, M., Branigan, H. P., Pickering, M. J., & Sorace, A. (2019). Language experience modulates bilingual language control: The effect of proficiency, age of acquisition, and exposure on language switching. *Acta Pathologica, Microbiologica et Immunologica Scandinavica, 193*, 160–170.

Borsci, S., Malizia, A., Schmettow, M., Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The chatbot usability scale: The design and pilot of a usability scale for interaction with AI-based conversational agents. *Personal and Ubiquitous Computing, 26*(1), 95–119. https://doi.org/10.1007/s00779-021-01582-9

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.(2020). Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H (eds.) Advances in neural information processing systems, vol. 33, pp. 1877–1901. Curran Associates, Inc., Virtual . Preprint at arXiv: org/abs/2005.14165

Cacioppo, J. T., Hughes, M. E., Waite, L. J., Hawkley, L. C., & Thisted, R. A. (2006). Loneliness as a specific risk factor for depressive symptoms: Cross-sectional and longitudinal analyses. *Psychology and Aging, 21*(1), 140–151. https://doi.org/10.1037/0882-7974.21.1.140

Cahyawijaya, S., Lovenia, H., Yu, T., Chung, W., Fung, P.(2023). Instruct-Align: Teaching novel languages with to llms through alignment-based cross-lingual instruction. Preprint at arXiv: org/abs/2305.13627

Caleb-Solly, P., Dogramadzi, S., Ellender, D., Fear, T., Heuvel, H.v.d.(2014). A mixed-method approach to evoke creative and holistic thinking about robots in a home environment. In: Proceedings of the 2014 ACM/IEEE international conference on human–robot interaction. HRI '14, pp. 374– 381. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2559636.2559681

Casper, S., Davies, X., Shi, C., Gilbert, T.K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E.J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., Hadfield-Menell, D.(2023). Open problems and fundamental limitations of reinforcement learning from human feedback. Preprint at arXiv: org/abs/1230.71521

Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., Pu, Y., Lei, Y., Chen, X., Wang, X., Lian, D., Chen, E.(2023). When large language models meet personalization: perspectives of challenges and opportunities. Preprint at arXiv: org/abs/2307.16376

Cherakara, N., Varghese, F., Shabana, S., Nelson, N., Karukayil, A., Kulothungan, R., Farhan, M., Nesset, B., Moujahid, M., Dinkar, T., Rieser, V., Lemon, O.(2023). Furchat: An embodied conversational agent using llms, combining open and closed-domain dialogue with facial expressions. In: Proceedings of the 24th annual meeting of the special interest group on discourse and dialogue (SigDIAL)

Chi, R.A., Kim, J., Hickmann, S., Li, S., Chi, G., Atchariyachanvanit, T., Yu, K., Chi, N.A., Dai, G., Rammoorthy, S., Wang, J.H., Sarthi, P., Adams, V., Xu, B.Y., Xu, B.Z., Park, K., Cao, S., Manning, C.D.(2023). Dialogue distillery: Crafting interpolable, interpretable, and introspectable dialogue from llms. In: Alexa Prize SocialBot grand challenge 5 proceedings . https://www.amazon.science/alexa-prize/proceedings/chirpy-cardinal-dialogue-distillery-crafting-interpolable-interpretable-and-introspectable-dialogue-from-llms

Chung, K., Oh, Y.H., Ju, D.Y.(2019). Elderly users' interaction with conversational agent. In: Proceedings of the 7th international conference on human-agent interaction, pp. 277–279. ACM, Kyoto Japan. https://doi.org/10.1145/3349537.3352791

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Cumbal, R., Moell, B., Lopes, J., & Engwall, O. (2021). "You don't understand me!": Comparing ASR results for L1 and L2 speakers of Swedish. In: *Interspeech 2021* (pp. 4463–4467). International Speech Communication Association (ISCA). https://doi.org/10.21437/Interspeech.2021-2140.

Danner, S. G., Krivokapić, J., & Byrd, D. (2021). Co-speech movement in conversational turn-taking. *Frontiers in Communication*. https://doi.org/10.3389/fcomm.2021.779814

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences, 362*(1480), 679–704. https://doi.org/10.1098/rstb.2006.2004

Davis, E. (2016). AI amusements: The tragic tale of Tay the chatbot. *AI Matters, 2*(4), 20–24. https://doi.org/10.1145/3008665.3008674

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cielibak, M. (2020). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review, 54*, 755–810. https://doi.org/10.1007/s10462-020-09866-x

Doğruöz, A.S., Skantze, G.(2021). How "open" are the conversations with open-domain chatbots? A proposal for speech event based evaluation. In: Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue, pp. 392–402. Association for computational linguistics, Singapore and Online . https://aclanthology.org/2021.sigdial-1.41

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., Sui, Z. (2023). A survey on in-context learning. Preprint at arXiv: org/abs/2301.00234

Dziri, N., Madotto, A., Zaïane, O., Bose, A.J.(2021). Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp. 2197– 2214. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic . https://doi.org/10.18653/v1/2021.emnlp-main.168 . https://aclanthology.org/2021.emnlp-main.168

Ekstedt, E., Skantze, G.(2020). TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In: Findings of the association for computational linguistics: EMNLP 2020, pp. 2981– 2990. Association for computational linguistics, Online . https://doi.org/10.18653/v1/2020.findings-emnlp.268 . https://aclanthology.org/2020.findings-emnlp.268

Ekstedt, E., Skantze, G.(2022). How much does prosody help turn-taking? investigations using voice activity projection models. In: Proceedings of the 23rd annual meeting of the special interest group on discourse and dialogue, pp. 541– 551. Association for Computational Linguistics, Edinburgh, UK . https://aclanthology.org/2022.sigdial-1.51

Elgarf, M., Skantze, G., Peters, C.(2021). 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, pp. 60– 67 . https://doi.org/10.1145/3472306.3478359

Fan, Y., Bowden, K.K., Cui, W., Chen, W., Harrison, V., Ramirez, A., Agashe, S., Liu, X.G., Pullabhotla, N., Jeshwanth Bheemanpally, N.Q., Garg, S., Walker, M., Wang, X.E.(2023). Athena 3.0: Personalized multimodal chatbot with neuro-symbolic dialogue generators. In: Alexa Prize SocialBot Grand challenge 5 proceedings . https://www.amazon.science/alexa-prize/proceedings/athena-3-0-personalized-multimodal-chatbot-with-neuro-symbolic-dialogue-generators

Fernández-Rodicio, E., Castro-González, A., Alonso-Martín, F., Maroto-Gómez, M., & Salichs, M. A. (2020). Modelling multimodal dialogues for social robots using communicative acts. *Sensors*. https://doi.org/10.3390/s20123440

Fischer K.(2011). Interpersonal variation in understanding robots as social actors. In: Proceedings of the 6th international conference on human–robot interaction. HRI '11, pp. 53– 60. Association for computing machinery, New York, NY, USA . https://doi.org/10.1145/1957656.1957672

Frennert, S., & Östlund, B. (2014). Review: Seven matters of concern of social robots and older people. *International Journal of Social Robotics, 6*(2), 299–310. https://doi.org/10.1007/s12369-013-0225-8

Funakoshi, K., Nakano, M., Torii, T., Hasegawa, Y., Tsujino, H., Kimura, N., Iwahashi, N.(2007). Robust acquisition and recognition of spoken location names by domestic robots. In: 2007 IEEE/RSJ International conference on intelligent robots and systems, pp. 1435– 1440 . IEEE

Gasteiger, N., Ahn, H. S., Lee, C., Lim, J., MacDonald, B. A., Kim, G. H., & Broadbent, E. (2022). Participatory design, development, and testing of assistive health robots with older adults: An international four-year project. *Transactions on Human-Robot Interaction*. https://doi.org/10.1145/3533726

Giorgi, I., Minutolo, A., Tirotto, F., Hagen, O., Esposito, M., Gianni, M., Palomino, M., & Masala, G. L. (2023). I am robot, your health adviser for older adults: Do you trust my advice? *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-023-01019-8

Goldsmith, D. J., & Baxter, L. A. (1996). Constituting relationships in talk: A taxonomy of speech events in social and personal relationships. *Human Communication Research, 23*(1), 87–114.

Gollasch, D., Weber, G.(2021). Age-related differences in preferences for using voice assistants. In: Mensch und Computer 2021, pp. 156– 167. ACM, Ingolstadt, Germany . https://doi.org/10.1145/3473856.3473889

Graaf, M. M. A., Allouch, S. B., & Dijk, J. A. G. M. (2019). Why would I use this in my home? A model of domestic social robot acceptance. *Human-Computer Interaction, 34*(2), 115–173. https://doi.org/10.1080/07370024.2017.1312406

Gupta, K., Thérien, B., Ibrahim, A., Richter, M.L., Anthony, Q., Belilovsky, E., Rish, I., Lesort, T. (2023). Continual pre-training of large language models: How to (re)warm your model? Preprint at arXiv: org/abs/2308.04014

Gupta, P., Wu, C.-S., Liu, W., Xiong, C.(2022). DialFact: A benchmark for fact-checking in dialogue. Preprint at arXiv: org/abs/2110.08222

Hancock, B., Bordes, A., Mazaré, P.-E., Weston, J.(2019). Learning from dialogue after deployment: Feed yourself, chatbot! Preprint at arXiv: org/abs/1901.05415

Haring, K., Mougenot, C., Ono, F., & Watanabe, K. (2014). Cultural differences in perception and attitude towards robots. *International Journal of Affective Engineering, 13*, 149–157. https://doi.org/10.5057/ijae.13.149

Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults: The Almere model. *International Journal of Social Robotics, 2*(4), 361–375. https://doi.org/10.1007/s12369-010-0068-5

Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y. (2020). The curious case of neural text degeneration. Preprint at arXiv: org/abs/1904.09751

Honig, S., & Oron-Gilad, T. (2018). Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2018.00861

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.(2022). LoRA: Low-rank adaptation of large language models. In: International conference on learning representations . https://openreview.net/forum?id=nZeVKeeFYf9

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. *Applied Sciences, 10*(4), 2076–3417. https://doi.org/10.3390/app10041522

Inoue, K., Lala, D., Yamamoto, K., Nakamura, S., Takanashi, K., Kawahara, T. (2020). An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. In: Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue, pp. 118– 127. Association for computational linguistics, 1st virtual meeting . https://aclanthology.org/2020.sigdial-1.15

Irfan, B., Hellou, M., Mazel, A., Belpaeme, T. (2020). Challenges of a real-world HRI study with non-native english speakers: Can personalisation save the day? In: Companion of the 2020 ACM/IEEE international conference on human–robot interaction. HRI '20, pp. 272– 274. Association for computing machinery, New York, NY, USA . https://doi.org/10.1145/3371382.3378278

Irfan, B., Narayanan, A., & Kennedy, J. (2020). Dynamic emotional language adaptation in multiparty interactions with agents. In: Proceedings of the 20th ACM international conference on intelligent virtual agents. IVA '20. Association for Computing machinery, New York, NY, USA https://doi.org/10.1145/3383652.3423881

Irfan, B., Hellou, M., & Belpaeme, T. (2021). Coffee with a hint of data: Towards using data-driven approaches in personalised long-term interactions. *Frontiers in Robotics and AI, 8*, 300. https://doi.org/10.3389/frobt.2021.676814

Irfan, B., Kennedy, J., Lemaignan, S., Papadopoulos, F., Senft, E., & Belpaeme, T. (2018). Social psychology and human-robot interaction: An uneasy marriage. In: Companion of the 2018 ACM/IEEE international conference on human-robot interaction, pp. 13–20. ACM, Chicago, IL, USA https://doi.org/10.1145/3173386.3173389

Irfan, B., Kuoppamäki, S., & Skantze, G. (2024). Recommendations for designing conversational companion robots with older adults through foundation models. *Frontiers in Robotics and AI*, 11. https://doi.org/10.3389/frobt.2024.1363713

Irfan, B., Miniota, J., Thunberg, S., Lagerstedt, E., Kuoppamäki, S., Skantze, G., & Pereira, A. (2024b). Human-robot interaction conversational user enjoyment scale (HRI CUES). arXiv:2405.01354

Irfan, B., Ortiz, M. G., Lyubova, N., & Belpaeme, T. (2021). Multimodal open world user identification. *Transactions on Human-Robot Interaction*. https://doi.org/10.1145/3477963

Irvine, R., Boubert, D., Raina, V., Liusie, A., Zhu, Z., Mudupalli, V., Korshuk, A., Liu, Z., Cremer, F., Assassi, V., Beauchamp, C.-C., Lu, X., Rialan, T., Beauchamp, W.(2023). Rewarding chatbots for real-world engagement with millions of users. Preprint at arXiv: org/abs/2303.06135

Irfan, B., & Skantze, G. (2025). Between you and me: Ethics of self-disclosure in human-robot interaction. In *Companion of the 2025 ACM/IEEE international conference on human-robot interaction*. IEEE.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*(12), 1211–1217. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Jang, J., Ye, S., Yang, S., Shin, J., Han, J., KIM, G., Choi, S.J., Seo, M.(2022). Towards continual knowledge learning of language models. In: International conference on learning representations, virtual https://openreview.net/forum?id=vfsRB5MImo9

Jang, J., Ye, S., Yang, S., Shin, J., Han, J., Kim, G., Choi, S.J., Seo, M.(2022). Towards continual knowledge learning of language models. Preprint at arXiv: org/abs/2110.03215

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement, 61*(2), 277–289. https://doi.org/10.1177/00131640121971239

Janssens, R., Pereira, A., Skantze, G., Irfan, B., & Belpaeme, T. (2025). Online prediction of user enjoyment in human-robot dialogue with LLMs. In *Companion of the 2025 ACM/IEEE international conference on human-robot interaction*. IEEE

Jenkins, S., & Draper, H. (2015). Care, monitoring, and companionship: Views on care robots from older people and their carers. *International Journal of Social Robotics, 7*(5), 673–683. https://doi.org/10.1007/s12369-015-0322-y

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 248. https://doi.org/10.1145/3571730

Jin, X., Zhang, D., Zhu, H., Xiao, W., Li, S.-W., Wei, X., Arnold, A., Ren, X.(2022). Lifelong pretraining: Continually adapting language models to emerging corpora. Preprint at arXiv: org/abs/2110.08534

Johansson, M., Skantze, G.(2015). Opportunities and obligations to take turns in collaborative multi-party human-robot interaction. In: Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue, pp. 305– 314. Association for Computational Linguistics, Prague, Czech Republic . https://doi.org/10.18653/v1/W15-4642 . https://aclanthology.org/W15-4642

Johnston, M., Flagg, C., Gottardi, A., Sahai, S., Lu, Y., Sagi, S., Dai, L., Goyal, P., Hedayatnia, B., Hu, L., Jin, D., Lange, P., Liu, S., Liu, S., Pressel, D., Shi, H., Yang, Z., Zhang, C., Zhang, D., Ball, L., Bland, K., Hu, S., Ipek, O., Jeun, J., Rocker, H., Vaz, L., Iyengar, A., Liu, Y., Mandal, A., Hakkani-Tür, D., Ghanadan, R.(2023). Advancing open domain dialog: The fifth alexa prize socialbot grand challenge. In: Alexa Prize SocialBot grand challenge 5 proceedings . https://www.amazon.science/alexa-prize/proceedings/advancing-open-domain-dialog-the-fifth-alexa-prize-socialbot-grand-challenge

Ju, D., Xu, J., Boureau, Y.-L., Weston, J. (2022). Learning from data in the mixed adversarial non-adversarial case: Finding the helpers and ignoring the trolls. Preprint at arXiv: org/abs/2208.03295

Kahn, P.H., Freier, N., Friedman, B., Severson, R., Feldman, E.(2004). Social and moral relationships with robotic others?, pp. 545– 550 . https://doi.org/10.1109/ROMAN.2004.1374819

Ke, Z., Shao, Y., Lin, H., Konishi, T., Kim, G., Liu, B. (2023). Continual pre-training of language models. Preprint at arXiv: org/abs/2302.03241

Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS), 2*(1), 26–41.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica, 26*, 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

Khoo, W., Hsu, L.-J., Amon, K.J., Chakilam, P.V., Chen, W.-C., Kaufman, Z., Lungu, A., Sato, H., Seliger, E., Swaminathan, M., Tsui, K.M., Crandall, D.J., Sabanović, S.(2023). Spill the tea: When robot conversation agents support well-being for older adults. In: Companion of the 2023 ACM/IEEE international conference on human–robot interaction. HRI '23, pp. 178– 182. Association for computing machinery, New York, NY, USA . https://doi.org/10.1145/3568294.3580067

Kim, J., Kim, S., Kim, S., Lee, E., Heo, Y., Hwang, C.-Y., Choi, Y.-Y., Kong, H.-J., Ryu, H., & Lee, H. (2021). Companion robots for older adults: Rodgers' evolutionary concept analysis approach. *Intelligent Service Robotics, 14*(5), 729–739. https://doi.org/10.1007/s11370-021-00394-3

Kobza, O., Čuhel, J., Gargiani, T., Herel, D., Marek, P.(2023). Alquist 5.0: Dialogue trees meet generative models. a novel approach for enhancing socialbot conversations. In: Alexa Prize SocialBot grand challenge 5 proceedings . https://www.amazon.science/alexa-prize/proceedings/alquist-5-0-dialogue-trees-meet-generative-models-a-novel-approach-for-enhancing-socialbot-conversations

Komeili, M., Shuster, K., Weston, J. (2021). Internet-augmented dialogue generation. Preprint at arXiv: org/abs/2107.07566

Kompatsiari, K., Ciardo, F., Tikhanoff, V., Metta, G., & Wykowska, A. (2021). It's in the eyes: The engaging role of eye contact in HRI. *International Journal of Social Robotics, 13*(3), 525–535. https://doi.org/10.1007/s12369-019-00565-4

Konrád, J., Pichl, J., Marek, P., Lorenc, P., Duy Ta, V., Kobza, O.(2021). Alquist 4.0: Towards social intelligence using generative models and dialogue personalization. In: Alexa Prize SocialBot grand challenge 4 proceedings . https://www.amazon.science/alexa-prize/proceedings/alquist-4-0-towards-social-intelligence-using-generative-models-and-dialogue-personalization

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology*. SAGE Publications Inc.

Kuoppamäki, S., Jaberibraheem, R., Hellstrand, M., & McMillan, D. (2023). Designing multi-modal conversational agents for the kitchen with older adults: A participatory design study. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-023-01055-4

Kuoppamäki, S., Tuncer, S., Eriksson, S., & McMillan, D. (2021). Designing kitchen technologies for ageing in place: A video study of older adults' cooking at home. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5*(2), 1–19. https://doi.org/10.1145/3463516

Lala, D., Inoue, K., Kawahara, T.(2019). Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In: 2019 International Conference on Multimodal Interaction. ICMI '19, pp. 226– 234. Association for computing machinery, New York, NY, USA . https://doi.org/10.1145/3340555.3353727

Lala, D., Milhorat, P., Inoue, K., Ishida, M., Takanashi, K., Kawahara, T.(2017). Attentive listening system with backchanneling, response generation and flexible turn-taking. In: Proceedings of the 18th annual sigdial meeting on discourse and dialogue, pp. 127– 136. Association for computational linguistics, Saarbrücken, Germany . https://doi.org/10.18653/v1/W17-5516 . https://aclanthology.org/W17-5516

Lazaro, M.J., Kim, S., Lee, J., Chun, J., Kim, G., Yang, E., Bilyalova, A., Yun, M.H.(2021). A review of multimodal interaction in intelligent systems. In: Kurosu, M. (ed.) Human-computer interaction. Theory, methods and tools, pp. 206– 219. Springer, Cham

Lee, K., Firat, O., Agarwal, A., Fannjiang, C., Sussillo, D.(2019). Hallucinations in neural machine translation. In: International conference on learning representations (ICLR)

Lee, A.N., Hunter, C.J., Ruiz, N. (2023). Platypus: Quick, cheap, and powerful refinement of LLMs. Preprint at arXiv: org/abs/2308.07317

Lee, Y.K., Jung, Y., Kang, G., Hahn, S.(2023). Developing social robots with empathetic non-verbal cues using large language models. In: 2023 32nd IEEE international conference on robot & human interactive communication (RO-MAN)

Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P.N., Shoeybi, M., Catanzaro, B.(2022). Factuality enhanced language models for open-ended text generation. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in neural information processing systems, vol. 35, pp. 34586– 34599. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/df438caa36714f69277daa92d608dd63-Paper-Conference.pdf

Lee, H.R., Šabanović, S., Chang, W.-L., Nagata, S., Piatt, J., Bennett, C., Hakken, D.(2017). Steps toward participatory design of social robots: mutual learning with older adults with depression. In: Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction, pp. 244– 253. ACM, Vienna, Austria . https://doi.org/10.1145/2909824.3020237

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents? The effects of physical embodiment, tactile interaction, and people's loneliness in human-robot interaction. *International Journal of Human-Computer Studies, 64*(10), 962–973. https://doi.org/10.1016/j.ijhcs.2006.05.002

Lekova, A., Tsvetkova, P., Andreeva, A.(2023). System software architecture for enhancing human-robot interaction by conversational AI. In: 2023 International conference on information technologies (InfoTech), pp. 1– 6 . https://doi.org/10.1109/InfoTech58664.2023.10266870

Leng, Y., Guo, Z., Shen, K., Tan, X., Ju, Z., Liu, Y., Liu, Y., Yang, D., Zhang, L., Song, K., He, L., Li, X.-Y., Zhao, S., Qin, T., Bian, J. (2023) PromptTTS 2: Describing and generating voices with text prompt. Preprint at arXiv: org/abs/2309.02285

Lin, J., Ma, Z., Gomez, R., Nakamura, K., He, B., & Li, G. (2020). A review on interactive reinforcement learning from human social feedback. *IEEE Access, 8*, 120757–120765. https://doi.org/10.1109/ACCESS.2020.3006254

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., Zhang, Y. (2023). an empirical study of catastrophic forgetting in large language models during continual fine-tuning. Preprint at arXiv: org/abs/2308.08747

Luo, Y., Hawkley, L. C., Waite, L. J., & Cacioppo, J. T. (2012). Loneliness, health, and mortality in old age: A national longitudinal study. *Social Science and Medicine, 74*(6), 907–914. https://doi.org/10.1016/j.socscimed.2011.11.028

Madaan, A., Tandon, N., Clark, P., Yang, Y.(2023). Memory-assisted prompt editing to improve GPT-3 after deployment. Preprint at arXiv: org/abs/2201.06009

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research, 15*(4), 336–355.

Maniscalco, U., Storniolo, P., & Messina, A. (2022). Bidirectional multi-modal signs of checking human-robot engagement and interaction. *International Journal of Social Robotics, 14*(5), 1295–1309.

Marge, M., Espy-Wilson, C., Ward, N. G., Alwan, A., Artzi, Y., Bansal, M., Blankenship, G., Chai, J., Daumé, H., Dey, D., et al. (2022). Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language, 71*, 101255.

Marion, G. (2017). How culture affects language and dialogue. *The routledge handbook of language and dialogue* (pp. 347–366). New York: Routledge.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation, 24*, 109–165. https://doi.org/10.1016/S0079-7421(08)60536-8. Academic Press

McLaughlin, M. L., & Cody, M. J. (1982). Awkward silences: Behavioral antecedent and consequences of the conversational lapse. *Human Communication Research, 8*(4), 299–316. https://doi.org/10.1111/j.1468-2958.1982.tb00669.x

McMillan, D., Brown, B., Kawaguchi, I., Jaber, R., Solsona Belenguer, J., Kuzuoka, H.(2019).: Designing with Gaze: Tama - a Gaze Activated Smart-Speaker. Proceedings of the ACM on human-computer interaction 3(CSCW), 1– 26 https://doi.org/10.1145/3359278

Meister, C., Vieira, T., & Cotterell, R. (2020). Best-first beam search. *Transactions of the Association for Computational Linguistics, 8*, 795–809. https://doi.org/10.1162/tacl_a_00346

Mishra, C., Verdonschot, R., Hagoort, P., & Skantze, G. (2023). Real-time emotion generation in human-robot dialogue using large language models. *Frontiers in Robotics and AI, 10*, 1271610. https://doi.org/10.3389/frobt.2023.1271610

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine, 19*(2), 98–100. https://doi.org/10.1109/MRA.2012.2192811

Moujahid, M., Hastie, H., Lemon, O. (2022). Multi-party interaction with a robot receptionist. In: 2022 17th ACM/IEEE international conference on human–robot interaction (HRI), pp. 927– 931. https://doi.org/10.1109/HRI53351.2022.9889641

Murali, P., Steenstra, I., Yun, H.S., Shamekhi, A., Bickmore, T.(2023). Improving multiparty interactions with a robot using large language models. In: Extended abstracts of the 2023 chi conference on human factors in computing systems. CHI EA '23. Association for computing machinery, New York, NY, USA . https://doi.org/10.1145/3544549.3585602

Nomura, T., Suzuki, T., Kanda, T., Kato, K.(2006). Measurement of anxiety toward robots. In: ROMAN 2006 - The 15th IEEE international symposium on robot and human interactive communication, pp. 372– 377 . https://doi.org/10.1109/ROMAN.2006.314462

Oertel, C., Castellano, G., Chetouani, M., Nasir, J., Obaid, M., Pelachaud, C., & Peters, C. (2020). Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI*. https://doi.org/10.3389/frobt.2020.00092

Ostrowski, A.K., Breazeal, C., Park, H.W.(2021).Long-term co-design guidelines: Empowering older adults as co-designers of social robots. In: 2021 30th IEEE international conference on robot & human interactive communication (RO-MAN), pp. 1165– 1172 . https://doi.org/10.1109/RO-MAN50785.2021.9515559

Ostrowski, A. K., DiPaola, D., Partridge, E., Park, H. W., & Breazeal, C. (2019). Older adults living with social robots: Promoting social connectedness in long-term communities. *IEEE Robotics & Automation Magazine, 26*(2), 59–70. https://doi.org/10.1109/MRA.2019.2905234

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R. (2022) Training language models to follow instructions with human feedback. Preprint at arXiv: org/abs/2203.02155

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.(2022). Training language models to follow instructions with human feedback. Preprint at arXiv: org/abs/2203.02155

Paetzel, M., Perugia, G., Castellano, G.(2020). The persistence of first impressions: The effect of repeated interactions on the perception of a social robot. In: Proceedings of the 2020 ACM/IEEE international conference on human–robot interaction. HRI '20, pp. 73– 82. Association for computing machinery, New York, NY, USA .https://doi.org/10.1145/3319502.3374786

Paradeda, R.B., Hashemian, M., Rodrigues, R.A., Paiva, A.(2016). How facial expressions and small talk may influence trust in a robot. In: Agah, A., Cabibihan, J.-J., Howard, A.M., Salichs, M.A., He, H. (eds.) Social robotics, pp. 169– 178. Springer, Cham . https://doi.org/10.1007/978-3-319-47437-3_17

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks, 113*, 54–71. https://doi.org/10.1016/j.neunet.2019.01.012

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., Launay, J. (2023). The RefinedWeb dataset for falcon LLM: Outperforming curated corpora with web data, and web data only. Preprint at arXiv: org/abs/2306.01116

Pereira, A., Marcinek, L., Miniota, J., Thunberg, S., Lagerstedt, E., Gustafson, J., Skantze, G., & Irfan, B. (2024). Multimodal user enjoyment detection in human-robot conversation: The power of large language models. In *Proceedings of the 26th international conference on multimodal interaction* (pp. 469–478). Association for Computing Machinery. https://doi.org/10.1145/3678957.3685729

Perera, V., Pereira, T., Connell, J., Veloso, M. (2017). Setting up pepper for autonomous navigation and personalized interaction with users. Preprint at arXiv: org/abs/1704.04797

Phillips, E., Zhao, X., Ullman, D., Malle, B.F.(2018). What is human-like? decomposing robots' human-like appearance using the anthropomorphic robot (abot) database. In: Proceedings of the 2018 ACM/IEEE international conference on human–robot interaction. HRI '18, pp. 105– 113. Association for computing machinery, New York, NY, USA . https://doi.org/10.1145/3171221.3171268

Pollmann, K., Ziegler, D.(2021). A pattern approach to comprehensible and pleasant human-robot interaction. Multimodal Technologies and Interaction https://doi.org/10.3390/mti5090049

Pradhan, A., Lazar, A., & Findlater, L. (2020). Use of intelligent voice assistants by older adults with low technology use. *ACM Transactions on Computer-Human Interaction, 27*(4), 1–27. https://doi.org/10.1145/3373759

Qin, Y., Zhang, J., Lin, Y., Liu, Z., Li, P., Sun, M., Zhou, J. (2022). ELLE: Efficient lifelong pre-training for emerging data. Preprint at arXiv: org/abs/2203.06311

Quinderé, M., Seabra Lopes, L., Teixeira, A.J.S.(2013). Evaluation of a dialogue manager for a mobile robot. In: 2013 IEEE RO-MAN, pp. 126– 132 . https://doi.org/10.1109/ROMAN.2013.6628466

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. Preprint at arXiv: org/abs/2212.04356

Randall, N., Šabanović, S., Chang, W.(2018). engaging older adults with depression as co-designers of assistive in-home robots. In: Proceedings of the 12th EAI international conference on pervasive computing technologies for healthcare, pp. 304– 309. ACM, New York, NY, USA . https://doi.org/10.1145/3240925.3240946

Randall, N., Joshi, S., Kamino, W., Hsu, L.-J., Agnihotri, A., Li, G., Williamson, D., Tsui, K., & Šabanović, S. (2022). Finding Ikigai: How robots can support meaning in later life. *Frontiers in Robotics and AI*. https://doi.org/10.3389/frobt.2022.1011327

Razdaibiedina, A., Mao, Y., Hou, R., Khabsa, M., Lewis, M., Almahairi, A. (2023) Progressive prompts: Continual learning for language models. Preprint at arXiv: org/abs/2301.12314

Reddy, R.G., Chandra, S., Singh Sidhu, M., Bai, H.J., Yao, W., Pillai, P., Aggarwal, K., Ren, L., Sonawane, P., Han, K., Goyal, V., Agrawal, S., Zhai, C.(2023). Charmbana: Progressive responses with real-time internet search for knowledge-powered conversations. In: Alexa Prize SocialBot grand challenge 5 proceedings . https://www.amazon.science/alexa-prize/proceedings/harmbana-progressive-responses-with-real-time-internet-search-for-knowledge-powered-conversations

Rehm, M., Krummheuer, A. L., Rodil, K., Nguyen, M., & Thorlacius, B. (2016). From social practices to social robots - user-driven robot development in elder care. In A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, & H. He (Eds.), *Social Robotics* (pp. 692–701). Cham: Springer.

Reimann, M.M., Kunneman, F.A., Oertel, C., Hindriks, K.V. (2023). A survey on dialogue management in human–robot interaction. Preprint at arXiv: org/abs/2307.10897

Riek, L. D. (2012). Wizard of Oz studies in HRI: A systematic review and new reporting guidelines. *Journal of Human and Robot Interact, 1*(1), 119–136. https://doi.org/10.5898/JHRI.1.1.Riek

Rogers, W. A., Kadylak, T., & Bayles, M. A. (2022). Maximizing the benefits of participatory design for human-robot interaction research with older adults. *Human Factors, 64*(3), 441–450. https://doi.org/10.1177/00187208211037465

Roller, S., Boureau, Y.-L., Weston, J., Bordes, A., Dinan, E., Fan, A., Gunning, D., Ju, D., Li, M., Poff, S., Ringshia, P., Shuster, K., Smith, E.M., Szlam, A., Urbanek, J., Williamson, M. (2020). Open-domain conversational agents: Current progress, open problems, and future directions. Preprint at arXiv: org/abs/2006.12442

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E.M., et al. (2020). Recipes for building an open-domain chatbot. Preprint at arXiv: org/abs/2004.13637

Šabanović, S., Chang, W.-L., Bennett, C.C., Piatt, J.A., Hakken, D.(2015). A robot of my own: Participatory Design of socially assistive robots for independently living older adults diagnosed with depression. In: Zhou, J., Salvendy, G. (eds.) Human aspects of IT for the aged population. Design for Aging vol. 9193, pp. 104– 114. Springer, Cham . https://doi.org/10.1007/978-3-319-20892-3_11

Šabanović, S. (2010). Robots in society, society in robots. *International Journal of Social Robotics, 2*(4), 439–450. https://doi.org/10.1007/s12369-010-0066-7

Sayago, S., Neves, B.B., Cowan, B.R.(2019). Voice assistants and older people: some open issues. In: Proceedings of the 1st international conference on conversational user interfaces - CUI '19, pp. 1– 3. ACM Press, Dublin, Ireland . https://doi.org/10.1145/3342775.3342803

Scialom, T., Chakrabarty, T., Muresan, S.(2022). Fine-tuned language models are continual learners. In: Proceedings of the 2022 conference on empirical methods in natural language processing, pp. 6107– 6122. Association for computational linguistics, Abu Dhabi, United Arab Emirates . https://doi.org/10.18653/v1/2022.emnlp-main.410

Segaert, K.(2020). In: Zeigler-Hill, V., Shackelford, T.K. (eds.) Priming Effects, pp. 4027– 4030. Springer, Cham . https://doi.org/10.1007/978-3-319-24612-3_479

Senaratna, H.A.S.D., Manawadu, U.A., Hansika, W.K.N., Samarasinghe, S.W.A.M.D., De Silva, P.R.S.(2020). Mucor: A multiparty conversation based robotic interface to evaluate job applicants. In: Stephanidis, C., Kurosu, M., Degen, H., Reinerman-Jones, L. (eds.) HCI International 2020—late breaking papers: Multimodality and intelligence, pp. 280– 293. Springer, Cham

Shahverdi, P., Tyshka, A., Trombly, M., Louie, W.-Y.G.(2022). Learning turn-taking behavior from human demonstrations for social human-robot interactions. In: 2022 IEEE/RSJ International conference on intelligent robots and systems (IROS), Kyoto, Japan, pp. 7643– 7649 .https://doi.org/10.1109/IROS47612.2022.9981243

Shen, Y., Qi, J., Wang, S., Yao, B.M., Liu, M., Xu, Z., Ashby, T., Huang, L.(2023). Hokiebot: Towards personalized open-domain chatbot with long-term dialogue management and customizable automatic evaluation. In: Alexa Prize SocialBot Grand challenge 5 proceedings . https://www.amazon.science/alexa-prize/proceedings/hokiebot-towards-personalized-open-domain-chatbot-with-long-term-dialogue-management-and-customizable-automatic-evaluation

Shervedani, A.M., Oh, K.-H., Abbasi, B., Monaikul, N., Rysbek, Z., Eugenio, B.D., Zefran, M.(2022). Evaluating multimodal interaction of robots assisting older adults. Preprint at arXiv:2212.10425

Shi, W., Li, Y., Sahay, S., Yu, Z.(2022). Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. Preprint at arXiv: org/abs/2012.15375

Shuster, K., Xu, J., Komeili, M., Ju, D., Smith, E.M., Roller, S., Ung, M., Chen, M., Arora, K., Lane, J., Behrooz, M., Ngan, W., Poff, S., Goyal, N., Szlam, A., Boureau, Y.-L., Kambadur, M., Weston, J. (2022) BlenderBot 3: A deployed conversational agent that continually learns to responsibly engage. Preprint at arXiv: org/abs/2208.03188

Sidnell, J., & Stivers, T. (2012). *The handbook of conversation analysis*. Wiley-Blackwell.

Skantze, G., Doğruöz, A.S.(2023). The Open-domain Paradox for chatbots: Common ground as the basis for human-like dialogue. Preprint at arXiv: org/abs/2303.11708

Skantze, G., & Irfan, B. (2025). Applying general turn-taking models to conversational human-robot interaction. In *2025 ACM/IEEE international conference on human-robot interaction*. IEEE.

Skantze, G., Johansson, M., Beskow, J.(2015).Exploring turn-taking cues in multi-party human-robot discussions about objects. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. ICMI '15, pp. 67– 74. Association for computing machinery, New York, NY, USA . https://doi.org/10.1145/2818346.2820749

Skantze, G. (2021). Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language, 67*, 101178. https://doi.org/10.1016/j.csl.2020.101178

Smedegaard CV (2019) Reframing the role of novelty within social HRI: From noise to information. In: 2019 14th ACM/IEEE International conference on human–robot interaction (HRI), Daegu, South Korea, pp. 411– 420 . https://doi.org/10.1109/HRI.2019.8673219

Søraa, R. A., Tøndel, G., Kharas, M. W., & Serrano, J. A. (2022). What do older adults want from social robots? A qualitative research approach to human–robot interaction (HRI) studies. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-022-00914-w

Stegner, L., Senft, E., Mutlu, B.(2023). Situated participatory design: A method for in situ design of robotic interaction with older adults. In: Proceedings of the 2023 CHI conference on human factors in computing systems. CHI '23. Association for computing machinery, New York, NY, USA . https://doi.org/10.1145/3544548.3580893

Syrdal, D.S., Dautenhahn, K., Koay, K., Walters, M.(2009). The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. In: Adaptive and emergent behaviour and complex systems, Edinburgh, United Kingdom

Tatarian, K., Stower, R., Rudaz, D., Chamoux, M., Kappas, A., & Chetouani, M. (2022). How does modality matter? investigating the synthesis and effects of multi-modal robot behavior on social intelligence. *International Journal of Social Robotics, 14*(4), 893–911.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. (2022). Lamda: Language models for dialog applications. Preprint at arXiv: org/abs/2201.08239

Thunberg, S., Thellman, S., Ziemke, T.(2017). Don't judge a book by its cover: A study of the social acceptance of NAO vs. pepper. In: Proceedings of the 5th international conference on human agent interaction. HAI '17, pp. 443– 446. Association for computing machinery, New York, NY, USA . https://doi.org/10.1145/3125739.3132583

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G. (2023). LLaMA: Open and efficient foundation language models. Preprint at arXiv: org/abs/2302.13971

Ung, M., Xu, J., Boureau, Y.-L.(2022). SaFeRDialogues: Taking feedback gracefully after conversational safety failures. In: Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp. 6462–6481. Association for computational linguistics, Dublin, Ireland . https://doi.org/10.18653/v1/2022.acl-long.447 . https://aclanthology.org/2022.acl-long.447

Vinyals, O., Le, Q.V.(2015). A neural conversational model. In: ICML deep learning workshop 2015, vol. 37 . Preprint at arXiv: org/abs/1506.05869

Wang, Y., Jiang, Z., Chen, Z., Yang, F., Zhou, Y., Cho, E., Fan, X., Huang, X., Lu, Y., Yang, Y. (2023) RecMind: Large language model powered agent for recommendation. Preprint at arXiv: org/abs/2308.14296

Wang, L., Lim, E.-P. (2023). Zero-shot next-item recommendation using large pretrained language models. Preprint at arXiv: org/abs/2304.03153

Wang, W., Lin, X., Feng, F., He, X., Chua, T.-S. (2023). Generative recommendation: Towards Next-generation recommender paradigm. Preprint at arXiv: org/abs/2304.03516

Wang, H., Wang, W., Saini, R., Zhukova, M., Yan, X.(2023). Gauchochat: Towards proactive, controllable, and personalized social conversation. In: Alexa Prize SocialBot Grand Challenge 5 Proceedings . https://www.amazon.science/alexa-prize/proceedings/gauchochat-towards-proactive-controllable-and-personalized-social-conversation

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., Anandkumar, A.(2023). Voyager: An open-ended embodied agent with large language models. Preprint at arXiv: org/abs/2305.16291

Wei, J., Kim, S., Jung, H., Kim, Y.-H.(2023) Leveraging large language models to power chatbots for collecting user self-reported data. Preprint at arXiv: org/abs/2301.05843

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. Preprint at arXiv: org/abs/2201.11903

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L.A., Isaac, W., Legassick, S., Irving, G., Gabriel, I.(2021) Ethical and social risks of harm from language models. arXiv. Preprint at arXiv: org/abs/2112.04359

Weiss, A., Bernhaupt, R., Tscheligi, M., Yoshida, E.(2009). Addressing user experience and societal impact in a user study with a humanoid robot. In: Adaptive and emergent behaviour and complex systems - proceedings of the 23rd convention of the society for the study of artificial intelligence and simulation of behaviour, AISB 2009, pp. 150– 157

Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., Weston, J. (2019). Neural text generation with unlikelihood training. Preprint at arXiv: org/abs/1908.04319

Werner, L., Huang, G., & Pitts, B. J. (2019). Automated speech recognition systems and older adults: A literature review and synthesis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63*(1), 42–46. https://doi.org/10.1177/1071181319631121

Whelan, S., Murphy, K., Barrett, E., Krusche, C., Santorelli, A., & Casey, D. (2018). Factors affecting the acceptability of social robots by older adults including people with dementia or cognitive impairment: A literature review. *International Journal of Social Robotics, 10*(5), 643–668. https://doi.org/10.1007/s12369-018-0471-x

Williams, T., Matuszek, C., Mead, R., & Depalma, N. (2023). Scarecrows in oz: The use of large language models in HRI. *ACM Transactions on Human-Robot Interaction*. https://doi.org/10.1145/3606261

Winkle, K., Caleb-Solly, P., Turton, A., Bremner, P.(2018). Social robots for engagement in rehabilitative therapies: Design implications from a study with therapists. In: 2018 13th ACM/IEEE international conference on human–robot interaction (HRI), pp. 289– 297

Wu, Z., Galley, M., Brockett, C., Zhang, Y., Gao, X., Quirk, C., Koncel-Kedziorski, R., Gao, J., Hajishirzi, H., Ostendorf, M., & Dolan, B. (2021). A controllable model of grounded response generation. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(16), 14085–14093. https://doi.org/10.1609/aaai.v35i16.17658

Xiang, J., Tao, T., Gu, Y., Shu, T., Wang, Z., Yang, Z., Hu, Z. (2023). Language models meet world models: Embodied experiences enhance language models. Preprint at arXiv: org/abs/2305.10626

Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M. (2023). Efficient streaming language models with attention sinks. Preprint at arXiv: org/abs/2309.17453

Xiao, Y., Wang, W.Y.(2021). On hallucination and predictive uncertainty in conditional language generation. In: Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main Volume, pp. 2734– 2744. Association for Computational Linguistics, Online https://doi.org/10.18653/v1/2021.eacl-main.236

Xu, J., Ju, D., Li, M., Boureau, Y.-L., Weston, J., Dinan, E. (2021). Recipes for safety in open-domain chatbots. Preprint at arXiv: org/abs/2010.07079

Xu, J., Szlam, A., Weston, J.(2022). Beyond goldfish memory: Long-term open-domain conversation. In: Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp. 5180– 5197. Association for computational linguistics, Dublin, Ireland . https://doi.org/10.18653/v1/2022.acl-long.356 . https://aclanthology.org/2022.acl-long.356

Xu, J., Ung, M., Komeili, M., Arora, K., Boureau, Y.-L., Weston, J.(2022). Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. Preprint at arXiv: org/abs/2208.03270

Yamazaki, T., Yoshikawa, K., Kawamoto, T., Mizumoto, T., Ohagi, M., & Sato, T. (2023). Building a hospitable and reliable dialogue system for android robots: A scenario-based approach with large language models. *Advanced Robotics*. https://doi.org/10.1080/01691864.2023.2244554

Yang, J., Wang, P., Zhu, Y., Feng, M., Chen, M., He, X. (2022). Gated multimodal fusion with contrastive learning for turn-taking prediction in human-robot dialogue. In: ICASSP 2022–2022 IEEE International conference on acoustics, speech and signal processing (ICASSP), Singapore, Singapore, pp. 7747– 7751 https://doi.org/10.1109/ICASSP43922.2022.9747056

Yao, Z., Li, C., Wu, X., Youn, S., He, Y. (2023). A comprehensive study on post-training quantization for large language models. Preprint at arXiv: org/abs/2303.08302

Yue, L., Liu, Q., Du, Y., Gao, W., Liu, Y., Yao, F. (2023). FedJudge: Federated legal large language model. Preprint at arXiv: org/abs/2309.08173

Zhang, M., Choi, E.(2021). SituatedQA: Incorporating extra-linguistic contexts into QA. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp. 7371– 7387. Association for computational linguistics, Online and Punta Cana, Dominican Republic . https://doi.org/10.18653/v1/2021.emnlp-main.586

Zhang, D., Finckenberg-Broman, P., Hoang, T., Pan, S., Xing, Z., Staples, M., Xu, X. (2023). Right to be forgotten in the era of large language models: Implications, challenges, and solutions. Preprint at arXiv: org/abs/2307.03941

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L.(2022) OPT: Open pre-trained transformer language models. Preprint at arXiv: org/abs/2205.01068

Zhang, B., Soh, H. (2023). Large language models as zero-shot human models for human–robot interaction. Preprint at arXiv: org/abs/2303.03548

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? ACL 2018–56th annual meeting of the association for computational linguistics. *Proceedings of the Conference, 1*, 2204–2213. https://doi.org/10.18653/v1/p18-1205. 1801.07243.

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., Wen, J.-R.(2023). A survey of large language models. Preprint at arXiv: org/abs/2303.18223

Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G(2020). Fine-tuning language models from human preferences. Preprint at arXiv: org/abs/1909.08593

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

Bahar Irfan is a Postdoctoral Researcher and Digital Futures fellow at KTH Royal Institute of Technology, Sweden. Her research focuses on creating personal robots that continually learn and adapt to assist in daily life. Her research interests include human-robot interaction, lifelong learning, conversational AI, and large language models. Prior to joining KTH, she held research positions at Disney Research Los Angeles and Evinoks Service Equipment Industry and Commerce Inc. She received her PhD (2020) in Robotics from the University of Plymouth (UK) and SoftBank Robotics Europe (France) as a Marie Skłodowska-Curie Actions fellow. She has an MSc (2016) in Computer Engineering and a BSc (2012) in Mechanical Engineering from Boğaziçi University, Turkey. Her research was published in several journals and conferences in robotics, receiving over 600 citations. Contact her at birfan@kth.se. Website: https://baharirfan.com.

Sanna Kuoppamäki is an Assistant Professor at the Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Sweden. Her research explores the use and design of interactive technologies for healthy ageing from the age and life course perspective, with application areas in social robotics, welfare technology and mobile health. She received her PhD in sociology from the University of Jyväskylä, Finland. She is a member of the Socio-Gerontechnology network. Contact her at sannaku@kth.se.

Aida Hosseini is an M.Sc. student at the Division of Health Informatics and Logistics at KTH Royal Institute of Technology, Sweden. Her research focuses on developing technical and business solutions for healthy living and active ageing. She previously held positions at A+ Science AB, Clinical Trial Consultants AB, and Karolinska University Hospital. She received her BSc (2022) in Biomedical Engineering and Health Systems from KTH Royal Institute of Technology, Sweden. Contact her at idaho@kth.se.

Gabriel Skantze is a Professor in Speech Technology and Communication at KTH Royal Institute of Technology, Sweden. He is leading several research projects related to conversational systems and human-robot interaction, investigating and modeling phenomena such as turn-taking, visual grounding, and multimodal feedback in dialogue. He is also co-founder and chief scientist at Furhat Robotics and President Emeritus of SIGDIAL, the ACL Special Interest Group on Discourse and Dialogue. Contact him at skantze@kth.se.