applied sciences

MDPI

Article

Personalization and Generative Dialogue in Social Robotics for Eldercare: A User Study

Luca Pozzi [1,2], Marco Nasato [1,3], Nicola Toscani [2], Francesco Braghin [1,2], Marta Gandolla [1,2,*]

[1] WE-COBOT Lab, Department of Mechanical Engineering, Politecnico di Milano, Polo Territoriale di Lecco, 23900 Lecco, Italy
[2] Department of Mechanical Engineering, Politecnico di Milano, 20156 Milano, Italy
[3] Department of Electronics, Information, and Bioengineering, Politecnico di Milano, 20133 Milano, Italy
* Correspondence: marta.gandolla@polimi.it

Abstract

Service robots have the potential to support cognitive and social well-being in long-term care facilities, yet their widespread adoption depends on intuitive interaction modalities that minimize user learning effort and the need for a technical expert on-ground. Spoken dialogue is a natural interface, and recent advances in large language models (LLMs) promise more flexible and engaging exchanges than traditional scripted systems. In this study, we implemented a modular speech-based architecture combining automatic speech recognition, text-to-speech synthesis, and a conversational agent capable of switching between a fully scripted and LLM-driven dialogue. The implemented architecture was embodied in a TIAGo robot (PAL Robotics) and tested to compare three conversational strategies: (1) scripted, predefined dialogue, (2) LLM-based free-form conversation, and (3) LLM-based conversation augmented with personal information provided through the prompt. Eighteen younger adults and eighteen older adults engaged in a five-minute interaction with the robot under all three conditions in a within-subject design, and subsequently completed the Almere model questionnaire. Across all subscales and both participant groups, differences between dialogue strategies were small and statistically non-significant, despite informal comments from several older participants indicating a perceived increase in intelligence or naturalness for the LLM conditions. The findings suggest that generative dialogue and basic personalization alone do not meaningfully shift perceived acceptance in brief, task-neutral encounters, underscoring the importance of longer-term deployment and functionally meaningful robot roles in future evaluations.

Keywords: human–robot interaction; service robotics; large language models; older adults

Academic Editors: Elisa Digo, Stefano Pastorelli and Valerio Cornagliotto

Received: 10 March 2026
Revised: 25 March 2026
Accepted: 27 March 2026
Published: 31 March 2026

1. Introduction

Population aging is becoming increasingly significant, particularly in developed countries. Life expectancy has risen steadily over recent decades and is expected to continue increasing through 2050 [1]. According to the World Health Organization, the number of people aged 60 years and older is expected to double by 2050, reaching 2.1 billion. In parallel, those aged 80 years and older are projected to triple [1]. Although aging does not represent a pathological condition per se, it is often accompanied by challenges. The World Health Organization highlights common age-related health issues such as hearing loss, vision impairment, chronic pain, osteoarthritis, diabetes, chronic obstructive pulmonary disease, depression, and dementia [1]. One way to mitigate the impact of these challenges

is by fostering social interaction and mental stimulation. These factors have been shown to improve the well-being of older adults, including individuals experiencing depression, cognitive decline, or loneliness, particularly in residential care settings [2,3]. Positive physical and social environments support aging in good health, enabling individuals to continue engaging in meaningful activities despite functional losses [2,3].

Despite this need, the growing older population is not matched by a corresponding increase in the availability of caregivers. Chronic understaffing in long-term care facilities, such as nursing homes, limits opportunities for caregivers to provide emotional or cognitive engagement, as immediate medical and physical needs must take priority. As a result, many residents receive limited access to social interaction and stimulating activities [4].

To address these gaps, new strategies are being explored, including the introduction of technological solutions in care environments [5]. Service and social robots have the potential to complement caregivers, reduce workload, and provide opportunities for physical, cognitive, and social engagement. Robots have already been deployed experimentally in nursing homes and private homes to support companionship [6,7], leisure activities [8,9], and cognitive stimulation [10,11]. Such interventions may offer an alternative means of mitigating loneliness, supporting emotional well-being, and promoting active lifestyles.

However, despite the existing technology and the increasing demand, these systems are still far from routine in everyday care environments. Barriers include economic constraints, logistical complexity, and ethical concerns [12]. Moreover, insight gained from interviews with Italian nursing home managers highlights an additional obstacle related to accessibility of interaction [13]. Managers expressed skepticism about the feasibility of fully autonomous robot-mediated engagement, noting that only a subset of residents might be capable of interacting with a robot independently. For residents with diminished cognitive capabilities or reduced familiarity with technology, interaction would require continuous mediation by staff, defeating the purpose of deployment [13]. In such scenarios, robots risk becoming an additional burden, demanding staff attention and training, rather than a resource.

Indeed, interaction with robotic systems is not trivial, particularly when the users are non-tech-savvy, frail, or may experience some degree of cognitive decline. One of the most common and reliable means of human–robot interaction (HRI) involves touchscreens, as well-designed graphical interfaces require little to no learning for users experienced with tablets, smartphones, or similar devices [5,14]. However, this modality is far less intuitive for people unfamiliar with such interfaces and raises additional accessibility demands, including considerations for visual impairment and motor limitations. Similarly, physical buttons or dedicated controllers share many of these drawbacks and often restrict the complexity and expressiveness of interaction [14]. In this context, verbal communication represents a valuable alternative. Speech is a natural and instinctive way of seeking assistance, even from systems that we do not consciously believe can "listen". In moments of frustration, it is a common experience to beg a smartphone or computer to work. Enabling such spontaneous, low-effort communication would substantially broaden the population of potential robot users.

Recent advancements in artificial intelligence models have enabled increasingly flexible, unstructured interaction with machines. Applications based on large language models (LLMs) are now widely available and enable people to access information, help, and services without requiring prior technical training. Generative dialogue systems also opened new possibilities for the interaction personalization. Traditional dialogue systems require predefined user models and ad hoc rules to tailor interaction, limiting scalability and adaptability. In contrast, LLMs can incorporate user attributes directly into prompts or context windows, allowing for a conversation that reflects a person's preferences, background, or interests with minimal additional programming. Such personalization may support rapport building, enhance engagement, and foster technology acceptance [15].

Unsurprisingly, robotics, and service robotics in particular, has started to adopt these models. LLMs have been explored for context awareness [16,17], task planning [18,19], and many other functions [20]. They are also used to implement dialogue systems embodied in robots, often under the assumption that their conversational capabilities exceed those of scripted or rule-based chatbots [20]. However, while these implementations are typically validated for feasibility or task performance, few studies explicitly compare LLM-based dialogue with more conventional systems. Importantly, moving toward generative conversation introduces risks alongside benefits as LLMs may hallucinate, provide misleading information, or produce content inappropriate for a user's abilities or context, e.g., suggesting outdoor activities to a resident with mobility limitations. Thus, although the flexibility and programmability of LLM-driven dialogue are clear from a development perspective, their impact on user perception remains poorly investigated. Systematic comparison of LLM-based interactions with more traditional approaches could help to clarify what benefits—if any—users experience, and thus inform the decision on whether to take the additional risks that come with generative agents.

1.1. Related Works

Despite the growing prevalence of LLM applications for conversational agents in HRI, relatively few studies have systematically investigated their impact on user experience. Schlesener and colleagues [21] investigated the impact of anthropomorphism (in behavior and embodiment) on acceptance in a virtual agent. Results showed that higher anthropomorphism increased acceptance, supporting LLM-based implementations and human-like embodiments, although the latter risked falling into the uncanny valley. In their work, Lo and co-workers [22] developed a GPT-4-based framework to simulate personality, capable of processing text and image inputs to generate responses, actions, and emotions. The framework, tested on a Mobi service robot, incorporated user preferences, personality models, long-term memory, and theory of mind (i.e., the ability to infer others' mental states and adjust behavior accordingly). LLM-based dialogue systems specifically designed for older adults have also been proposed. Pinto-Bernal and her group [23] integrated an LLM, with automatic speech recognition (ASR), and text-to-speech (TTS) to enable vocal interaction. The dialogue system, capable of memory retention across multiple interactions, was embodied in a Pepper robot. Technical evaluations highlighted issues such as interruptions and occasional disclosures of the LLM's nature, but older adults generally found the robot engaging, with some expressing interest in it as a companion. Similarly, Irfan et al. [24] explored challenges in LLM-based companions for older adults, including managing latency, avoiding repetitive or inappropriate responses, personalizing interactions, and monitoring user engagement. Finally, Khoo et al. [25] tested a fine-tuned GPT-3 on a personalized QT robot with seven older adults. While most participants reported positive, enjoyable interactions, a few deemed the system to be more suitable for older adults living alone or with dementia, and some noted slow responses or limited personal relevance.

1.2. Paper Contribution

In this framework, this paper contributes to the current state of the art by introducing and evaluating a speech-based interaction architecture embodied in a service robot. The system allows seamless switching between different conversational agent back-ends, enabling a direct comparison of interaction modes. To this end, the impact of different dialogue strategies on user experience is studied by engaging two distinct age groups (young adults, i.e., less than 30 years old, and older adults, i.e., 70 and older) in three five-minute chats with the robot. Each interaction relies on a different conversational logic, namely a scripted dialogue, an LLM-driven dialogue, and an LLM-driven dialogue in which the agent is

provided with user-profiling information. By analyzing users' perceptions across these conditions, the study aims at providing early insights on whether, and to what extent, generative and personalized conversational agents influence acceptance and experience in spoken HRI. In summary, the main contributions of this research are:

- A comparison between scripted and agent-based vocal interaction on a service robot;
- An evaluation of the feasibility of off-the-shelf personalization of an agentic conversational agent through user profiling information;
- An analysis of user acceptance in HRI across younger and older adults.

2. Materials and Methods

2.1. Robot Hardware

In this study, the conversational agent is embodied on a TIAGo robot (PAL Robotics), shown in Figure 1.

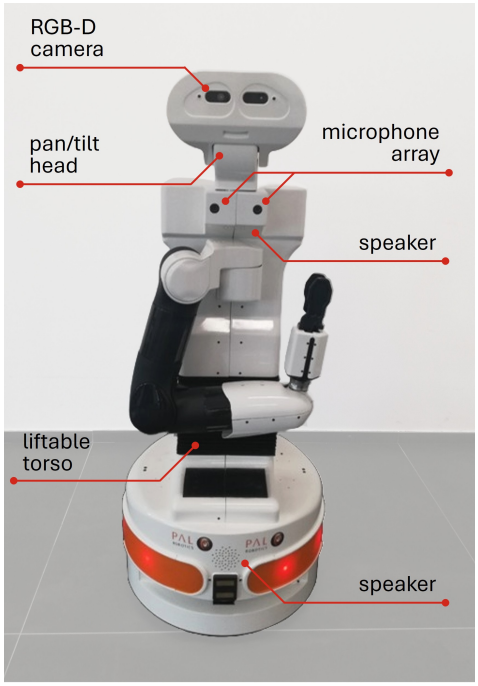

Figure 1. The TIAGo robot (PAL Robotics) used in the experimental evaluation as the embodiment of the proposed HRI framework. Robot features and peripherals that are relevant for either the interaction or the validation procedure are highlighted in the image.

TIAGo is a mobile manipulator designed as a versatile research platform. As such, in addition to a mobile base and a 7-degree-of-freedom arm, the robot is equipped with hardware supporting HRI, including an SuperBeam^{TM} (Andrea Electronics, Bohemia, NY, USA) stereo array microphone, integrated in the robot's upper torso, and speakers located in both the upper torso and mobile base. The two joints controlling the head orientation enable the robot to orient the embedded camera toward the user. Together with the liftable torso, these joints are exploited to perform random, non-functional movements during the interaction, to enhance the robot animacy.

2.2. Human–Robot Interaction Framework

The HRI software framework developed in this work is schematically represented in Figure 2.

The interaction pipeline includes an ASR module that converts the user's vocal input, captured through the microphone array, into a textual string provided to the conversational

agent. The output from the conversational agent, either generative or predefined depending on the selected operation modality, is fed into the TTS module. The TTS converts the string into an audio stream reproduced through the robot's speakers. Individual components of the system are described in the following Sections 2.2.1–2.2.3. The HRI framework is implemented in ROS, with nodes written in Python 3.8. As depicted in Figure 2, nodes responsible for audio acquisition and conversion to ROS-compatible formats are deployed on the robot's onboard computer. The remaining processing, including conversational logic and language model inference, is executed on an external desktop computer equipped with an Intel® Core™ i7 CPU, 16GB RAM, and an NVIDIA GeForce GTX 4GB GPU.

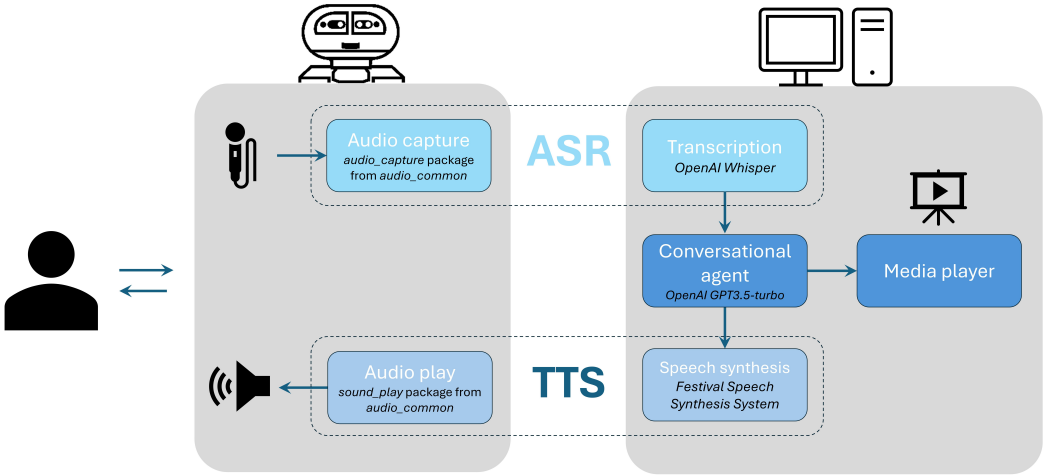

Figure 2. Schematic representation of the HRI framework. The ROS nodes for sound acquisition and playback are deployed onto the robot internal computer. The core of the HRI processing (i.e., transcription, response generation, and speech synthesis) is executed on an external computer connected via Ethernet to the robot. Also, the media player application, for the entertainment activity that concludes the interaction, is executed on the external computer. In each block, the relevant software package, module or API used in the implementation is mentioned. For the conversational agent, only the model for the LLM and pLLM modality (i.e., OpenAI GPT3.5-turbo) is reported, as the implementation of the script modality is based on a simple text file.

2.2.1. Automatic Speech Recognition

The ASR pipeline consists of two main stages. The first stage performs audio acquisition, chunking, and conversion into ROS messages using the `audio_capture` node from the ROS `audio_common` (https://github.com/ros-drivers/audio_common, accessed on 19 March 2024) package. This node runs on the robot's onboard computer, records audio from the microphone array, and broadcasts it to the ROS network. The second stage performs automatic transcription of the acquired audio into text. For transcription, the OpenAI Whisper model (multilingual `small` model, provided via https://github.com/openai/whisper, accessed on 19 April 2024) [26] was employed in a streaming implementation (https://github.com/TIAGo-WE-COBOT/whisper_streaming, forked from https://github.com/ufal/whisper_streaming with minor modifications, accessed on 19 April 2024), i.e., enabling incremental processing of incoming audio and the generation of partial transcriptions without requiring the user to complete an utterance. The core of the ASR node is an update loop that receives audio chunks and appends them to an internal buffer. Optionally, incoming chunks can be filtered by a voice activity detection (VAD) module to discard segments classified as non-speech, resulting in a cleaner buffer for transcription. Whisper is trained to process audio segments of up to approximately 30 s containing a full sentence [26]. To maintain contextual consistency across consecutive segments, despite the streaming approach, the transcription function is provided with the

last 200 words of confirmed previous output. This mechanism promotes coherence in style, terminology, and inter-sentence references.

Several practical issues were observed during integration. Occasionally, silence segments triggered the generation of short, predefined utterances, likely due to biases in the Italian training data. This behavior was mitigated through a blacklist of fixed transcriptions, i.e., discarding output exactly matching a blacklisted string. In addition, short pauses in user speech were sometimes interpreted as utterance termination, leading to incomplete transcriptions and premature robot responses, possibly interrupting the user speech. To reduce this effect, partial transcriptions were concatenated until two consecutive empty outputs (i.e., silence) were received. Only then the accumulated text was forwarded to the conversational agent. While this strategy reduced premature interruptions, it introduced an increase in response latency.

2.2.2. Text-to-Speech

Analogously to the ASR, the TTS pipeline is divided into two components: the synthesis stage executed on the desktop computer, and a playback stage handling the audio streaming through the robot peripherals, which runs on the robot's onboard computer. The robot's speech output is produced using the `sound_play` package, part of the ROS `audio_common` (https://github.com/ros-drivers/audio_common, accessed on 19 March 2024) metapackage Utterances are synthesized through the Festival speech synthesis system [27] and transmitted to the robot speakers for playback. Although the experimental evaluation reported in this work is conducted exclusively in Italian (see Section 2.3), Festival supports operation in several languages, provided that the corresponding language models are downloaded.

2.2.3. Conversational Agent

The conversational agent module is the core of the dialogue framework (https://github.com/TIAGo-WE-COBOT/hri_conversational_agency.git, branch `v1`, accessed on 2 May 2024). The interaction is conceived as a structured dialogue with alternating turns between the embodied agent (i.e., the robot) and the user. In each conversation loop, the user's spoken input is processed by the ASR module and converted into text. Upon receiving a textual input, the conversational agent processes it according to its internal logic and generates an appropriate textual response. This response is then published to the TTS module, enabling the robot to deliver a vocal response. The modular architecture of the framework, as well as the design of the conversational agent, allows one to seamlessly switch the control logic. For the purpose of the present work, three conversational agents are implemented:

- **Script-Based Dialogue (*SBD*).** The robot follows a predefined dialogue script. Upon receiving a user input, the conversational agent simply outputs the next predefined utterance. Although no semantic understanding of the user input is performed in this agent modality, careful script design helps to mask this limitation. Indeed, each robot utterance begins with a generic acknowledgment (e.g., *"Oh, cool"*, *"I see"*, *"Yeah, I understand"*), and ends with a question to encourage the user engagement. In most conversational flows, this strategy maintains the impression that the robot is responding to the user's content. The full script is reported in the Supplementary Materials.
- **Generic LLM (*gLLM*).** The LLM-based agent is implemented based on OpenAI's GPT3.5-turbo model [28], accessed through the chat-completion API. The temperature parameter is set to 0.8 (out of a [0–2] range) to introduce moderate variability in the responses while limiting excessive randomness and potential hallucinations.

The system message is defined according to the *persona pattern* prompting technique, which sets the style or perspective that the assistant should adopt when generating responses [29]. According to this structure, the role of the assistant is defined. In addition, the prompt informs the model that user inputs originate from an ASR system and may contain recognition errors, and instructs the agent to ask for repetition in case of unclear input. A maximum word constraint is also included to limit response length. All these elements were combined in the prompt below:

gLLM prompt

You are a service robot named TIAGo. Your task is to engage in a conversation entirely in `[language]` to entertain a person. If no question is asked, initiate one yourself. The person's input comes from a speech transcription system and may not be accurate. If you don't understand the input, apologize and politely ask for it to be repeated. Your response should be under `[max_words_num]` words.

Anticipating the content of Section 2.3, the `language` was set to Italian for all the participants of the study, and the `max_words_num` to 500 words. In order to maintain a coherent dialogue flow, it is important to keep the agent informed of the conversation that has happened so far. Given the limited temporal extent of the target conversation, the agent memory is obtained by simply feeding to the agent the full conversation log.

- **Personalized LLM-based (*pLLM*).** The pLLM modality extends the gLLM implementation by augmenting the same base prompt with additional information describing the user. In addition to basic demographic information (i.e., gender and age range), the user profile includes information about the educational and professional background, collected through short open-ended questions regarding the highest level of education attained and the most relevant occupations of the job career. Furthermore, the profile includes the outcome of a personality assessment based on the Big Five Inventory (BFI), which describes the personality of a subject along five dimensions, called traits: *extraversion*, *agreeableness*, *conscientiousness*, *neuroticism*, *openness*.

pLLM prompt

You are a service robot named TIAGo. Your task is to engage in a conversation entirely in `[language]` to entertain a person described in the user profile below. If no question is asked, initiate one yourself. The person's input comes from a speech transcription system and may not be accurate. If you don't understand the input, apologize and politely ask for it to be repeated. Your response should be under `[max_words_num]` words.
User profile:

- Gender: `[gender]`
- Age range: `[age]`
- Background:
  - Education level: `[education]`
  - Working experience: `[work]`
- Interests:
  - `[interest_1]`
  - `[interest_2]`
  - `[interest_3]`
- Personality (accoding to Big Five Inventory)
  - Extraversion: `[bfi_extra]`
  - Agreeableness: `[bfi_agree]`
  - Conscientiousness: `[bfi_consc]`
  - Neuroticism: `[bfi_neuro]`
  - Openness: `[bfi_open]`

In all agent modalities, the open-domain conversation is concluded by the agent after approximately 5 min in order to proceed with a short entertainment activity. In the generative modalities (i.e., gLLM and pLLM), the conversation duration is monitored by a timer. When the timer elapses, the agent concludes the dialogue on its next turn and invites the user to select a media content (movie clip, song, or audiobook excerpt) from a three-item list. Similarly, in the SBD condition, the same choice is proposed after the user answers the final question in the script. This functionality is intended to provide a simple example of the type of entertainment and stimulation services that the robot could provide in a real-world deployment. Finally, after the completion of the selected activity, the robot ends the interaction with a farewell.

2.3. Experimental Protocol

The setup described in Section 2.2 was tested at the Polo territoriale di Lecco-Politecnico di Milano within the Wearable and Collaborative Robotics (WE-COBOT) from 7th May to 15th May 2024, under the approval of the Ethical Committee of Politecnico di Milano (approval no. 46/2022).

The study involved 36 participants, divided into two age groups: older adults (O65; 7 M, 11 F; age 76.11 ± 3.41), and younger adults (U30; 9 M, 9 F; age 21.67 ± 1.88). Participants were recruited through a promotional video presenting the research team and the project, dissemination of a flyer with general project information, and word-of-mouth referrals.

The structure of the experimental protocol is shown in Figure 3. It consists of three main stages, namely intake and profiling, interaction with the embodied conversational agent, and user experience assessment. As depicted in Figure 3, the latter two stages are repeated three times, with both the conversational agent modality (i.e., SBD, gLLM, pLLM) and the entertainment activity (i.e., movie clip, audiobook excerpt, song) varying across repetitions and assigned in random order.

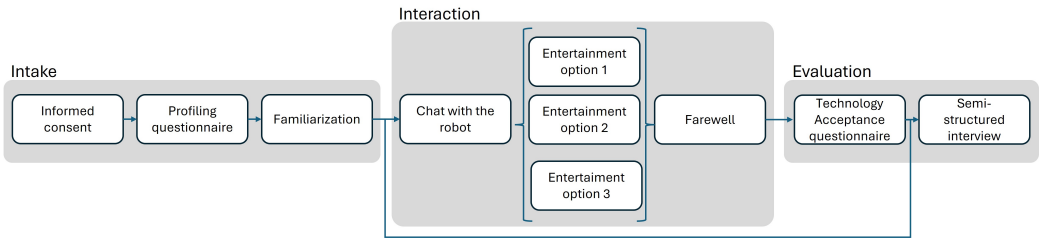

Figure 3. Schematic representation of the experimental session. After providing informed consent, participants complete the profiling questionnaire and engage in a brief familiarization interaction with the robot, to adjust the voice volume and experience the dialogue latencies. The main interaction is then repeated three times—each one under a different conversational agent condition—each time followed by a technology acceptance questionnaire based on the Almere model. After completion of the third interaction, a semi-structured interview is conducted to collect qualitative feedback on the user experience.

The experiments were conducted in a dedicated room divided by a partition into two separate areas. The partition allowed the participant to interact with the robot in isolation, while still enabling communication with the researchers if needed. The interaction area was arranged with a chair positioned in front of the robot for the participant to sit during the interaction, and a table holding the computer to play the multimedia content (visible in Figure 4, top-right panel). A camera was also placed on the table to record the interaction. In the control room, one or two researchers monitored the sessions and verified correct system operation.

Prior to the assessments, the experimental setting was prepared with all required equipment. Upon arrival, participants received a brief introduction to the general aims of the project and signed the consent. The participants were asked to fill out the profiling questionnaire in digital format, with assistance from a researcher if needed. Participants were then invited to sit in front of TIAGo and engage in a short familiarization interaction (i.e., one or two conversational turns, with the agent operating in SBD modality informed by a dedicated script), intended to adjust to the robot's voice volume and let the user experience the dialogue timing. Subsequently, each participant performed three independent trials. Each trial consisted of a five-minute conversation structured as an alternating exchange of turns between TIAGo and the user by using a different interaction modality of the ones described in Section 2.2.3 (i.e., SBD, gLLM, or pLLM). As reported in Section 2.2.3, at the end of each five-minute interaction, TIAGo proposed an entertainment activity, consisting on either watching a short movie clip, listening to a song, or hearing an excerpt from an audiobook. For each activity, the media to be played could be selected among three options. After playback of the multimedia content, TIAGo concluded the interaction with a farewell. The three interaction modalities were presented in randomized order across the three trials to prevent order effects. Similarly, the three entertainment activities were randomly ordered. As a result, each participant experienced all interaction modalities and all entertainment activities once, while the resulting modality–activity combinations varied across participants.

Figure 4. Samples from the interactions during the experimental campaign. User seated in front of the robot. A support PC was placed on a table, in the subject's view (see top-right panel) to play the entertainment content. Subjects are anonymized using a non-photorealistic sketch-based rendering, designed to suppress biometric identity while maintaining spatial configuration, posture, and interaction context.

After each interaction, participants completed a technology acceptance questionnaire referring to the interaction that had just taken place. Overall, each participant engaged in three conversations with TIAGo and completed three questionnaires, one per conversational modality.

Acceptance was assessed using an adaptation of the Almere model, a technology acceptance framework specifically developed for evaluating social assistive agents with older adults [30]. For this study, the Almere model was tailored to the experimental context. Specifically, the *social influence* and *use* constructs were excluded, as interaction with the robot was limited to a single experimental session rather than prolonged use. The remaining are thus *anxiety* (ANX), *attitude* (ATT), *facilitating conditions* (FC), *perceived adaptivity* (PAD), *perceived enjoyment* (PENJ), *perceived ease of use* (PEOU), *perceived sociability* (PS), *perceived*

usefulness (PU), *social presence* (SP), *trust* (TR), and *intention to use* (ITU). Said constructs were represented by two or three items each, resulting in a total of 25 items (the questionnaire is provided in the Supplementary Materials). Items were translated into Italian and presented in a randomized order, which was kept consistent across trials and participants. Each item was rated on a five-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*).

Questionnaire responses were converted to numerical scores according to the Almere scoring procedures. Internal consistency of Almere constructs was assessed in an exploratory manner by computing inter-item Pearson correlations. For constructs composed of three items, if the average inter-item correlation was below 0.3 [31], pairwise correlations were inspected, and the least correlated item was removed. After this refinement, responses were averaged within each construct to obtain a single score per participant, interaction modality, and construct. A mixed-design ANOVA was performed to analyze the within-subject (i.e., agent modality) and between-subject (i.e., age group) factors in terms of main effect and their interaction. When significant effects were observed, post hoc paired *t*-tests were conducted to further explore within-group differences, with correction for multiple comparisons. Finally, relationships between constructs were examined by computing Pearson's correlation coefficients in accordance with the Almere technology acceptance model.

3. Results

The analysis of internal consistency resulted in the item *"I find the robot easy to use"* to be removed from the perceived ease of use (PEOU) construct for further analysis. Even after the item removal, the internal construct consistency remained below-threshold (i.e., average item–item correlation lower than 0.3). Moreover, the anxiety (ANX) construct showed low internal consistency. As the construct included two items only, it was not possible to remove any items from it. The results about the PEOU and ANX constructs are thus reported for completeness, with the caveat that they should be interpreted with caution, due to their limited reliability. All remaining constructs met the internal consistency criteria.

After the consistency check, average scores were computed for each construct, separately for each interaction modality and age group. The results of the mixed-design ANOVA (shown in Table 1) revealed no significant main effects of age group ($F(1,34) = 0.77$, $p = 0.39$, $\eta^2 = 0.02$) or interaction modality ($F(2,68) = 1.50$, $p = 0.23$, $\eta^2 = 0.01$). However, a significant interaction between age group and agent modality was observed ($F(2,68) = 5.74$, $p = 0.005$, $\eta^2 = 0.04$), indicating that the effect of the interaction modality differed between the two age groups, although its effect size was *small-to-medium* according to the interpretation of Cohen [32].

Table 1. Mixed-design ANOVA results reporting F, df_1, df_2, p-values, and η^2 effect sizes for age group and agent modality main effects and interactions.

Factor	F	df_2	df_2	p	η^2
age group	0.772157	1	34	0.385719	0.015417
agent	1.497289	2	68	0.231019	0.011099
interaction	5.744578	2	68	0.004953	0.042585

Following these results, post hoc paired *t*-tests were performed to compare all pairwise combinations of interaction modalities within each age group and for each construct. To control for multiple comparisons, p-values were adjusted using a Bonferroni correction. After correction, none of the post hoc comparisons reached statistical significance, with median intention to use (ITU) being equal to the median value for both groups in each modality ($ITU_{median} = 3$), with an exception given for pLLM for the U30 group ($ITU_{median} = 4$). The

score distribution for each construct, age group, and agent modality is summarized in boxplots reported in Figure 5.

Relationships between constructs were examined by computing Pearson correlation coefficients in accordance with the structure of the Almere technology acceptance model. Correlations were computed separately for the two age groups. The resulting correlation patterns are summarized in Figure 6, where statistically significant associations are highlighted. Overall, several expected relationships are confirmed, while others did not reach statistical significance.

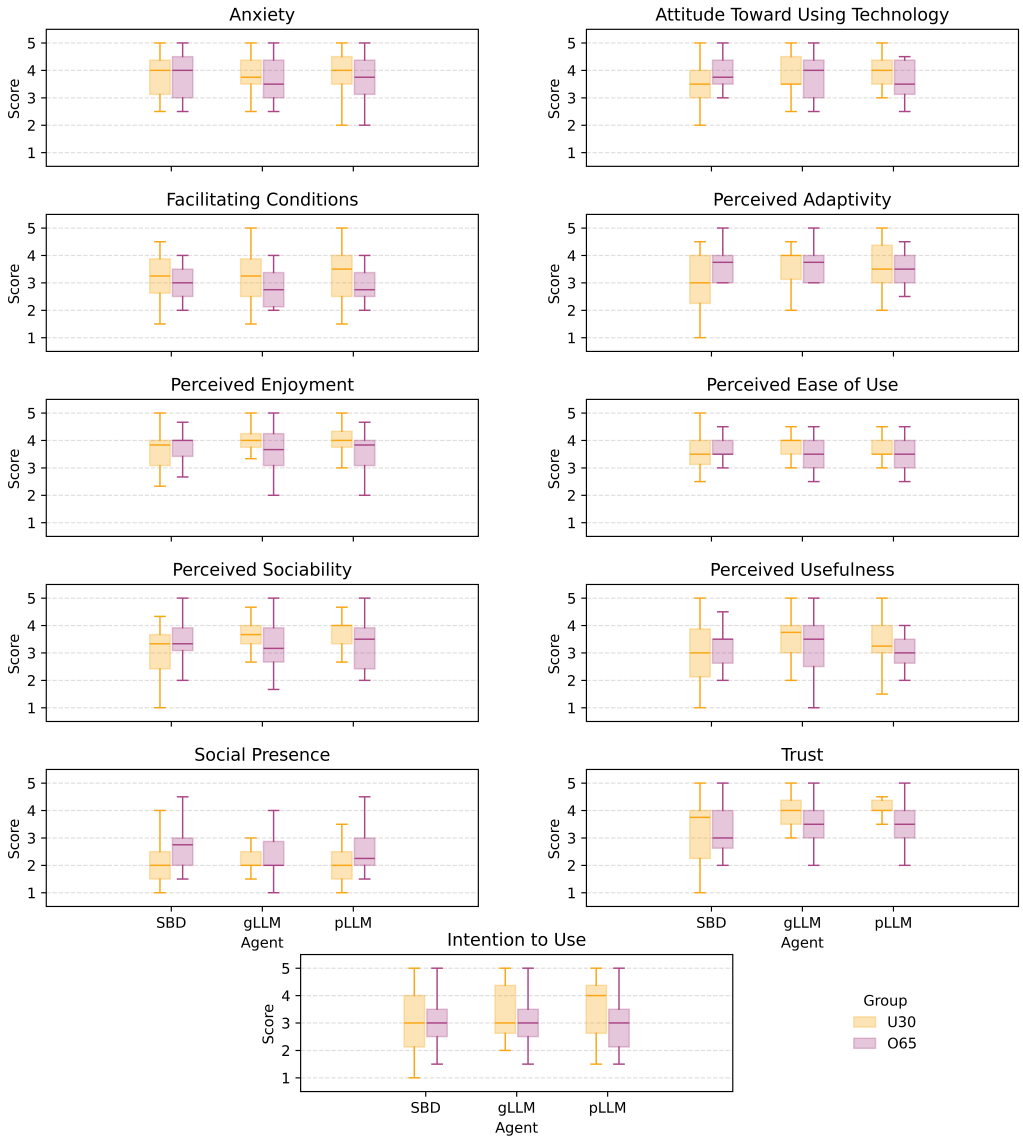

Figure 5. Boxplot representing the distribution of the average scores for each Almere model construct. Boxes are color-coded per age group, and grouped per agent modality used in the interaction (namely script-based dialogue, generic LLM, and personalized LLM). Results from the ANX construct are reported for completeness, despite the construct items having a below-threshold internal consistency.

The post-test interviews revealed several recurring themes. Focusing on the O65 group, difficulties related to audio quality were frequently reported, with 8 out of 18 participants indicating that the robot's speech was sometimes hard to understand. Delays in the conversation were also reported as a barrier to a smooth interaction flow. In this regard, the response delay time was 9.70 ± 3.28 s in the SBD modality, 12.56 ± 3.24 s in gLLM, and 13.18 ± 2.58 s in pLLM. Several participants (5/18) emphasized the potential usefulness

of the robot as a form of support for individuals experiencing loneliness or difficulties in independently using common technological devices (e.g., smartphones or tablets). Some participants (3/18) compared the robot to commercial voice assistants, noting the added value of physical embodiment in fostering engagement and empathy. With respect to the interaction dynamics, a subset of users (4/18) reported difficulty in proposing conversation topics, underlining the importance of the robot being proactive in guiding the dialogue. Finally, positive impressions were expressed by several participants (4/18), who reported surprise at the robot's perceived "culture".

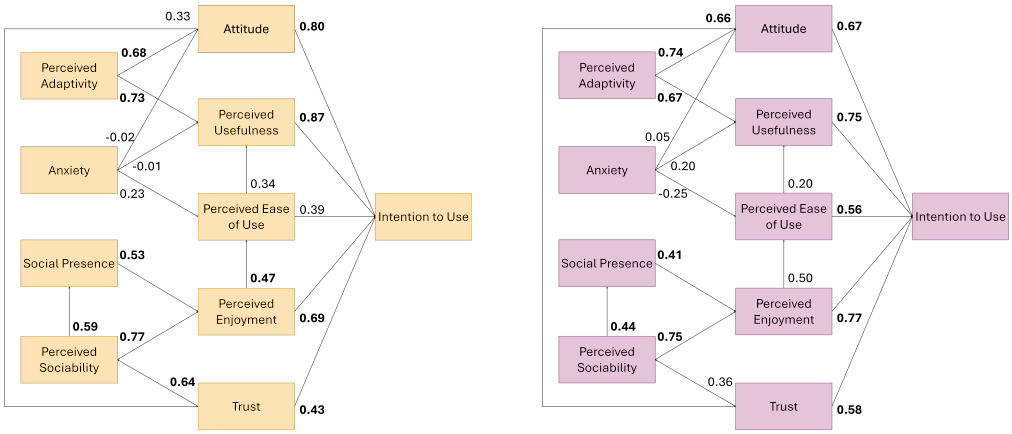

Figure 6. Correlation structure of the Almere technology acceptance model by age group. Pearson correlations between acceptance constructs are computed separately for the U30 (**left**) and O65 (**right**) groups. Significant associations are highlighted.

4. Discussion

This study aimed to investigate whether different implementations of a conversational agent (scripted or agentic, with or without user profile information) affect the perceived quality of human–robot interaction in younger and older users. Particular attention was devoted to understanding whether recent advances in generative dialogue systems, and the inclusion of user profiling information, translate into measurable differences in technology acceptance within a short, speech-only interaction.

The mixed-design ANOVA indicated a potential effect of interaction modality, emerging only within age groups and not as a global main effect. Given the small-to-moderate effect sizes observed, the lack of statistically significant differences in the post hoc pairwise comparisons is not unexpected. This suggests that, while modality-related trends may exist, their magnitude is limited within the present experimental setting. The interpretation of these results may be related to the younger participants' greater ability to distinguish the non-agentic modality and evaluate it accordingly. Indeed, the SBD modality was, on average, perceived by the U30 group as less sociable, adaptive, and useful than the gLLM and pLLM modalities, a pattern that did not emerge in the O65 group. However, these differences remain relatively small and therefore do not reach statistical significance in the post hoc analysis after the stringent Bonferroni correction applied to the nominal significance level of 0.05 (i.e., $p_{adj} = 0.05/33 = 1.5 * 10^{-3}$). Despite the absence of strong statistical effects, some qualitative patterns can be noted. Restricting the discussion to constructs with satisfactory internal consistency (i.e., excluding ANX and PEOU), median scores for PAD, PENJ, PU were consistently above the midpoint across conditions. These results indicate that the robot was generally perceived as enjoyable and useful, and as a system that could plausibly respond to users' needs. In addition, both age groups reported relatively high levels of trust (TR), suggesting a potential use of the robot in promoting positive behaviors, such as encouragement of healthy habits or engagement in cognitively

stimulating activities. Conversely, the SP construct received low scores across all modalities and age groups. This outcome can be translated into a limited perceived human-likeliness of the robot. Such an interpretation well aligns with feedback during the post-test interviews, clearly classifying the robot as a surrogate of human presence, and identifying the lack of richer social expressiveness (e.g., smile, facial expression) as a notable limitation. Finally, the generally positive scores observed for the ATT construct may reflect a bias in the population taking part in the study, particularly within the O65 group. Indeed, older adult volunteers for a robotics study are likely to hold a priori positive expectations toward technology, and are thus more prone to favorable evaluations. This effect should be taken into account when interpreting the overall acceptance levels reported in this study.

Looking at the constructs correlation, for both age groups, ATT emerged as strongly correlated to ITU (O65: $\rho_{ATT \rightarrow ITU} = 0.80$, $p = 1.65 * 10^{-10}$, U30: $\rho_{ATT \rightarrow ITU} = 0.67$, $p = 1.81 * 10^{-8}$) confirming the impact of the potential favorite subjects preconception toward technology. Among other constructs, PU and PENJ emerged as the most strongly correlated with ITU. Within the O65 group, PU ($\rho_{ATT \rightarrow ITU} = 0.75$, $p = 5.80 * 10^{-11}$) showed a particularly strong association with ITU, suggesting that older adults greatly value the perceived practical utility of the system. In the U30 group, while PU remained an important predictor of ITU ($\rho_{ATT \rightarrow ITU} = 0.87$, $p = 1.60 * 10^{-18}$), it was more balanced with PENJ ($\rho_{ATT \rightarrow ITU} = 0.69$, $p = 1.02 * 10^{-8}$). This pattern is expected, as it may reflect generational differences, as younger adults tend to associate technology more with entertainment, rather than support and assistance. In this regard, and with respect to the experimental protocol, it should be noted that the entertainment component offered by the robot was primarily illustrative and thus very limited. Participants were therefore required to infer potential real-world applications of the robot rather than directly experience them. While this choice was motivated by the desire to provide a concise example of everyday use, keeping the focus on the interaction component, the simplicity of the proposed activities may have hindered the possibility for users to understand the robot's broader potential. This limitation may help to explain the substantial variability observed in PU responses, which, in turn, likely affected the inconclusive results in the ITU. Moreover, the limited duration of the conversation (i.e., 5 min) likely did not provide sufficient time for users to explore the depth of the robot/agent's reasoning, making it more difficult to appreciate its potentiality and adaptability. Similarly, the short interaction time probably reduced participants' opportunity to detect differences between modalities, in particular between the gLLM and pLMM ones.

Despite the statistical analysis not revealing remarkable results, the responses collected from participant interviews after the tests revealed positive aspects and areas for improvement. Participants appreciated the robot's friendly appearance, perceived knowledge, and empathetic conversational style. They deemed TIAGo to be a valuable companion, especially for lonely subjects, possibly struggling with the independent use of technology. In this regard, TIAGo's ability to adapt to user inputs and provide engaging content like audiobooks and music was well-received. However, critical issues emerged that help to complement the quantitative findings. Participant reported limited expressiveness, confusing interaction times, and poor audio quality as major drawbacks. The conversation flow was sometimes considered too slow and lacking fluidity, also due to occasional speech misinterpretations. The TTS quality was considered a limiting factor by both the age groups. The use of standard software packages included in default ROS installations was motivated by simplicity and ease of system replication. However, these results suggest that adopting more advanced TTS solutions, capable of producing more natural and expressive speech, would substantially improve the perceived quality of interaction and overall user experience. In addition, several users reported difficulties in understanding the robot's transitions

between speaking and listening states. This lack of clear feedback led to uncertainty during turn-taking and overlaps, particularly among older participants. Notably, some users from the O65 group suggested the introduction of simple visual cues, such as colored lights, to explicitly signal the robot's current interaction state. Such mechanisms could significantly enhance interaction clarity without increasing system complexity.

Finally, it is worth noting that, during the post-test interviews, participants were explicitly asked whether they perceived differences across trials. Some users provided responses suggesting that they recognized distinctions between the agentic (gLLM, pLLM) and scripted (SBD) conditions. Nevertheless, even for these participants, questionnaire responses did not differentiate between interaction modalities. This fact may be partly explained by a novelty effect, due to which subtle conversational differences are masked by the overall experience of interacting with a social robot. At the same time, this discrepancy between interviews and questionnaire feedback might also be a consequence of a suboptimal selection of the evaluation tools, lacking explicit closed questions on differences among agents. In this regard, similar studies would benefit from repeated exposures over an extended period of time, and from a more structured familiarization phase to mitigate the novelty effect. However, the personalization was consistently observed to be effective, in particular driving the conversation topics of interest from the user's profile. When relevant to the ongoing dialogue, occasional references to the user's background were also observed in the responses. Anecdotal examples of such behaviors are provided in the Supplementary Materials.

All in all, the present study suggests that, within short, verbal HRI, variations in dialogue implementation may have a limited impact on technology acceptance measures, particularly when compared to broader factors such as baseline user attitude, perceived usefulness, and interaction quality. In particular, coupling the conversational interaction with a meaningful functional task appears to play an important role toward acceptance. At the same time, qualitative feedback highlights design elements, including audio quality, turn-taking feedback, and transcription delays, as critical factors in user experience. These findings indicate that, before subtle differences between conversational agents can be meaningfully assessed, technical interaction aspects must be robustly addressed, especially when targeting older user populations.

5. Conclusions

This paper investigated the impact of different conversational agent implementations on users' acceptance of a social robot, comparing a scripted dialogue approach (SBD) with generic and personalized LLM-based agents (gLLM and pLLM, respectively). The study involved younger (U30) and older (O65) participants, assessing user experience through a questionnaire grounded in the Almere technology acceptance model and qualitative post-test interviews. The experimental design focused on short, speech interactions, followed by a simple entertainment activity, intentionally limiting task complexity to isolate conversational aspects of the interaction.

Overall, the results indicate that differences in dialogue implementation had a limited effect on acceptance measures. While some interaction effects emerged within age groups, no consistent or robust differences were observed across modalities in post hoc comparisons. The questionnaire findings, together with qualitative feedback, suggest that user attitude toward technology, together with the implementation of both the interaction framework and the experimental setting, are crucial factors toward an effective HRI. In particular, speech quality emerged as a critical factor, indicating the need for enhanced TTS solutions. In this view, the findings of the present pilot study suggest that, for users to grasp nuanced aspects of the interaction, the robot should be capable of producing equally nuanced output. This could be achieved, e.g., by introducing a TTS module capable of emotional speech

synthesis, enabling the agent to modulate not only the content of the interaction, but also the prosody [33]. Similarly, a more advanced model, outperforming the GPT3.5-turbo used in this work in terms of empathy and complex reasoning, would likely contribute to better differentiating the different interaction modalities. Moreover, the findings from this research suggest that conversational interaction should be more proactively guided by the robot and more tightly coupled with goal-oriented tasks. User feedback also helped in identifying simple interventions to improve the current system, such as the introduction of visual cues to represent the robot's internal states. Finally, future improvement to the current framework should include a more effective implementation of the agent memory, together with the ones identified above. Indeed, an important limitation of the proposed system concerns the memory mechanism adopted in the current implementation. The system relies on a rudimentary approach in which the entire conversational log is provided as context, a solution that does not scale to longer interactions and is unsuitable for real-world deployment. Moreover, this approach only supports short-term memory within a single session, with no information retained across interactions. The integration of a long-term memory component, e.g., periodically summarizing relevant user information through additional LLM calls [23], emerges as a necessary step toward long-term, personalized HRIs.

In conclusion, this work suggests that, before fine-grained differences between conversational agent architectures can be meaningfully evaluated, fundamental interaction aspects must be carefully addressed. Embedding dialogue within functional tasks, ensuring high-quality speech output, and supporting longer interactions, both in terms of session number and duration, appear to be key prerequisites for fostering acceptance, particularly in older user populations. These insights contribute to informing future integrations of conversational agents into social robots and their evaluation in the interaction with older adults, toward a real-world deployment.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app16073369/s1.

Author Contributions: Conceptualization, M.G., L.P., and F.B.; methodology, M.G., N.T., and L.P.; software, M.N. and L.P.; investigation, M.N., and L.P.; formal analysis, L.P., M.N., N.T., and M.G.; writing—original draft preparation, L.P. and M.N.; writing—review and editing, M.G., N.T., and F.B.; visualization, L.P.; supervision, M.G. and F.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Politecnico di Milano (46/2022) on 16 November 2022.

Informed Consent Statement: Informed consent to the participation in the study and the publication of the present paper was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy reasons.

Conflicts of Interest: M.G. and F.B. hold shares in AllyArm srl and AGADE srl.

References

1. World Health Organization. Ageing and Health. Available online: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health (accessed on 31 October 2025).
2. Macdonald, B.; Luo, M.; Hülür, G. Daily social interactions and well-being in older adults: The role of interaction modality. *J. Soc. Pers. Relatsh.* **2021**, *38*, 3566–3589. [CrossRef]

3. Kelly, M.E.; Loughrey, D.; Lawlor, B.A.; Robertson, I.H.; Walsh, C.; Brennan, S. The impact of cognitive training and mental stimulation on cognitive and everyday functioning of healthy older adults: A systematic review and meta-analysis. *Ageing Res. Rev.* **2014**, *15*, 28–43. [CrossRef]
4. Clemens, S.; Wodchis, W.; McGilton, K.; McGrail, K.; McMahon, M. The relationship between quality and staffing in long-term care: A systematic review of the literature 2008–2020. *Int. J. Nurs. Stud.* **2021**, *122*, 104036. [CrossRef]
5. Holland, J.; Kingston, L.; McCarthy, C.; Armstrong, E.; O'Dwyer, P.; Merz, F.; McConnell, M. Service Robots in the Healthcare Sector. *Robotics* **2021**, *10*, 47. [CrossRef]
6. Abdollahi, H.; Mahoor, M.H.; Zandie, R.; Siewierski, J.; Qualls, S.H. Artificial Emotional Intelligence in Socially Assistive Robots for Older Adults: A Pilot Study. *IEEE Trans. Affect. Comput.* **2023**, *14*, 2020–2032. [CrossRef]
7. Khosla, R.; Nguyen, K.; Chu, M.T. Human Robot Engagement and Acceptability in Residential Aged Care. *Int. J. Hum.–Comput. Interact.* **2016**, *33*, 510–522. [CrossRef]
8. Mishra, N.; Tulsulkar, G.; Li, H.; Thalmann, N.M.; Er, L.H.; Ping, L.M.; Khoong, C.S. Does Elderly Enjoy Playing Bingo with a Robot? A Case Study with the Humanoid Robot Nadine. In *Advances in Computer Graphics*; Springer International Publishing: Cham, Switzerland, 2021; pp. 491–503. [CrossRef]
9. Thompson, C.; Mohamed, S.; Louie, W.Y.G.; He, J.C.; Li, J.; Nejat, G. The robot Tangy facilitating Trivia games: A team-based user-study with long-term care residents. In Proceedings of the 2017 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS), Ottawa, ON, Canada, 5–7 October 2017; pp. 173–178. [CrossRef]
10. Pou-Prom, C.; Raimondo, S.; Rudzicz, F. A Conversational Robot for Older Adults with Alzheimer's Disease. *ACM Trans. Hum.-Robot Interact.* **2020**, *9*, 1–25. [CrossRef]
11. Luperto, M.; Monroy, J.; Renoux, J.; Lunardini, F.; Basilico, N.; Bulgheroni, M.; Cangelosi, A.; Cesari, M.; Cid, M.; Ianes, A.; et al. Integrating Social Assistive Robots, IoT, Virtual Communities and Smart Objects to Assist at-Home Independently Living Elders: The MoveCare Project. *Int. J. Soc. Robot.* **2022**, *15*, 517–545. [CrossRef]
12. Silvera-Tawil, D. Robotics in Healthcare: A Survey. *SN Comput. Sci.* **2024**, *5*, 189. [CrossRef]
13. Pozzi, L.; Gheduzzi, E.; Lettieri, E.; Braghin, F.; Gandolla, M. Socially Assistive Robots in Nursing Homes: A Literature Review and Demand-Supply Analysis. Under preparation, 2025.
14. Bonarini, A. Communication in Human-Robot Interaction. *Curr. Robot. Rep.* **2020**, *1*, 279–285. [CrossRef]
15. Robert, L.P.J. Personality in the Human Robot Interaction Literature: A Review and Brief Critique. In Proceedings of the Twenty-fourth Americas Conference on Information Systems, New Orleans, LA, USA, 16–18 August 2018.
16. Chen, Y.; Cui, W.; Chen, Y.; Tan, M.; Zhang, X.; Liu, J.; Li, H.; Zhao, D.; Wang, H. RoboGPT: An LLM-Based Long-Term Decision-Making Embodied Agent for Instruction Following Tasks. *IEEE Trans. Cognit. Dev. Syst.* **2025**, *17*, 1163–1174. [CrossRef]
17. Honerkamp, D.; Büchner, M.; Despinoy, F.; Welschehold, T.; Valada, A. Language-Grounded Dynamic Scene Graphs for Interactive Object Search With Mobile Manipulation. *IEEE Robot. Autom. Lett.* **2024**, *9*, 8298–8305. [CrossRef]
18. Chalvatzaki, G.; Younes, A.; Nandha, D.; Le, A.T.; Ribeiro, L.F.R.; Gurevych, I. Learning to reason over scene graphs: A case study of finetuning GPT-2 into a robot language model for grounded task planning. *Front. Robot. AI* **2023**, *10*, 1221739. [CrossRef]
19. Singh, I.; Blukis, V.; Mousavian, A.; Goyal, A.; Xu, D.; Tremblay, J.; Fox, D.; Thomason, J.; Garg, A. ProgPrompt: Generating Situated Robot Task Plans using Large Language Models. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 11523–11530. [CrossRef]
20. Pashangpour, S.; Nejat, G. The Future of Intelligent Healthcare: A Systematic Analysis and Discussion on the Integration and Impact of Robots Using Large Language Models for Healthcare. *Robotics* **2024**, *13*, 112. [CrossRef]
21. Schlesener, E.A.; Ziolkowski, M.; Wong, S.K.; Westmoreland, B.; Babu, S.V. 'Am I Understood?': How the Interplay Between Embodiment and Theory of Mind Behavior Affects LLM-based Conversational Agents on Perceived Trust, Anthropomorphism, Presence, Usability, and User Experience. *ACM Trans. Interact. Intell. Syst.* **2025**, *16*, 1–45. [CrossRef]
22. Lo, J.H.; Huang, H.P.; Lo, J.S. LLM-based robot personality simulation and cognitive system. *Sci. Rep.* **2025**, *15*, 16993. [CrossRef]
23. Pinto-Bernal, M.; Biondina, M.; Belpaeme, T. Designing Social Robots with LLMs for Engaging Human Interaction. *Appl. Sci.* **2025**, *15*, 6377. [CrossRef]
24. Irfan, B.; Kuoppamäki, S.; Hosseini, A.; Skantze, G. Between reality and delusion: Challenges of applying large language models to companion robots for open-domain dialogues with older adults. *Autonom. Robots* **2025**, *49*, 9. [CrossRef]
25. Khoo, W.; Hsu, L.J.; Amon, K.J.; Chakilam, P.V.; Chen, W.C.; Kaufman, Z.; Lungu, A.; Sato, H.; Seliger, E.; Swaminathan, M.; et al. Spill the Tea: When Robot Conversation Agents Support Well-being for Older Adults. In *HRI '23: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*; ACM: New York, NY, USA, 2023; pp. 178–182. [CrossRef]
26. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv* **2022**, arXiv:2212.04356. [CrossRef]
27. Taylor, P.A.; Black, A.W.; Caley, R. The architecture of the Festival speech synthesis system. In Proceedings of the Speech Synthesis Workshop, Jenolan Caves House, Blue Mountains, Australia, 26–29 November 1998.

28. Ye, J.; Chen, X.; Xu, N.; Zu, C.; Shao, Z.; Liu, S.; Cui, Y.; Zhou, Z.; Gong, C.; Shen, Y.; et al. A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models. *arXiv* **2023**, arXiv:2303.10420. [CrossRef]
29. White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; Schmidt, D.C. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv* **2023**, arXiv:2302.11382. [CrossRef]
30. Heerink, M.; Kröse, B.; Evers, V.; Wielinga, B. Assessing Acceptance of Assistive Social Agent Technology by Older Adults: The Almere Model. *Int. J. Soc. Robot.* **2010**, *2*, 361–375. [CrossRef]
31. Ratner, B. The correlation coefficient: Its values range between +1/−1, or do they? *J. Target. Meas. Anal. Mark.* **2009**, *17*, 139–142. [CrossRef]
32. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Routledge: New York, NY, USA, 2013. [CrossRef]
33. Ma, F.; Xie, Y.; Li, Y.; He, Y.; Zhang, Y.; Ren, H.; Liu, Z.; Yao, W.; Ren, F.; Yu, F.R.; et al. A Review of Human Emotion Synthesis Based on Generative Technology. *IEEE Trans. Affect. Comput.* **2025**, *16*, 2579–2598. [CrossRef]