Two-Wave Comparative Study With Independent Samples on Integrating a Large Language Model Into a Socially Assistive Robot in a Hospital Geriatric Unit: Performance, Engagement, and User Perceptions

Lauriane Blavette, Sébastien Dacunha, Xavier Alameda-Pineda, Jeanne Cattoni, Maribel Pino, Anne-Sophie Rigaud

Submitted to: JMIR Human Factors
on: August 13, 2025

Table of Contents

Two-Wave Comparative Study With Independent Samples on Integrating a Large Language Model Into a Socially Assistive Robot in a Hospital Geriatric Unit: Performance, Engagement, and User Perceptions

Lauriane Blavette[1, 2, 3] MSc; Sébastien Dacunha[1, 2, 4] MSc; Xavier Alameda-Pineda[5] PhD; Jeanne Cattoni[1, 2, 4] MSc; Maribel Pino[1, 2, 4] PhD; Anne-Sophie Rigaud[1, 2, 4] Prof Dr Med

[1] Institut national de la santé et de la recherche médicale - Optimisation thérapeutique en pharmacologie OTEN U1144, Université Paris Cité Paris FR
[2] Assistance Publique - Hôpitaux de Paris, Hôpital Broca, Centre Mémoire de Ressources et Recherches Île-de-France-Broca, Service gériatrie 1&2 Paris FR
[3] Broca Living Lab, Hôpital Broca Paris FR
[4] Institut national de recherche en sciences et technologies du numérique de l'Université Grenoble Alpes Grenoble FR

Corresponding Author:
Lauriane Blavette MSc

Institut national de la santé et de la recherche médicale - Optimisation thérapeutique en pharmacologie OTEN U1144, Université Paris Cité
Case 15, Faculté de Pharmacie, 4, avenue de l'observatoire
Paris
FR

Abstract

Background: Addressing the complex medical and psychosocial needs of older adults (OAs) is increasingly difficult in resource-limited care settings. In this context, socially assistive robots (SARs) provide support and practical functions such as orientation and information delivery. Integrating large language models (LLMs) into SARs dialogue systems offers opportunities to improve interaction fluency and adaptability. Yet in real-world use, acceptability also depends on minimizing both technical and conversational errors, ensuring successful user interactions, and adapting to individual user characteristics.

Objective: This study aimed to evaluate the impact of integrating a large language model into a SAR dialogue system in a hospital geriatric unit by (1) comparing system performance and interaction success across two experimental waves, (2) examining the links between robot errors, interaction success, and multidimensional user engagement, and (3) exploring how user characteristics influence performance and perceptions of acceptability and usability.

Methods: Over an 8-month period, 28 older adults (OAs; ?60 years) attending a geriatric day care hospital (Paris, France) participated in a single-session evaluation of a SAR. Interactions took place in the DCH and were video-recorded across two waves: wave 1 (basic dialogue system) and wave 2 (LLM- enhanced system). From the recordings, system performance (error types, interaction success) and user engagement (verbal, physical, and emotional dimensions) were coded. Acceptability and usability were measured using the Acceptability E-scale and the System Usability Scale. Sociodemographic data were collected, and quantitative results were supplemented with a thematic analysis of qualitative observations.

Results: Following LLM integration, error-free interactions increased from 27.8% to 70.2% (P<.001), comprehension failures decreased from 47.2% to 17% (P<.001), and interaction success rose from 25.0% to 74.5% (P<.001). Acceptability (AES: 12.8 vs 20.8, P=.003) and usability (SUS: 40.0 vs 60.4, P=.04) were significantly higher in wave 2. Engagement scores did not differ significantly between waves, though emotional engagement correlated positively with interaction success (r=0.28, P<.01), and age was negatively associated with both physical engagement (r=–0.30, P<.001) and acceptability (r=–0.20, P<.05).

Conclusions: Behavioral engagement with a SAR in geriatric care is shaped by both system performance and individual user characteristics. Improvements in dialogue quality, particularly through the integration of a LLM, were associated with higher interaction success and enhanced user experience. These findings highlight the importance of combining multimodal behavioral analysis with self-reported measures to inform the iterative, user-centered design of socially responsive robots in clinical contexts. Clinical Trial: The study was approved by the French National Ethics Committee (“Comité de Protection des

Personnes, CPP Ouest II, Maison de la Recherche Clinique – CHU Angers"; Institutional Review Board [IRB] 2021/20) and complied with the General Data Protection Regulation (GDPR). Data processing was registered with the Data Protection Officer (DPO) under reference number 20210114153645 in the AP-HP registry.

(JMIR Preprints 13/08/2025:81936)
DOI: https://doi.org/10.2196/preprints.81936

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Original Paper

Two-Wave Comparative Study With Independent Samples on Integrating a Large Language Model Into a Socially Assistive Robot in a Hospital Geriatric Unit: Performance, Engagement, and User Perceptions

Lauriane Blavette[1,2,3,*], Sébastien Dacunha[1,2,3], Xavier Alameda-Pineda[4], Jeanne Cattoni[1,2,3], Anne-Sophie Rigaud[1,2,3], Maribel Pino[1,2,3]

[1]Université Paris-Cité, Institut national de la santé et de la recherche médicale - Optimisation thérapeutique en pharmacologie OTEN U1144, 75006 Paris, France

[2]Service gériatrie 1&2, Centre Mémoire de Ressources et Recherches Île-de-France-Broca, Assistance publique – Hôpitaux de Paris, Hôpital Broca, F-75013 Paris, France.

[3] Broca Living Lab, Hôpital Broca, Paris, France

[4]Institut national de recherche en sciences et technologies du numérique de l'Université Grenoble Alpes, France

[*] Corresponding author: lauriane.blavette@aphp.fr

Abstract

Background: Addressing the complex medical and psychosocial needs of older adults (OAs) is increasingly difficult in resource-limited care settings. In this context, socially assistive robots (SARs) provide support and practical functions such as orientation and information delivery. Integrating large language models (LLMs) into SARs dialogue systems offers opportunities to improve interaction fluency and adaptability. Yet in real-world use, acceptability also depends on minimizing both technical and conversational errors, ensuring successful user interactions, and adapting to individual user characteristics.

Objective: This study aimed to evaluate the impact of integrating a large language model into a SAR dialogue system in a hospital geriatric unit by (1) comparing system performance and interaction success across two experimental waves, (2) examining the links between robot errors, interaction success, and multidimensional user engagement, and (3) exploring how user characteristics influence performance and perceptions of acceptability and usability.

Methods: Over an 8-month period, 28 OAs (OAs; ≥60 years) attending a geriatric day care hospital (Paris, France) participated in a single-session evaluation of a SAR. Interactions took place in the DCH and were video-recorded across two waves: wave 1 (basic dialogue system) and wave 2 (LLM-enhanced system). From the recordings, system performance (error types, interaction success) and user engagement (verbal, physical, and emotional dimensions) were coded. Acceptability and usability were measured using the Acceptability E-scale and the System Usability Scale. Sociodemographic data were collected, and quantitative results were supplemented with a thematic analysis of qualitative observations.

Results: Following LLM integration, error-free interactions increased from 27.8% to 70.2% (P<.001), comprehension failures decreased from 47.2% to 17% (P<.001), and interaction success rose from 25.0% to 74.5% (P<.001). Acceptability (AES: 12.8 vs 20.8, P=.003) and usability (SUS: 40.0 vs 60.4, P=.04) were significantly higher in wave 2. Engagement scores did not differ significantly between waves, though emotional engagement correlated positively with interaction success (r=0.28, P<.01), and age was negatively associated with both physical engagement (r=–0.30,

P<.001) and acceptability (r=–0.20, *P*<.05).
Conclusion: Behavioral engagement with a SAR in geriatric care is shaped by both system performance and individual user characteristics. Improvements in dialogue quality, particularly through the integration of a LLM, were associated with higher interaction success and enhanced user experience. These findings highlight the importance of combining multimodal behavioral analysis with self-reported measures to inform the iterative, user-centered design of socially responsive robots in clinical contexts.
Keywords: Human-robot interaction; Socially assistive robot; Older adults, Behavioral engagement; Multimodal analysis; Geriatric care; Large language models.

Introduction

The rapid aging of the population worldwide, combined with a shortage of health care professionals, is placing growing pressure on care systems, particularly in geriatric hospitals and long-term care settings [1]. This strain is amplified when caring for older adults (OAs) with neurocognitive disorders, such as Alzheimer's disease and related dementias, whose needs extend beyond medical supervision to include sustained social, emotional, and cognitive support. However, many care institutions, including hospitals, long-term care facilities, and other residential settings, often lack the resources and structures needed to address these complex needs, resulting in care experiences characterized by unmet psychosocial demands, ineffective communication strategies, and increased vulnerability to distress and disorientation [2-4].

Socially assistive robots (SARs) have emerged as a promising technological complement to human-delivered care in such contexts. SARs are designed primarily for social rather than physical interaction, using speech, gesture, and expressive behaviors to engage users and deliver cognitive, emotional, or informational support [5, 6]. Building on their social capabilities, SARs have shown potential in enhancing emotional well-being, reducing anxiety and apathy, and promoting social interaction among OAs, particularly in institutional settings, including but not limited to those living with dementia [7-9]. Beyond therapeutic applications, SARs are increasingly explored as tools to assist healthcare professionals in non-clinical tasks, such as welcoming patients, providing orientation, facilitating recreational activities, or delivering health-related information [10-12].

These institutional applications often encompass reception, visitor guidance, and wayfinding. For instance, studies have shown that robots acting as medical receptionists or lobby greeters can facilitate orientation, manage patient flow, and enhance the perceived quality of service, particularly in high-traffic environments [13,14]. A systematic review by González-González et al. [15]. further suggests that SARs in hospital contexts often adopt hybrid roles, combining informational support with elements of emotional engagement, especially in waiting rooms or ambulatory care settings. Such use cases broaden the vision of SARs beyond therapeutic agents to include communication and orientation facilitators within care environments [16], aligning with broader efforts to improve patient experience where staff availability is limited.

However, despite advances in robotics, including improved sensor technologies, natural language processing, and adaptive user interfaces, the integration of SARs into real-world care environments remains limited, with technical limitations representing one of the most significant barriers [17]. Healthcare providers have reported frustration with issues such as complex operational steps, short battery life, slow system responses, and rigid dialogue systems [18, 19]. These limitations may discourage routine use, particularly in dynamic care settings where efficiency is essential. In addition, persistent challenges in personalization and speech recognition further limit deployment, as current systems often fail to accommodate the wide range of capabilities, preferences, and interaction styles found among geriatric users [20, 21].

From the end user's perspective, errors in human–robot interaction (HRI), including comprehension failures, incoherent or incomplete responses, and misaligned task execution, can erode trust in the robot and reduce engagement [22]. For OAs, even minor conversational breakdowns (e.g., misrecognitions or incorrect answers) may trigger confusion, frustration, and diminished trust, particularly when the task is perceived as important, underscoring the need for error-tolerant, clearly structured dialogue [23]. As shown by Kim et al. [24], OAs can also struggle to adapt their speech to the rigid input formats required by conversational systems, sometimes realizing only after delivering a lengthy command that the system had failed to recognize or process their request. Such breakdowns not only cause irritation but also expose the gap between the intended "natural" interaction and the constrained, one-directional exchanges these systems often afford. Similar patterns have been observed elsewhere: Khosla et al. [25] in a three-month deployment of a SAR with five older adults, reported that negative emotions (e.g., anger, sadness, anxiety) occasionally emerged during interactions with the robot, most often when it failed to respond as expected.

Recent advances in large language models (LLMs) offer the potential to mitigate some of these conversational limitations by improving fluency, contextual relevance, and adaptability in robot dialogue systems [26-28]. While early results from other domains (e.g., customer service, museum guidance, educational robotics) are promising, there is limited empirical evidence on how integrating LLMs into SARs affects real-time interactions, error rates, and user experience in real-world geriatric care contexts.

Another critical but underexamined factor in SAR evaluation is user engagement, which is inherently multidimensional, encompassing *verbal behaviors* (e.g., speech production, turn-taking), *physical involvement* (e.g., gestures, posture), and *emotional signals* (e.g., facial expressions or affective cues) [29]. In HRI, engagement not only serves as a proxy for interaction quality but also predicts subsequent outcomes, such as willingness to re-engage with the robot and the perceived value of the system [30]. Research in care environments, including work with OAs by Hebesberger et al [19], shows that sustained engagement is essential for acceptance and that both technical reliability and interaction fluency shape the depth and duration of participation. Despite its theoretical and practical importance, little is known about how these engagement dimensions are influenced by robot performance, error frequency, or task success in clinical and geriatric contexts, representing a notable gap in the literature.

Finally, user characteristics, including age, socio-educational background, and cognitive functioning, are likely to influence both observable HRI behaviors and subjective evaluations of SARs [31-34]. For instance, older age may be associated with reduced physical expressiveness, while higher cognitive functioning could support more complex conversational exchanges. Understanding these relationships is essential for informing adaptive robot design and deployment strategies that are inclusive and responsive to diverse user needs.

Given these gaps, the present study investigated the deployment of a SAR in a hospital geriatric unit, focusing on three main objectives:

1. *To assess changes in system performance and interaction success following the integration of an LLM into the robot's dialogue system*, comparing two experimental waves.
2. *To examine the relationships between robot errors, interaction success, and user engagement*, considering verbal, physical, and emotional dimensions.
3. *To explore how user characteristics relate to system performance and subjective evaluations* of acceptability and usability.

By combining quantitative measures of performance, engagement, and user ratings with qualitative analysis of participant perceptions, this study provides a real-world grounded perspective on the opportunities and challenges of SAR deployment in aging care contexts.

Methods

Participants

The study involved OAs attending consultations at the geriatric DCH of Broca Hospital (Assistance Publique–Hôpitaux de Paris, AP-HP), France.

Inclusion criteria were (1) being ≥60 years old; (2) participation in a scheduled DCH consultation; (3) having a Mini-Mental State Examination (MMSE) [35] score above10, indicating the absence of severe cognitive impairment; (4) no current symptoms of altered reality (e.g., delusions or hallucinations); and (5) fluent comprehension and expression in French.

No *exclusion criteria* were applied based on gender, socio-economic backgrounds, or ethnicity.

Recruitment

Recruitment was carried out using the DCH database, and participants were prescreened prior to enrollment, contacted over the phone, and invited to participate in the study the day of their next consultation. An information letter was sent by post and informed consent was collected onsite.

A total of 41 OAs were recruited for the study. However, the present analysis focuses on a subset of 28 participants who met the following criteria: (1) provided consent to be filmed; (2) completed a full interaction session with the robot; and (3) generated usable video data, defined by adequate visibility and recording quality. The remaining participants were excluded due to one or more of the following reasons: refusal to engage with the robot, technical malfunctions during the session, or withdrawal of consent for video recording.

Setting

The study was conducted between May 2023 and December 2023 in the DCH of a geriatric hospital in Paris (France). The DCH provides specialized outpatient care for OAs with physical or cognitive impairments, offering a wide range of consultations, including neurology, oncology, cardiology, psychiatry, and memory assessments.

Data collection occurred in two waves during this period, with distinct participant samples recruited for each phase. All interactions were carried out in a quiet, dedicated room located near the DCH waiting area. This space, typically used for rest and informal activities, was chosen to provide a calm and comfortable environment for testing, while maintaining proximity to the clinical setting.

Study Design

This cross-sectional observational study was conducted as part of a broader research protocol evaluating the integration of a SAR in geriatric care [36, 37]. The present analysis focuses on real-world HRI, combining behavioral observations with user-reported data. Each participant engaged in a single, non-scripted interaction session with the robot, followed by questionnaires. Data were collected across two experimental waves. The first wave involved a baseline version of the robot without LLM integration, while the second wave used an updated LLM-enabled version. Each wave included a different group of OAs.

Material

ARI Robot

The ARI robot developed by Pal Robotics (Spain) is a 1.65-meter (5 ft 5 in) wheeled humanoid platform equipped with a touchscreen, cameras, microphones, animated eyes, and articulated arms (Figure 1). In this study, ARI was used as a socially assistive agent in a geriatric hospital setting.

For this study, ARI was programmed with interaction modules developed as part of the European H2020 SPRING (Socially Pertinent Robots in Gerontological Healthcare) project, aimed at enabling SAR in real-world clinical environments [36]. Two different configurations of ARI's conversational system were tested during this study, as detailed in the following section ("System Evolution"). As ARI had not yet been commercially deployed in care institutions at the time of the study, its use in this context was part of a controlled, exploratory evaluation conducted within the hospital environment.

Figure 1. Front and side views of the ARI robot. (Photo credit: PAL Robotics)

System Evolution

To assess the impact of iterative improvements to ARI's interaction capabilities, we conducted two experimental waves, each following a major system update informed by participant feedback.

- Wave 1 (May–July 2023) deployed a first version of ARI featuring a modular dialogue system based on traditional rule-based intent handling, retrieval-based responses, and basic open-domain generation. This initial architecture was supported by core diagnostic tools, on-screen speech transcription, and perceptual modules (e.g., person tracking, facial recognition).
- Wave 2 (September–December 2023) introduced a redesigned conversational module centered on a LLM architecture. Specifically, the system was upgraded to integrate Vicuna-13B-v1.5, a 13-billion parameter LLM optimized for offline use. A custom prompt in French was developed to align with hospital-based use cases, ensuring context-sensitive, coherent interactions. Other system improvements included a refined transcription interface and more efficient vision module processing.

A detailed technical overview of these updates is available in SPRING Deliverable D1.6. [38]

Assessment Instruments

Demographic Data Collection

Basic demographic information was collected using a standardized paper-based form prior to the start of the interaction session. Participants were asked to report their age, gender, level of education (years of formal education) and global level of cognitive functioning (MMSE).

System Performance

To assess system performance all interactions were recorded and annotated based on robot behavior. Each interaction was defined as a conversational unit, beginning when the participant initiated a verbal input and ending when the robot responded, or failed to do so. For example, if a participant asked, *"Where is the restroom?"* and the robot replied, *"To your right as you exit this room,"* the exchange was considered complete, regardless of its duration. Participants could then initiate a new exchange, continue the dialogue, or end the session. This approach produced interactions of widely varying length, from brief question–answer sequences to longer, multi-turn conversations, depending on participant intent and conversational flow.

Robot behaviors were annotated using a predefined classification scheme consisting of four categories:

1. Comprehension failure – the robot fails to interpret or meaningfully respond to the user's input, typically resulting in silence or a fallback message (e.g., *"I don't understand"*).
2. Inappropriate verbal response – the robot produces a reply that is unrelated, incoherent, or socially inappropriate in context (e.g., To the question "*What time is my appointment?*", the robot replies "*Of course. The hospital is open every day from 9 a.m. to 5.30 p.m.*").
3. Incomplete utterances - The robot produces an unfinished or abruptly cut-off response, resulting in a message that lacks necessary information (e.g., For example, when asked "*Can you tell me where the toilets are?"*, the robot replies *"The toilets are on your left as you come out of..."*).
4. Technical error – system-level malfunctions such as speech synthesis failure, audio dropout, or interface freezing.

The first three robot behaviors, comprehension failures, inappropriate verbal responses, and incomplete utterances, were classified as *verbal-related errors.* This categorization allowed for exploratory correlation analyses between robot behavior and user engagement, enabling a distinction between technical malfunctions and conversational breakdowns.

Multiple error types could be assigned to a single interaction when applicable (e.g., a comprehension failure accompanied by a technical issue). This allowed for the analysis of both error frequencies and error co-occurrence patterns, as well as comparisons across the two experimental waves to examine changes in system performance over time.

Interaction Success

Interaction success was defined as the robot providing a relevant and coherent response that appropriately addressed the user's request or corresponded to the information provided.

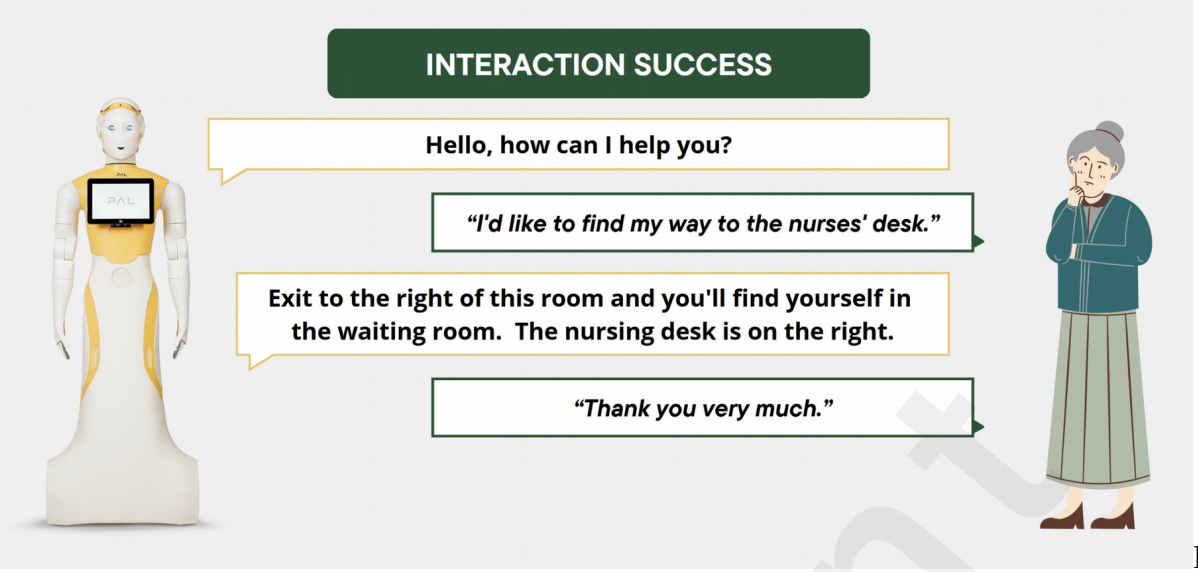


Figure 2.
Example of a successful interaction between the ARI robot and a participant

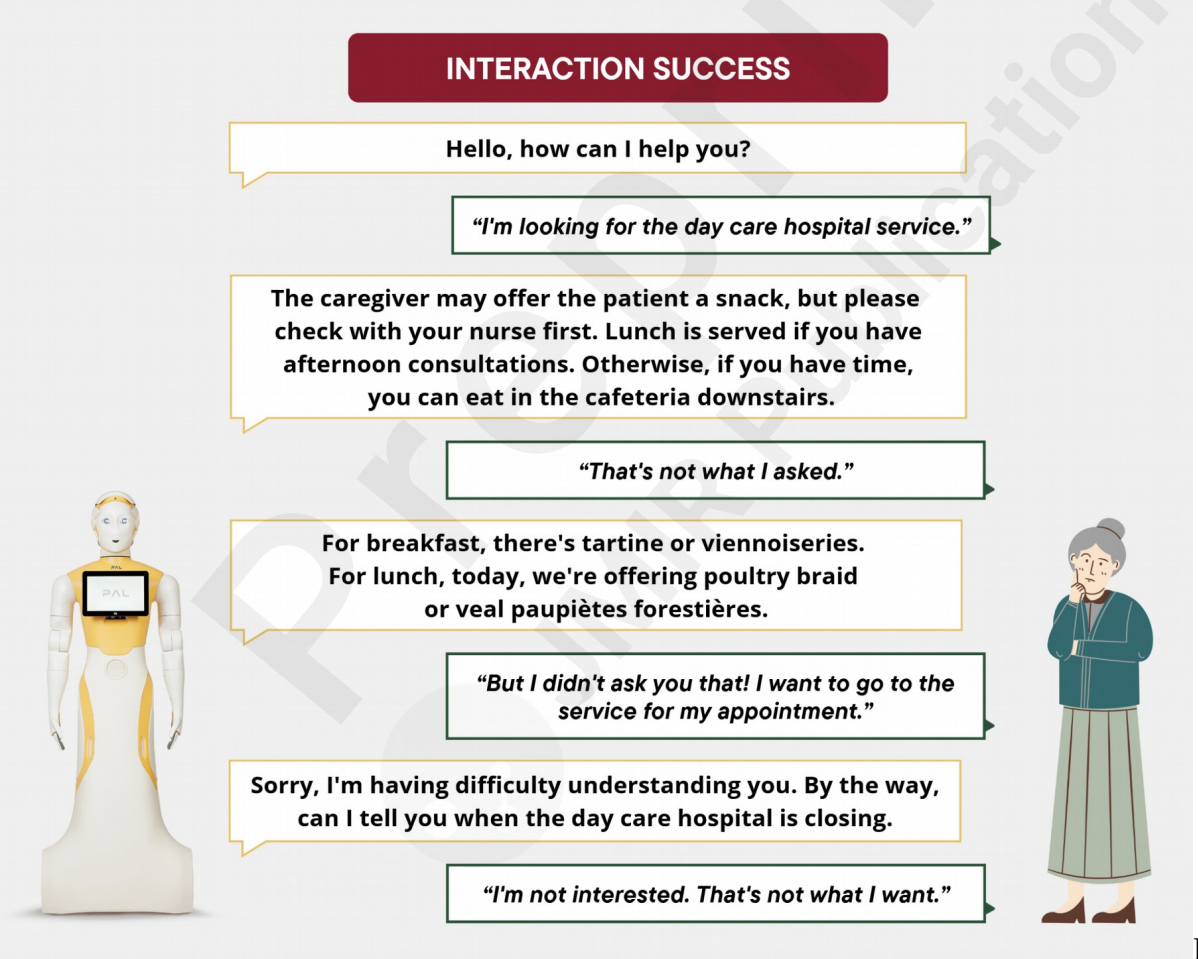


Figure 3.
Example of a failed interaction between the ARI robot and a participant

User Engagement Coding Framework

User engagement was assessed using a structured behavioral coding scheme (Table 1), developed to capture three dimensions: verbal, physical, and emotional engagement. The *verbal dimension* reflected speech production and conversational responsiveness; the *physical dimension* encompassed posture, gesture use, body orientation, and other forms of physical involvement; and the *emotional*

dimension captured visible affect and emotional expressivity. Each dimension was rated on a 5-point scale, ranging from 1 (minimal engagement or rejection) to 5 (high involvement and expressivity).

The coding grid was designed by the research team, grounded in prior work on affective robotics and social signal processing [39], and iteratively refined through preliminary testing to ensure clarity, consistency, and applicability in naturalistic HRI settings. It incorporates both affective and interactional markers, such as speech spontaneity, posture alignment, gesture expressiveness, and emotional display.

All interactions were independently coded by two trained researchers using a detailed coding manual to reduce subjectivity inherent in qualitative behavioral analysis [40]. In cases of disagreement, coders discussed the item to reach consensus; if consensus could not be achieved, a third researcher acted as an arbitrator to determine the final score. This multi-coder approach aligns with best practices in HRI research, which emphasize the use of multiple raters to mitigate individual bias and enhance intersubjective reliability [39, 41].

Emotional engagement scores were not assigned for 28 out of 130 interactions due to *limited facial visibility* caused by surgical masks, which impeded reliable interpretation of affective expressions. These interactions involved five participants, three who remained masked throughout and two who removed their masks partway. While excluded from emotional engagement analysis, these sessions were retained for all other measures.

In addition to the quantitative ratings, coders were encouraged to document any ambiguous, unexpected, or contextually meaningful behaviors using an open comment field. These qualitative annotations served to refine and contextualize the interpretation of engagement patterns observed in the coded data.

Table 1. User engagement dimensions and rating scale

Dimension	Score	Description
Verbal engagement	1	Persistent silence or clearly disengaged verbal behavior (e.g., hostile or rejecting remarks).
	2	Minimal verbal output (e.g., one-word replies such as “yes” or “no”), delivered in a flat tone, with no spontaneous elaboration or follow-up.
	3	Brief, functional responses that are appropriate and clearly articulated, but emotionally neutral and lacking initiative.
	4	Active participation through contextually appropriate answers and spontaneous follow-up contributions.
	5	Fluent, sustained conversation featuring regular verbal initiative, co-constructed dialogue, and spontaneous, personally meaningful comments or anecdotes.
Physical engagement	1	Defensive or avoidant posture (e.g., crossed limbs, leaning away), evasive gaze, and rejecting or dismissive gestures.
	2	Passive physical presence, with infrequent or minimal gestures and body oriented away from the robot.
	3	Upright, neutral posture with limited physical expressivity; gaze may be directed toward the robot but lacks clear engagement; few gestures.
	4	Open body orientation toward the robot, accompanied by

		expressive gestures and visibly responsive physical behavior (e.g., nodding, leaning in).
	5	Strong physical involvement, including spontaneous gestures, mimicry of the robot's actions, or voluntary physical contact (e.g., touching the robot).
Emotional engagement	1	Clear negative emotional expression (e.g., anger, frustration, disgust), often accompanied by signs of interaction breakdown or withdrawal.
	2	Very low emotional reactivity, flat facial affect, and visible signs of fatigue, boredom, or disinterest.
	3	Neutral but attentive facial expression, stable gaze toward the robot, with limited or absent emotional expressivity.
	4	Moderate social-affective signals, such as smiling, nodding, or short expressive reactions, indicating active but contained emotional engagement.
	5	Strong affective display, including laughter, animated facial or physical expressions, and moments of emotional synchrony with the robot.

Assessment Scales

Two standardized self-report questionnaires were administered immediately after the interaction to assess participants' perceptions of the robot's acceptability and usability.

The *Acceptability E-scale (AES)*, adapted from Heerink et al. [42] and translated into French by Micoulaud-Franchi et al. [43], measures perceived acceptability across six dimensions: trust, perceived usefulness, enjoyment, sociability, ease of use, and intention to use. The scale includes six items, each rated on a 5-point Likert scale, yielding a total score ranging from 6 to 30. AES scores above 25.81/30 are considered to reflect high perceived acceptability, while lower scores indicate limited acceptability. The adapted version of the scale used in this study is provided in Multimedia Appendix 1.

The System Usability Scale (SUS) [44] is a widely used, validated ten-item questionnaire designed to assess perceived usability. Items are rated on a 5-point Likert scale, and the final score is calculated according to standard scoring procedures, yielding a composite usability score between 0 and 100. According to standard interpretive guidelines, SUS scores below 50.9 are considered poor, scores between 51 and 71.4 are rated as OK to good, and scores above 71.4 reflect excellent perceived usability [45]. The adapted version of the scale used in this study is provided in Multimedia Appendix 1.

Assessment Procedure

Each session (~45 minutes) followed a standardized sequence consisting of a short reminder of the objectives of the study, informed consent, interaction with the robot, and post-interaction assessments. Participants were first introduced to the robot's general capabilities (e.g., hospital orientation, entertainment, appointment assistance) and informed that they could ask the robot questions freely related to these functions. No detailed interaction instructions or scripted prompts were provided, in order to preserve a naturalistic interaction context.

A fixed-position camera recorded the entire session for behavioral coding. A researcher was present throughout the entire session to provide assistance if needed, while minimizing interference.

Following the interaction, participants completed two standardized self-report questionnaires, the AES and the SUS, with support from the research team when necessary.

Data Analysis

Quantitative analyses were conducted using the JAMOVI software [46] and focused on four key dimensions: sociodemographic data, system performance, user engagement, and acceptability and usability.

For *sociodemographic characteristics* (age, gender, socio-educational level, and MMSE scores), descriptive statistics (mean, standard deviation, minimum, and maximum) were first computed. Comparisons across experimental waves were then conducted using the Mann–Whitney U test, due to violations of normality and homogeneity of variance assumptions, as assessed by Shapiro–Wilk and Levene's tests.

For the *analysis of HRIs*, the unit of analysis was the individual interaction. Interaction duration was calculated for each HRI session based on video recordings. Mean durations were computed separately for each experimental wave, and the total cumulative duration of all recorded interactions was calculated to quantify the overall volume of video data analyzed. The definition of an interaction used in this study is provided in the previous section, "System Performance."

A descriptive analysis was first conducted for each wave, covering the following elements: robot behavior categorized by four error types (comprehension failures, inappropriate verbal responses, incomplete responses, and technical errors) and the proportion of interactions with and without errors.

Subsequently, between-wave comparisons were performed. Robot behavior was compared across waves based on the average error rate per interaction. Analyses were first conducted considering all error types, followed by a focused analysis on the three verbal-related errors: comprehension failures, inappropriate verbal responses, and incomplete responses. These comparisons were conducted using Mann–Whitney U tests, due to non-normal distributions and violations of homogeneity of variance in the error frequency data.

User engagement was then analyzed using multimodal behavioral coding across three dimensions: verbal, physical, and emotional engagement. A descriptive analysis was first conducted for each dimension. Subsequently, engagement scores were compared between experimental waves using Kruskal–Wallis tests, due to violations of normality and homogeneity of variance assumptions.

Interaction success was operationalized as the completion of a communicative goal, specifically, when the user successfully obtained the information requested from the robot. Success rates were compared between experimental waves using a chi-squared test of independence, to assess whether system upgrades influenced the likelihood of a successful exchange.

The dimensions of *acceptability* and *usability,* assessed using the AES and SUS questionnaires respectively, were first examined through descriptive analysis. Subsequently, scores were compared between experimental waves using independent samples *t*-tests, following confirmation of normality (Shapiro–Wilk test) and homogeneity of variance (Levene's test).

To complement group-level comparisons, additional analyses were conducted to explore relationships between the previously defined dimensions of engagement, system performance, user experience, and participant characteristics. Pearson correlation analyses were used to examine

associations between user engagement scores and interaction outcomes (i.e., successful vs. unsuccessful interactions), as well as between engagement scores and perceived acceptability (AES) and usability (SUS). Correlation analyses were also performed on the full sample to investigate relationships between sociodemographic characteristics (age, socio-educational level, MMSE), system performance indicators (total and conversational error counts), user engagement dimensions (verbal, physical, emotional), and subjective evaluations of the robot (AES and SUS). These analyses aimed to determine whether individual differences were associated with user experience, robot performance, or behavioral engagement during interactions.

Finally, qualitative data from video recordings were analyzed thematically to identify recurring patterns and themes. This analysis was intended to complement the quantitative findings by providing deeper insights into user engagement and interaction dynamics [47].

Ethical Considerations

The study was approved by the French National Ethics Committee (“Comité de Protection des Personnes, CPP Ouest II, Maison de la Recherche Clinique – CHU Angers”; Institutional Review Board [IRB] 2021/20) and complied with the General Data Protection Regulation (GDPR). Data processing was registered with the Data Protection Officer (DPO) under reference number 20210114153645 in the AP-HP registry. The study did not involve randomization or clinical intervention. Informed consent was obtained from all participants on site, and they were informed that they could withdraw from the study at any time. The original consent included approval for secondary analyses without requiring additional consent. All participant data were anonymized, and no compensation was provided.

Results

Participant Characteristics and Baseline Equivalence

The sample included 28 OAs, with 10 participants in wave 1 and 18 in wave 2. The mean age was 78.2 years (SD=6.25; range: 67–93), and most participants were women (n=20). The average Mini-Mental State Examination (MMSE) score was 26.3 out of 30 (SD=3.73; range: 18–30), with scores ≥26 generally indicating normal cognitive function. The mean socio-educational level was 12.8 years (SD=1.94), ranging from 9 years (completion of lower secondary education or vocational training) to 14 or more years (university-level qualifications such as a bachelor's degree or higher).

Baseline equivalence between groups was assessed using Mann–Whitney U tests for MMSE scores and socio-educational level, as both variables were non-normally distributed. No significant differences were found between wave 1 and wave 2 for MMSE (U=49.5, P=.05) or socio-educational level (U=79.5, P=.57), supporting the comparability of the two experimental groups.

Table 2. Sociodemographic characteristics of participants by experimental wave

Experimental Wave (n)	**Gender** n (%)	**Age** years (± SD)	**Years of Education** years (± SD)	**MMSE** score (± SD)
Wave 1 10	M: 3 (20%); F: 7 (80%)	78.6 ± 7.6	13.1 ± 1.7	28.0 ± 1.9
Wave 2 18	M: 6 (33.3%); F: 12 (66.7%)	77.9 ± 5.6	12.6 ± 2.1	25.3 ± 4.1
Total 28	M: 9 (32.1%); F: 19 (67.9%)	78.2 ± 6.3	12.8 ± 1.9	26.3 ± 3.7

Note. *n*=number of participants. Gender is reported as the number and percentage of male (M) and female (F) participants. MMSE=Mini-Mental State Examination (range: 0–30); scores≥26 are generally considered within the normal cognitive range. Years of education refer to total years of formal schooling completed.

Interaction Duration and Data Set Composition

A total of 130 HRIs were analyzed (wave 1=36 HRIs; wave 2=94 HRIs), comprising the full dataset used for multimodal behavioral coding. Interaction durations ranged from 7 seconds to 3 minutes and 14 seconds, with a mean duration of 52 seconds (wave 1=40 seconds; wave 2=58 seconds). In total, the recordings resulted in 1 hour, 53 minutes, and 55 seconds of video data.

Figure 4. Experimental Context: Participant interacting with the SAR

Robot Error Patterns and System Performance Across Waves

Table 3 presents the distribution of robot error types across the two experimental waves. Comprehension failures were the most frequent error overall, occurring in 25.4% of all interactions, with a notably higher rate in wave 1 (47.2%) compared to wave 2 (17%). Incomplete utterances were absent in wave 1 but appeared in 14.9% of interactions in wave 2. Inappropriate verbal responses were also more common in wave 1 (36.1%) than in wave 2 (14.9%), contributing to a combined rate of 20.8%. Technical errors remained relatively stable between conditions, occurring in 11.1% of interactions in wave 1 and 11.7% in wave 2. For analysis, each error type was binary-coded per interaction (0=absent, 1=present).

Table 3. Distribution of Robot Error Types by Experimental Wave

Robots behaviors	Wave 1 n (%)	Wave 2 n (%)	Total n (%)
Comprehension failures	17 (47.2%)	16 (17%)	33 (25.4%)
Incomplete utterances	0 (0%)	14 (14.9%)	14 (10.8%)
Inappropriate verbal responses	13 (36.1%)	14 (14.9%)	27 (20.8%)
Technical error	4 (11.1%)	11 (11.7%)	15 (11.5%)

Robot error types were compared across the two experimental waves based on total error counts. Verbal-related errors, including comprehension failures (U=989, $P<.001$), incomplete utterances (U=1440, $P=.015$), and inappropriate responses (U=1333, $P=.008$), differed significantly between wave 1 and wave 2. A higher number of verbal-related errors was recorded in wave 1, reflecting reduced conversational performance in this condition and suggesting improved system behavior following the updates implemented in wave 2.

An additional analysis was conducted at the level of individual interactions to compare error rates between experimental waves. Results showed higher error rates per interaction in wave 1 than in wave 2: comprehension failures occurred on average 2.03 times per interaction in wave 1 compared to 0.61 in wave 2; inappropriate responses occurred at a rate of 0.36 in wave 1 versus 0.15 in wave 2; and incomplete utterances, which were absent in wave 1, appeared with a mean frequency of 0.16 per interaction in wave 2. In contrast, technical error rates did not differ significantly between waves (U=1682, $P=.93$), indicating stable hardware performance across conditions.

The proportion of interactions without any system errors increased significantly in wave 2 following the introduction of the LLM (U=995, $P<.001$), indicating improved system stability and interaction reliability after the update. As shown in Figure 5, only 27.8% of interactions in wave 1 were error-free, compared to 70.2% in wave 2. This substantial increase in error-free interactions suggests that the integration of the LLM contributed positively to the overall robustness and consistency of the system's performance.

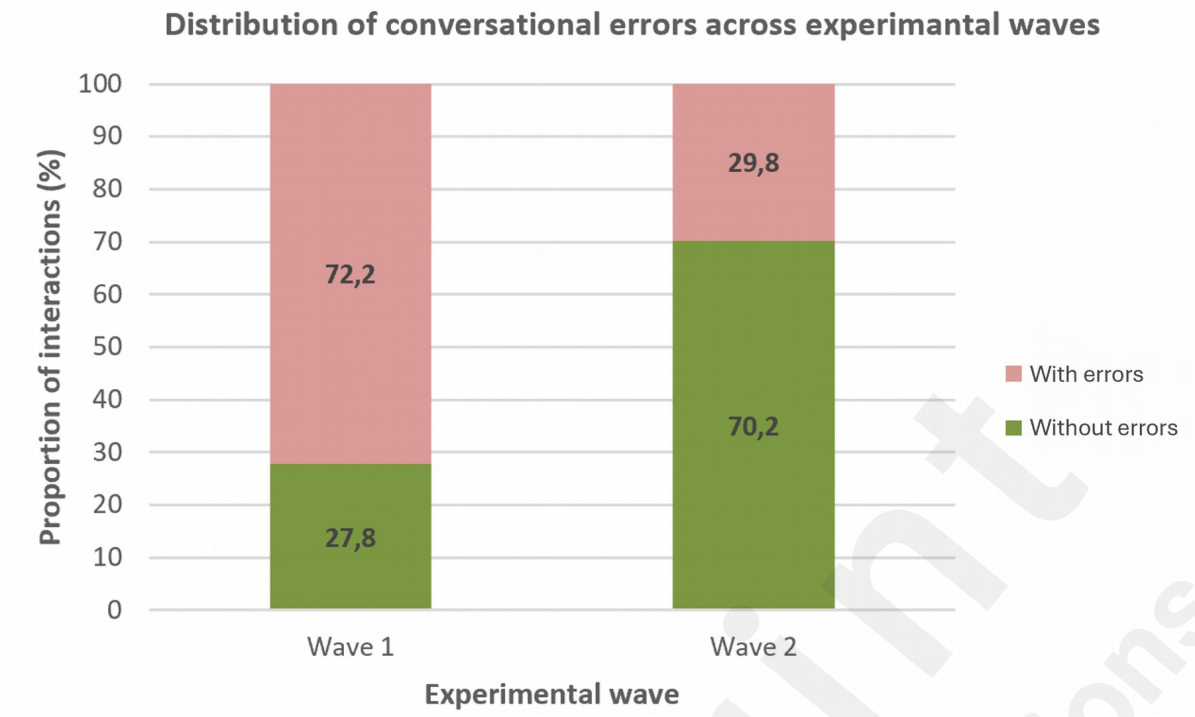


Figure 5. Proportion of error-free and problematic interactions by experimental wave.

User Engagement

Participant engagement was evaluated using multimodal behavioral coding across three dimensions: verbal, physical, and emotional. Each dimension was scored on a 5-point scale, with higher scores indicating greater levels of engagement.

The mean verbal engagement score was 3.64 (SD=0.64) in wave 1 and 3.47 (SD=0.77) in wave 2. A Kruskal–Wallis test revealed no significant difference in verbal engagement between the two waves ($\chi^2(1)=2.54$, $P=.11$).

For physical engagement, scores remained stable across waves, with a mean of 3.03 (SD=0.77) in wave 1 and 2.99 (SD=0.73) in wave 2. No significant difference was observed ($\chi^2(1)=0.063$, $P=.80$).

Emotional engagement showed a slight increase, from 2.92 (SD=1.02) in wave 1 to 3.33 (SD=0.86) in wave 2. No significant difference was observed ($\chi^2(1)=3.40$, $P=.06$). Descriptive statistics by wave are presented in Table 3.

Table 3. Comparison of acceptability and usability scores, user-engagement metrics and interaction success across waves

Dimension/ Wave	Wave 1	Wave 2	Difference
AES score out of 30 (±SD)	12.8 (4.58)	20.8 (6.52)	*P*<.001
SUS score out of 100 (±SD)	40.0 (24.04)	60.4 (23.11)	*P*<.001
Duration of interaction mean (±SD)	40 s (±32 s)	57 s (±32 s)	*P*<0.01
Verbal engagement mean score out of 5 (±SD)	3.64 (0.64)	3.47 (0.77)	*P*=0.24
Physical engagement mean score out of 5 (±SD)	3.03 (0.77)	2.99 (0.73)	*P*=0.79
Emotional engagement mean score out of 5 (±SD)	2.92 (1.02)	3.33 (0.86)	*P*<0.05
Interaction success % (n/N)	25% (9/36)	74.5% (70/94)	*P*<.001

Interaction Success and Perceived Acceptability and Usability

Interaction success, defined as the completion of a communicative goal without breakdown, improved significantly across waves: 25.0% in wave 1 (9/36) versus 74.5% in wave 2 (49/66), $\chi^2(1)=26.7$, *P*<.001.

Participants who experienced a successful interaction reported significantly higher ratings of both acceptability and usability. For the AES, mean scores were 21.6/30 (SD=5.84) in successful interactions, compared to 17.4/30 (SD=7.15) in failed ones (U=1261.5, *P*<.001). SUS scores followed the same pattern, with a mean of 63.3/100 (SD=22.5) for successful exchanges versus 51.1/100 (SD=26.5) for unsuccessful ones (U=1456.5, *P*=.01).

Importantly, no significant differences were found between participants who experienced successful versus unsuccessful interactions in terms of age (U=1898.5, *P*=.66), years of education (U=1915.5, *P*=.58), or MMSE scores (U=1921, *P*=.65), suggesting that interaction success was not dependent on participants' demographic or cognitive profiles.

Acceptability Assessment

In wave 1, acceptability scores ranged from 9 to 23, with a mean of 12.8 (SD=4.58). In wave 2, scores ranged from 7 to 30, with a mean of 20.8 (SD=6.52). The AES ranges from 6 to 30, with higher scores indicating greater acceptability; a commonly used threshold of 25.81 denotes high acceptability. An independent sample *t*-test revealed a significant difference between waves, $t(25)=-3.28$, *P*=.003, with higher acceptability observed in wave 2 compared to wave 1.

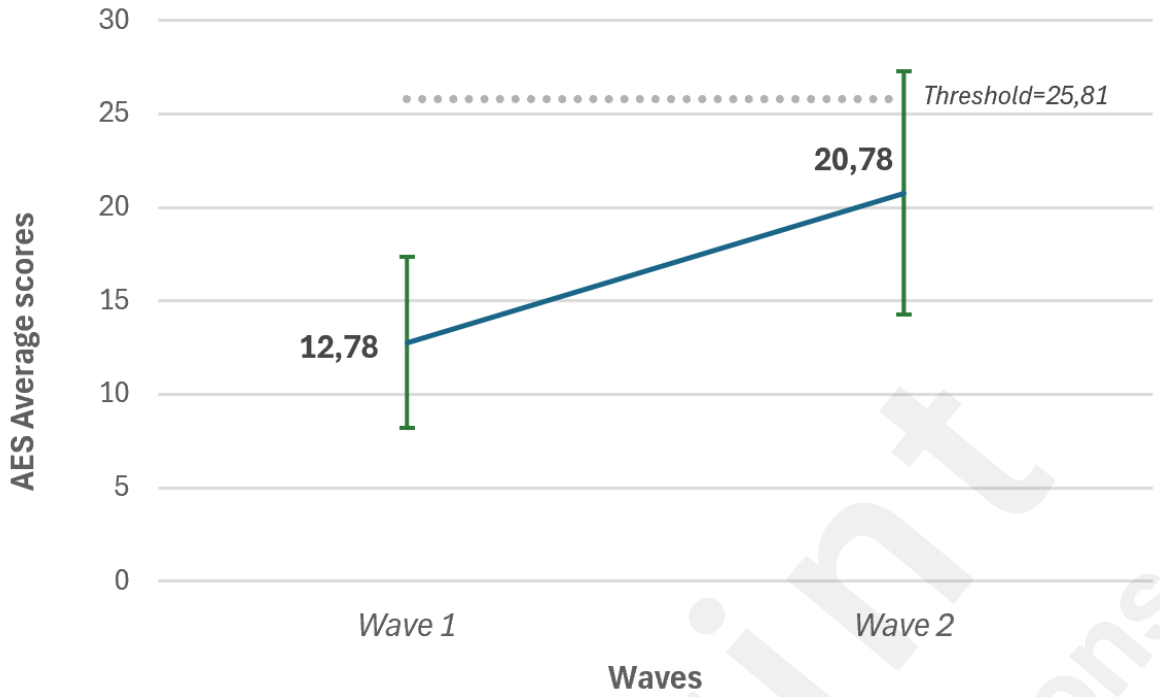


Figure 6.

AES scores across the two experimental waves

Usability Assessment

In wave 1, SUS scores ranged from 12.5 to 67.5, with a mean of 40.00 (SD=24.04). In wave 2, scores ranged from 2.5 to 92.5, with a mean of 60.42 (SD=23.11). The SUS ranges from 0 to 100, with 50.9 commonly cited as the minimum threshold for acceptable usability; scores below this indicate insufficient usability. An independent sample *t*-test revealed a significant difference between waves, *t*(25)=–2.14, *P*=.043, indicating greater perceived usability in wave 2 compared to wave 1.

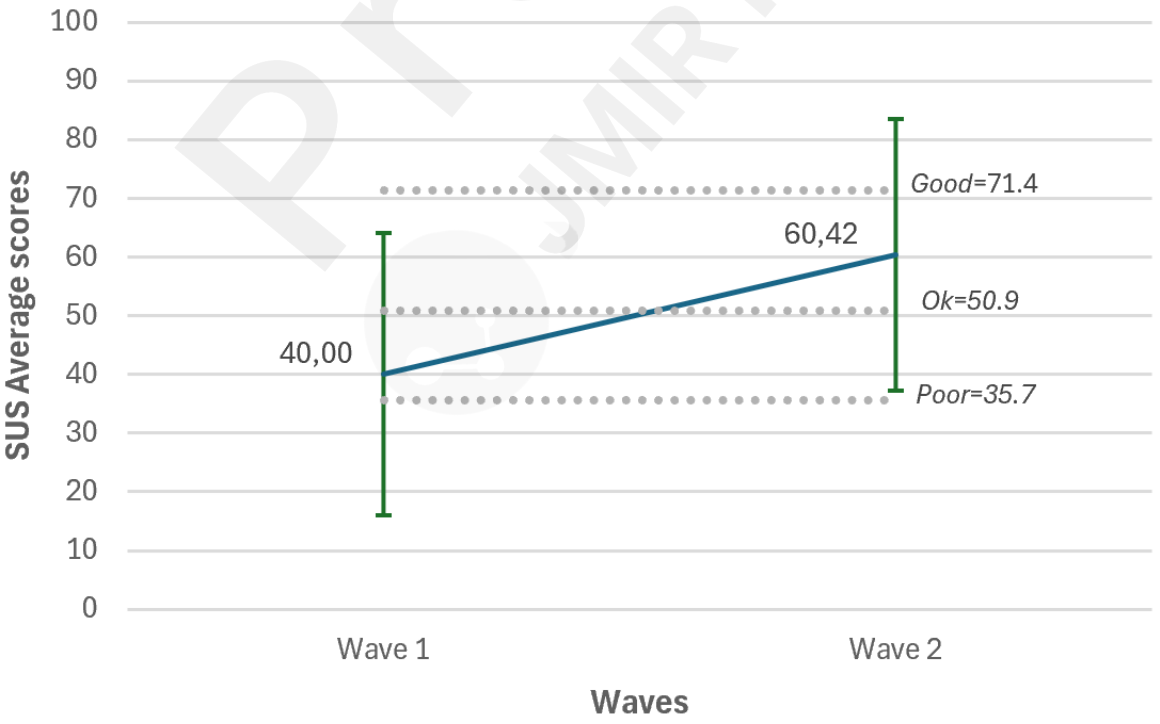


Figure 7.

SUS scores across the two experimental waves

Associations Between Engagement, Outcomes, and User Characteristics

A Pearson correlation analysis was conducted to explore the relationships between user engagement (verbal, physical, and emotional), interaction outcomes, system evaluations, and participant characteristics. Emotional and physical engagement were strongly positively correlated (r=0.63, *P*<.001), indicating that increased emotional expressiveness was closely associated with greater physical involvement. Emotional engagement was also positively associated with interaction success (r=0.28, *P*<.01), suggesting that successful interactions tended to elicit higher emotional responsiveness.

A moderate positive correlation was observed between physical and verbal engagement (r=0.23, *P*<.01), indicating that participants who were more physically expressive also tended to engage in more sustained or contextually responsive verbal behavior. Conversely, verbal engagement was negatively correlated with perceived acceptability (AES; r=–0.27, *P*<.01), suggesting that more frequent or proactive verbal contributions were associated with lower acceptability ratings.

Among participant characteristics, age showed a moderate negative correlation with physical engagement (r=–0.30, *P*<.001), indicating that older participants tended to be less physically expressive during interactions. Age was also negatively associated with acceptability ratings (r=–0.20, *P*<.05), suggesting slightly reduced system acceptability among older users. No significant correlations were found between cognitive functioning, as measured by MMSE scores, or socio-educational level (NC), and the other engagement or evaluation measures.

Qualitative Analysis of Spontaneous Verbalisations

During the interactions, participants occasionally produced spontaneous comments in reaction to specific events, most often triggered by the robot's conversational or technical errors. Captured in real time, these remarks provide insight into how OAs evaluated the exchange, revealing patterns of frustration, adaptation, humor, and expectation management. Thematic analysis identified six recurrent themes.

Silence, misunderstandings, and humor as coping

Moments when the robot failed to respond, misinterpreted input, or produced irrelevant answers often provoked visible reactions from participants. These ranged from mild irritation to playful humor, sometimes using sarcasm to mask disappointment.

> *"Well, yes, but it [the robot] is not answering me"* (wave 1)
> *"There, we asked it [the robot] a question [it] is not very good at."* (wave 1)
> *"[laughs] Wasn't that question planned? [in the robot database]"* (wave 1)
> *"Does it [the robot] have blocked ears? [laughs]"* (wave 1)
> *"Yet [the robot] understood (points to the screen with the transcript) but it [the robot] doesn't answer my question."* (wave 1)
> *"But I'm going to... (sign of slap) it [the robot]. [laughs]"* (wave 1)

Self-attribution of blame

In several cases, participants assumed that the communication breakdown was due to their own shortcomings rather than the robot's limitations. They speculated about their own speaking volume, speed, or enunciation, and even about environmental factors such as mask-wearing or hearing loss.

> *"I can't hear very well, as I have a lot of trouble with my ears."* (wave 1)
> *"It [the robot] is having trouble understanding me. Is it the mask?"* (wave 2)
> *"Do I speak fast enough [for the robot to understand]?"* (wave 1)
> *"Am I speaking loud enough [for the robot to understand]?* (wave 1)
> *"Ah! Because I have to ask a question? [the user did not understand how to interact with the*

robot] (wave 2)

Perceived repetition and circular conversations

In situations where the same phrases were repeated or where mutual understanding failed to progress, participants expressed the feeling of "going in circles". While some framed it with amusement, others voiced subtle impatience.

"We're going round in circles here. [laughs]" (wave 1)
"Shall I continue?" (wave 2)
"Wake up ARI! [because the robot did not answer to user's question]" (wave 1)

Expectation–response mismatch and and technology comparisons

Some participants noted a misalignment between what they expected the robot to do and the actual content or relevance of its responses. This sometimes led to dismissive judgments about its [the robot]'s role or capacity. Others drew parallels with earlier technological experiences, often highlighting how limitations persist despite progress.

"It [the robot] is a catering robot. It is not a robot for guiding patients." (wave 1)
"Doesn't it [the robot] understand anything? [the robot did not give an accurate answer]" (wave 1)
"We're not going to upset it [the robot]" [laughs] (wave 1)
"It [the robot] looks at me with strange eyes [the robot did not give an accurate answer]" (wave 1)
"It [the robot] reminds me of when they first introduced voice word processing. We had some absolutely incredible things written down. It didn't make any sense at all." (wave 2)

Seeking external guidance

When the conversation stalled, some participants looked to the experimenter for help, clarification, or reassurance about the "rules" of interaction. These moments highlighted uncertainty about the interaction boundaries and the available options.

"Come along please [to the experimenter]."(wave 2)
"How can I find out what questions I can ask the robot? [to the experimenter]" (wave 2)
"Do I have to talk to it [the robot]?" (wave 2)
"So now I can start again? Do I have to restart from zero or should I step out of the field and come back? [to the experimenter]" (wave 2)

Positive reactions to robot-initiated exchanges

Some comments reflected surprise or curiosity when the robot took the conversational initiative, suggesting that this behavior was unexpected or novel. In several cases, participants expressed positive reactions, noting that such behavior made the interaction feel more dynamic, personal, and engaging.

"It's impressive [the robot]!" (wave 2)
"Oh! It [the robot] talks to me first, just like a person!" (wave 2)
"So if I don't say anything, it [the robot] will keep asking me things?" (wave 2)

Discussion

This study evaluated the integration of a SAR into the routine activities of a geriatric hospital unit through a two-wave, real-world comparative design. In both waves, OAs interacted freely with the robot within its functional domains (e.g., hospital orientation, entertainment, appointment assistance), without scripted prompts, in order to preserve the spontaneity. Quantitative measures included system performance indicators (error types and rates), user engagement (verbal, physical, and emotional dimensions), interaction success, and subjective evaluations of acceptability and

usability; these were complemented by a thematic analysis of post-interaction interviews focusing on perceptions of robot errors.

Results showed that the introduction of a LLM between waves was associated with marked improvements in conversational performance, including a substantial reduction in verbal errors, an increase in error-free interactions, and higher rates of task success. These gains in system reliability were mirrored by higher acceptability (AES) and usability (SUS) ratings in the second wave. Qualitative analysis showed that participants' spontaneous reactions to the robot were shaped by both system performance and their own expectations. Breakdowns often prompted humor, self-blame, or requests for clarification, while successful or unexpected robot-initiated exchanges could elicit surprise and positive engagement. These findings suggest that emotional and behavioural responses in SAR–older adult interactions are influenced as much by expectation management and novelty as by technical reliability.

The following discussion interprets these findings in light of existing literature, examining how the observed changes in performance, engagement, and user perceptions contribute to current knowledge on SAR deployment in geriatric care and highlighting implications for future system design and implementation

LLM Integration and Conversational Reliability

The results suggest that the integration of the LLM, introduced between wave 1 (without LLM) and wave 2 (with LLM), enhanced the SAR's ability to process diverse user input and produce contextually relevant responses. This improvement was evident not only in lower rates of comprehension failures and incoherent replies but also in greater interactional fluidity, characteristics essential for sustaining user engagement and trust. The marked improvement in interaction success across waves (25.0% in wave 1 vs. 74.5% in wave 2) further underscores the practical benefits of LLM integration for enabling communicative goals to be achieved without breakdowns. These gains align with the observed reductions in verbal errors and the higher proportion of error-free interactions, indicating that technical reliability is closely tied to the system's capacity to support smooth exchanges. Importantly, successful interactions were associated with higher acceptability (AES) and usability (SUS) ratings, reinforcing prior evidence that perceived system competence is a key determinant of user satisfaction and trust in SARs

Such outcomes align with initial evidence from other domains, including customer service [48], museum guidance [49], and the use of social robots for education [50, 51], where LLMs have been reported to improve language comprehension, contextual adaptation, and response coherence [52]. Compared to rule-based or task-specific systems, systems equipped with LLM seem to offer more fluid, adaptive, and context-sensitive interactions [53, 54], which is consistent with our findings showing a significant improvement in interaction success and perceived acceptability and usability after LLM integration.

Recent work with OAs further demonstrates the potential of LLM-powered systems to support health and well-being in real-life care contexts. For example, conversational agents have been used in home environments to monitor safety risks, verify symptoms, and initiate alerts during emergencies [55]. Other studies have leveraged LLMs to collect richer health information with less provider effort [56] or to deliver cognitive stimulation through structured dialogue, resulting in improved task performance, increased social engagement, and high acceptance among OAs [57].

System and User Constraints in LLM-Enhanced OA–Robot Interaction

However, LLM-based applications, whether implemented in clinical or everyday settings involving OAs, also reveal limitations that can arise from both the system and the use. On the *system side*, latency from cloud-based speech recognition and LLM processing, as observed in Lima et al. [57], can interrupt conversational flow and disrupt turn-taking, sometimes being perceived as inattentiveness. These challenges are consistent with broader observations in spoken dialogue systems for robotics, where developers must balance trade-offs between accuracy and latency [58]. While cloud-based systems often deliver higher accuracy, they introduce delays that can break the rhythm of interaction; conversely, on-premises systems avoid latency but are typically less accurate and more limited in vocabulary. Furthermore, pre-trained recognizers are usually optimized on datasets that differ from the spontaneous, fragmentary, and context-dependent speech, common in real-world robot use, making adaptation resource-intensive. In our study, similar conversational breakdowns occurred in verbal-related errors, such as incomplete, irrelevant, or incoherent responses, which interrupted the exchange and occasionally led to repeated, circular interactions. We believe this might be due to a combination of several factors. First, even if we adapted the speech recognition model with French data, our on-premises solution has its limitations in terms of accuracy, and confuses some words. Second, even if the LLM solution operated on a partner's cloud with sufficient resources, it could not always provide accurate and appropriate answers, even when the speech recognition worked flawlessly. Third, the system worked in a half-duplex fashion, listening to the user only when the system was not speaking, which limited the fluidness of the interaction. Finally, even if these limitations are mild individually, their combination could cause misinterpretation on the robot's side, and generate user frustration, which did not ease the next steps of the interaction.

A further system-related limitation is that current LLMs are not specifically trained for interactions with OAs. As noted by Díaz et al. [59] and Chu et al. [60], speech and interaction data from OAs are scarce in AI training corpora, and existing datasets often contain age-related biases. As a result, characteristic features of OA communication, such as changes in prosody, vocabulary, conversational pacing, and the presence of hesitations or fragmented discourse in noisy environments, are underrepresented. This underrepresentation may reduce the ability of LLM-based systems to accurately interpret OA speech, anticipate their communicative needs, and adapt to their interaction style in real-world settings.

From the *user perspective*, common age-related factors such as slower speech tempo, pauses, and fluctuations in volume can pose difficulties for speech recognition, particularly in the presence of cognitive decline [61]. Beyond these physiological and cognitive aspects, there is also an expectation gap in how OAs perceive conversational systems. Mahmood et al. [62] describe how OAs often approach such technologies through the lens of human conversational norms, anticipating richer and more contextually adaptive exchanges than the system can actually deliver. This tendency is reinforced, as noted by Mahmood et al. [63] and Liu et al. [64], by two recurring influences: (a) the way many voice interfaces are designed and promoted to mirror human conversational styles, and (b) the inherent affordance of speech as an interaction modality, which can implicitly suggest that open-ended, human-like dialogue is possible when, in practice, the system's scope is more constrained and task-oriented.

These perceptions can be resistant to change, even with repeated exposure, and may lead to frustration when the robot's responses fall short of these implicit promises [65, 66]. In our study, this was evident in qualitative data: some participants persisted in treating the robot as a human interlocutor, even after repeated conversational errors. For example, one participant humorously asked, *"Does it [the robot] have blocked ears?"* (wave 1), while another, after receiving no reply, remarked, *"Well, yes, but it [the robot] is not answering me."* (wave 1). Others attributed

breakdowns to their own performance, asking, *"Am I speaking loud enough [for the robot to understand]?"* (wave 1) or *"Do I speak fast enough [for the robot to understand]?"* (wave 1). In some cases, repeated failures prompted disengagement or the need for external guidance, as illustrated by the query, *"How can I find out what questions I can ask the robot?"* (wave 2).

User engagement and its relationship with interaction outcomes

Across verbal, physical, and emotional dimensions, engagement scores in our study remained relatively stable between waves, with no statistically significant differences. This indicates that although LLM integration in wave 2 markedly improved interaction success, acceptability, and usability, these technical gains did not directly translate into measurable increases in overt engagement behaviors.

While emotional engagement showed a modest upward trend from wave 1 to wave 2, potentially reflecting more fluid exchanges with the LLM-enhanced system, this change did not reach statistical significance. Since participants in the two waves were different individuals, this pattern is unlikely to reflect individual-level adaptation over time. Instead, it may be linked to situational factors or differences in interaction dynamics. One possibility is that the novelty of directly engaging with a social robot can initially elicit a positive effect, as suggested by Naneva et al. [67], regardless of specific system performance. However, sustaining such engagement over time is more challenging, as robots that are not designed to cater to the needs and subjective experiences of OAs may limit both their engagement and the robot's sustainable use [25]. This aligns with findings from Hebesberger et al. [19] in long-term care contexts, where maintaining social relevance and personal meaning was key to continued interaction. Our findings revealed strong correlations between emotional and physical engagement, indicating that when users are emotionally responsive, they also tend to be more physically or behaviorally expressive. However, unlike system performance indicators, these engagement dimensions were not directly associated with interaction success. This partly contrasts with the work of Lin et al. [68], who found that emotional and behavioral engagement varied depending on the type of activity proposed by the SAR, although their study did not clarify whether these variations were influenced by the overall quality of the HRI. Further research is needed to examine more precisely how different dimensions of user engagement relate to interaction quality, and whether these patterns are consistent across varying activities and contexts.

Interestingly, verbal engagement was negatively correlated with acceptability, echoing findings by Mahmood et al. [63] that higher verbal activity in older adult–robot interactions can sometimes stem from the need to overcome conversational breakdowns rather than from enjoyment. In our context, this may indicate that some users spoke more in an attempt to repair misunderstandings, potentially lowering their subjective evaluations of the system.

Limitations of the study

This study was conducted in a single geriatric hospital unit, which limits the generalizability of the findings to other care contexts or to community-dwelling older adults. The relatively small sample size may have reduced the statistical power to detect certain associations, particularly regarding user characteristics.

While the two-wave design enabled a comparison before and after LLM integration, the participants in wave 1 and wave 2 were not the same individuals. As such, we cannot fully disentangle the impact of the LLM from potential differences in participant characteristics, hospital activity, or patient case mix between the two periods. A within-subject design, where the same participants assess the system before and after LLM integration, would allow for a more precise evaluation of the LLM's

contribution to observed improvements.

In addition, this study focused on stationary, face-to-face interactions and did not explore the potential contribution of robot mobility (e.g., autonomous navigation) or expressive gestures to user engagement. Prior research has shown that these dimensions can significantly enhance social presence and interactional involvement in SAR–older adult encounters, particularly in stimulating physical engagement and sustained attention.

Finally, although the mean interaction duration (52 seconds) provided a valuable snapshot of spontaneous use, the short time frame limits conclusions about long-term engagement and sustained acceptance. Longitudinal deployments are needed to capture how perceptions, trust, and interaction patterns evolve over extended use.

Recommendations for Future Development and Implementation of Conversational SARs Using LLMs

This study highlights the potential of conversational SARs powered by LLMs to substantially improve conversational reliability with OAs. Future development should build on these gains while addressing key limitations and contextual needs observed in our findings.

Adapt AI Architectures for OA Use, Including Future VLM and VLAM Applications

Beyond LLMs, future conversational SARs are likely to incorporate vision–language models (VLMs) and vision–language–action models (VLAMs), enabling richer multimodal understanding and more context-aware behaviors. However, all such models should be specifically adapted for OAs, particularly those with cognitive impairment, whose speech patterns, vocabulary, cognitive processing, and sensory capabilities differ from the datasets typically used to train general-purpose systems. Responsible adaptation must comply with regulatory and ethical frameworks to ensure safe, equitable, and trustworthy use in care settings.

Address the Limited Prior Experience of OAs with Conversational SARs Through Structured Familiarization Periods

Many OAs, particularly those without prior exposure to conversational agents, may require time and guided support to understand the robot's capabilities, limitations, and "rules of engagement." Future implementations should include a facilitator-led familiarization phase, where OAs can safely explore functions, interaction styles, and question formats without performance pressure. This early stage can build confidence, reduce uncertainty, and lay the groundwork for more meaningful and sustained engagement once independent use begins.

Strengthen the Facilitator's Role During Familiarization

For both novice and cognitively impaired OAs, a human facilitator can play a critical role in bridging early HRI challenges. This includes demonstrating effective input styles, explaining the robot's capabilities and limitations, and supporting OAs during breakdowns. Incorporating a structured familiarization period may accelerate learning, reduce frustration, and foster more positive first impressions.

Implement Adaptive Error-Handling Strategies

Conversational SARs should detect and repair interaction breakdowns early, such as prolonged silences, repeated input, or off-topic responses, by offering clarifying prompts, rephrased questions, or multimodal alternatives (e.g., visual selection, touch inputs). This can maintain engagement and reduce the negative emotional impact of errors.

Enhance HRI Personalization Across Sensory, Cognitive, and Motivational Dimensions

Personalization should include tuning speech recognition thresholds, adjusting interaction pacing, adapting vocabulary complexity, and aligning content with individual preferences and abilities. For OAs with varying cognitive, sensory, and mobility profiles, this is critical for both accessibility and engagement.

Integrate proactive and context-aware conversational strategies

Initiative-taking behaviors, such as relevant follow-up questions, personalized activity suggestions, and reference to prior interactions, can help sustain OA interest. Care should be taken to ensure such behaviors remain contextually appropriate and non-intrusive.

Evaluate future systems with multi-dimensional frameworks

Assessment should combine technical measures (e.g., error rates, latency), behavioral indicators (e.g., engagement scores, turn-taking dynamics), and experiential outcomes (e.g., trust, perceived usefulness). Such an approach ensures that improvements target both system performance and user experience in real-world contexts.

Conclusions

In this real-world, two-wave study in a geriatric hospital day care unit, integrating an LLM into a conversational SAR improved interaction success, conversation quality, and perceived system performance. Comparing a baseline dialogue system (Wave 1) with an LLM-enhanced system (Wave 2) showed that advanced AI can produce more coherent, contextually relevant exchanges. These improvements were reflected in higher acceptability and usability ratings from OAs. Conducting the study in a naturalistic care environment was key to capturing spontaneous interactions and real-world constraints, factors often absent in laboratory settings.

Our findings confirm the potential of LLMs to enhance HRI in care contexts. However, they also highlight the need to better understand how interaction success relates to engagement quality and to user sociodemographic characteristics. To maximize these benefits, future systems should be tuned to OA speech, cognitive, and sensory profiles. Strategies such as structured familiarization periods, adaptive error recovery, and personalized interactions could help maintain engagement over time. Aligning technical capabilities with the needs and preferences of older adults will be essential for making conversational SARs trusted and effective tools for geriatric care.

Acknowledgements

The authors confirm contribution to the paper as follows: study conception and design: LB, XAP, A-SR, MP; data collection: LB; analysis and interpretation of results: LB, JC; draft manuscript

preparation: LB, MP, A-SR. All authors reviewed the results and approved the final version of the manuscript.

This research was funded by the EU H2020 program under grant agreement no. 871245 [69].

The authors would like to sincerely thank all the participants who took part in this study, as well as all the health care professionals whose involvement contributed to the successful conduct of the SPRING project.

Conflicts of Interest

None declared

Abbreviations

AES: acceptability e-scale
AI: artificial intelligence
LLM: large language model
MMSE: mini-mental state examination
OA: older adult(s)
SAR: socially assistive robot(s)
SUS: system usability scale

Data Availability

Requests to access the datasets should be sent to the corresponding author.

References

1. Ageing and health. Available from: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed Aug 12, 2025]
2. Loukissa D. Understanding and Addressing Aggressive and Related Challenging Behaviors in Individuals with Dementia. AJH 2018 Dec 1;5(4):245–264. doi: 10.30958/ajh.5-4-1
3. Abley C. Developing a dementia care leaders' toolkit for older patients with cognitive impairment. Nursing Older People; Available from: https://journals.rcni.com/nursing-older-people/evidence-and-practice/developing-a-dementia-care-leaders-toolkit-for-older-patients-with-cognitive-impairment-nop.2020.e1250/abs [accessed Aug 12, 2025]
4. Hendlmeier I, Bickel H, Heßler-Kaufmann JB, Schäufele M. Care challenges in older general hospital patients. Z Gerontol Geriat 2019 Nov 1;52(4):212–221. doi: 10.1007/s00391-019-01628-x
5. Feil-Seifer D, Mataric MJ. Defining socially assistive robotics. 9th International Conference on Rehabilitation Robotics, 2005 ICORR 2005 2005. p. 465–468. doi: 10.1109/ICORR.2005.1501143
6. Tapus A, MATARIC´ AJ, Scassellati B. Socially assistive robotics [Grand Challenges of Robotics]. IEEE Robotics & Automation Magazine 2007; Available from: https://www.semanticscholar.org/paper/Socially-assistive-robotics-%5BGrand-Challenges-of-Tapus-MATARIC%C2%B4/6b29199b5f7e796f7f266cd6e957fadb4494e0e3 [accessed Aug 12, 2025]
7. Hsu T-C, Hsieh C-J. Socially Assistive Robots Assisting Older Adults in an Internet and Smart Healthcare Era: A Literature Review. 2022 IEEE 4th Eurasia Conference on IOT,

Communication and Engineering (ECICE) 2022. p. 591–594. doi: 10.1109/ECICE55674.2022.10042947
8. Abdollahi H, Mahoor MH, Zandie R, Siewierski J, Qualls SH. Artificial Emotional Intelligence in Socially Assistive Robots for Older Adults: A Pilot Study. IEEE Transactions on Affective Computing 2023 Jul;14(3):2020–2032. doi: 10.1109/TAFFC.2022.3143803
9. Dosso JA, Bandari E, Malhotra A, Guerra GK, Hoey J, Michaud F, Prescott TJ, Robillard JM. User perspectives on emotionally aligned social robots for older adults and persons living with dementia. Journal of Rehabilitation and Assistive Technologies Engineering SAGE Publications Ltd STM; 2022 Jun 1;9:20556683221108364. doi: 10.1177/20556683221108364
10. Rigaud A-S, Dacunha S, Harzo C, Lenoir H, Sfeir I, Piccoli M, Pino M. Implementation of socially assistive robots in geriatric care institutions: Healthcare professionals' perspectives and identification of facilitating factors and barriers. Journal of rehabilitation and assistive technologies engineering 2024 Sep 20;11:20556683241284765. doi: 10.1177/20556683241284765
11. Aymerich-Franch L, Ferrer I. Socially Assistive Robots' Deployment in Healthcare Settings: A Global Perspective. Int J Human Robot World Scientific Publishing Co.; 2023 Feb;20(01):2350002. doi: 10.1142/S0219843623500020
12. Graf B, Eckstein J. Service Robots and Automation for the Disabled and Nursing Home Care. In: Nof SY, editor. Springer Handbook of Automation Cham: Springer International Publishing; 2023. p. 1331–1347. doi: 10.1007/978-3-030-96729-1_62ISBN:978-3-030-96729-1
13. Sutherland CJ, Ahn BK, Brown B, Lim J, Johanson DL, Broadbent E, MacDonald BA, Ahn HS. The Doctor will See You Now: Could a Robot Be a medical Receptionist? 2019 International Conference on Robotics and Automation (ICRA) 2019. p. 4310–4316. doi: 10.1109/ICRA.2019.8794439
14. Babu A, K. S, Babu A, K. S. Empowering Elderly Care with Social Robots. rrrj 2024 Dec 6;3(2):410–423. Available from: https://irojournals.com/rrrj/article/view/3/2/8 [accessed Aug 12, 2025]
15. González-González CS, Violant-Holz V, Gil-Iranzo RM. Social Robots in Hospitals: A Systematic Review. Applied Sciences Multidisciplinary Digital Publishing Institute; 2021 Jan;11(13):5976. doi: 10.3390/app11135976
16. Romero-Garcés A, Bandera JP, Marfil R, González-García M, Bandera A. CLARA: Building a Socially Assistive Robot to Interact with Elderly People. Designs Multidisciplinary Digital Publishing Institute; 2022 Dec;6(6):125. doi: 10.3390/designs6060125
17. Wong KLY, Hung L, Wong J, Park J, Alfares H, Zhao Y, Mousavinejad A, Soni A, Zhao H. Adoption of Artificial Intelligence–Enabled Robots in Long-Term Care Homes by Health Care Providers: Scoping Review. JMIR Aging 2024 Aug 27;7(1):e55257. doi: 10.2196/55257
18. Huisman C, Kort H. Two-Year Use of Care Robot Zora in Dutch Nursing Homes: An Evaluation Study. Healthcare (Basel) 2019 Feb 19;7(1):31. PMID:30791489
19. Hebesberger D, Koertner T, Gisinger C, Pripfl J. A Long-Term Autonomous Robot at a Care Hospital: A Mixed Methods Study on Social Acceptance and Experiences of Staff and Older Adults. International Journal of Social Robotics 2017 Jun 1;9. doi: 10.1007/s12369-016-0391-6
20. Ghafurian M, Hoey J, Dautenhahn K. Social Robots for the Care of Persons with Dementia: A Systematic Review. J Hum-Robot Interact 2021 Dec 31;10(4):1–31. doi: 10.1145/3469653

21. Lima MR, Wairagkar M, Gupta M, Rodriguez Y Baena F, Barnaghi P, Sharp DJ, Vaidyanathan R. Conversational Affective Social Robots for Ageing and Dementia Support. IEEE Trans Cogn Dev Syst 2022 Dec;14(4):1378–1397. doi: 10.1109/TCDS.2021.3115228
22. Wald S, Puthuveetil K, Erickson Z. Do Mistakes Matter? Comparing Trust Responses of Different Age Groups to Errors Made by Physically Assistive Robots. arXiv; 2024. doi: 10.48550/arXiv.2408.13153
23. Giorgi I, Tirotto FA, Hagen O, Aider F, Gianni M, Palomino M, Masala G. Friendly But Faulty: A Pilot Study on the Perceived Trust of Older Adults in a Social Robot. IEEE Access IEEE Explore; 2022 Aug 29;10:92084–92096. doi: 10.1109/access.2022.3202942
24. Kim S. Exploring How Older Adults Use a Smart Speaker-Based Voice Assistant in Their First Interactions: Qualitative Study. JMIR Mhealth Uhealth 2021 Jan 13;9(1):e20427. PMID:33439130
25. Khosla R, Chu M-T, Khaksar SMS, Nguyen K, Nishida T. Engagement and experience of older people with socially assistive robots in home care. Assist Technol 2021 Mar 4;33(2):57–71. PMID:31063044
26. Miyake T, Wang Y, Yang P chu, Sugano S. Feasibility Study on Parameter Adjustment for a Humanoid Using LLM Tailoring Physical Care: 15th International Conference on Social Robotics, ICSR 2023. Ali AA, Cabibihan J-J, Meskin N, Rossi S, Jiang W, He H, Ge SS, editors. Social Robotics - 15th International Conference, ICSR 2023, Proceedings Springer Science and Business Media Deutschland GmbH; 2024;230–243. doi: 10.1007/978-981-99-8715-3_20
27. Esteban-Lozano I, Castro-González Á, Martínez P. Using a LLM-Based Conversational Agent in the Social Robot Mini. In: Degen H, Ntoa S, editors. Artificial Intelligence in HCI Cham: Springer Nature Switzerland; 2024. p. 15–26. doi: 10.1007/978-3-031-60615-1_2
28. Pinto-Bernal M, Biondina M, Belpaeme T. Designing Social Robots with LLMs for Engaging Human Interaction. Available from: https://www.mdpi.com/2076-3417/15/11/6377 [accessed Aug 13, 2025]
29. Oertel C, Castellano G, Chetouani M, Nasir J, Obaid M, Pelachaud C, Peters C. Engagement in Human-Agent Interaction: An Overview. Front Robot AI Frontiers; 2020 Aug 4;7. doi: 10.3389/frobt.2020.00092
30. Del Duchetto F, Baxter P, Hanheide M. Are You Still With Me? Continuous Engagement Assessment From a Robot's Point of View. Front Robot AI Frontiers; 2020 Sep 16;7. doi: 10.3389/frobt.2020.00116
31. Rossi S, Conti D, Garramone F, Santangelo G, Staffa M, Varrasi S, Di Nuovo A. The Role of Personality Factors and Empathy in the Acceptance and Performance of a Social Robot for Psychometric Evaluations. Robotics Multidisciplinary Digital Publishing Institute; 2020 Jun;9(2):39. doi: 10.3390/robotics9020039
32. Kuo IH, Rabindran JM, Broadbent E, Lee YI, Kerse N, Stafford RMQ, MacDonald BA. Age and gender factors in user acceptance of healthcare robots. RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication 2009. p. 214–219. doi: 10.1109/ROMAN.2009.5326292
33. Flandorfer P. Population Ageing and Socially Assistive Robots for Elderly Persons: The Importance of Sociodemographic Factors for User Acceptance. International Journal of Population Research 2012;2012(1):829835. doi: 10.1155/2012/829835

34. Lee SH, Kim JS, Yu S. The impact of care robots on older adults: A systematic review. Geriatric Nursing 2025 Sep 1;65:103507. doi: 10.1016/j.gerinurse.2025.103507
35. Folstein MF, Folstein SE, McHugh PR. “Mini-mental state”. A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975 Nov;12(3):189–198. PMID:1202204
36. Alameda-Pineda X, Addlesee A, García DH, Reinke C, Arias S, Arrigoni F, Auternaud A, Blavette L, Beyan C, Camara LG, Cohen O, Conti A, Dacunha S, Dondrup C, Ellinson Y, Ferro F, Gannot S, Gras F, Gunson N, Horaud R, D’Incà M, Kimouche I, Lemaignan S, Lemon O, Liotard C, Marchionni L, Moradi M, Pajdla T, Pino M, Polic M, Py M, Rado A, Ren B, Ricci E, Rigaud A-S, Rota P, Romeo M, Sebe N, Sieińska W, Tandeitnik P, Tonini F, Turro N, Wintz T, Yu Y. Socially Pertinent Robots in Gerontological Healthcare. arXiv; 2025. doi: 10.48550/arXiv.2404.07560
37. Blavette L, Dacunha S, Alameda-Pineda X, Hernández García D, Gannot S, Gras F, Gunson N, Lemaignan S, Polic M, Tandeitnik P, Tonini F, Rigaud A-S, Pino M. Acceptability and Usability of a Socially Assistive Robot Integrated With a Large Language Model for Enhanced Human-Robot Interaction in a Geriatric Care Institution: Mixed Methods Evaluation. JMIR Hum Factors 2025 Aug 1;12:e76496. PMID:40750072
38. SPRING D1.6 – user feedback from the final validation in relevant environments. SPRING Consortium. 2024. URL: https://spring-h2020.eu/wp-content/uploads/2024/07/SPRING_D1.6_-User-feedback-from-the-final-validation-relevant-environments_vFinal_31.05.2024.pdf [accessed 2025-08-13]
39. Anzalone S, Boucenna S, Ivaldi S, Chetouani M. Evaluating the Engagement with Social Robots. International Journal of Social Robotics 2015 Aug 29;7. doi: 10.1007/s12369-015-0298-7
40. Hemmler VL, Kenney AW, Langley SD, Callahan CM, Gubbins EJ, Holder S. Beyond a coefficient: an interactive process for achieving inter-rater consistency in qualitative coding. Qualitative Research 2022 Apr;22(2):194–219. doi: 10.1177/1468794120976072
41. Indurkhya X, Venture G. Lessons in Developing a Behavioral Coding Protocol to Analyze In-the-Wild Child–Robot Interaction Events and Experiments. Electronics Multidisciplinary Digital Publishing Institute; 2024 Jan;13(7):1175. doi: 10.3390/electronics13071175
42. Heerink M, Krose B, Evers V, Wielinga B. The Influence of a Robot’s Social Abilities on Acceptance by Elderly Users. ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication 2006. p. 521–526. doi: 10.1109/ROMAN.2006.314442
43. Micoulaud-Franchi J-A, Sauteraud A, Olive J, Sagaspe P, Bioulac S, Philip P. Validation of the French version of the Acceptability E-scale (AES) for mental E-health systems. Psychiatry Res 2016 Mar 30;237:196–200. PMID:26809367
44. Bangor A, Kortum PT, Miller JT. An Empirical Evaluation of the System Usability Scale. International Journal of Human–Computer Interaction Taylor & Francis; 2008 Jul 29;24(6):574–594. doi: 10.1080/10447310802205776
45. Bangor A, Kortum P, Miller J. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale - JUX. JUX - The Journal of User Experience. 2009. Available from: https://uxpajournal.org/determining-what-individual-sus-scores-mean-adding-an-adjective-rating-

scale/ [accessed Aug 12, 2025]
46. Jamovi - open statistical software for the desktop and cloud. Available from: https://www.jamovi.org/ [accessed Aug 12, 2025]
47. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology Routledge; 2006 Jan 1;3(2):77–101. doi: 10.1191/1478088706qp063oa
48. Peddinti SR, Katragadda SR, Pandey BK, Tanikonda A. Utilizing Large Language Models for Advanced Service Management: Potential Applications and Operational Challenges. 2023 Mar 12; Available from: https://thesciencebrigade.com/jst/article/view/517 [accessed Aug 12, 2025]
49. Vasic I, Fill H-G, Quattrini R, Pierdicca R. LLM-Aided Museum Guide: Personalized Tours Based on User Preferences. Extended Reality: International Conference, XR Salento 2024, Lecce, Italy, September 4–7, 2024, Proceedings, Part III Berlin, Heidelberg: Springer-Verlag; 2024. p. 249–262. doi: 10.1007/978-3-031-71710-9_18
50. Li, C, Xing, W, Leite, W. Building socially responsible conversational agents using big data to support online learning: A case with Algebra Nation - Li - 2022 - British Journal of Educational Technology - Wiley Online Library. Available from: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13227 [accessed Aug 12, 2025]
51. Voultsiou E, Vrochidou E, Moussiades L, Papakostas GA. The potential of Large Language Models for social robots in special education. Prog Artif Intell 2025 Jun 1;14(2):165–189. doi: 10.1007/s13748-025-00363-2
52. Chen J, Liu Z, Huang X, Wu C, Liu Q, Jiang G, Pu Y, Lei Y, Chen X, Wang X, Zheng K, Lian D, Chen E. When large language models meet personalization: perspectives of challenges and opportunities. World Wide Web 2024 Jun 28;27(4):42. doi: 10.1007/s11280-024-01276-1
53. Wang C. Comparing Rule-based and LLM-based Methods to Enable Active Robot Assistant Conversations. Available from: https://www.semanticscholar.org/paper/Comparing-Rule-based-and-LLM-based-Methods-to-Robot-Wang/54bffd8096740732ce4bf6486b513a261f05e313 [accessed Aug 12, 2025]
54. Kim CY, Lee CP, Mutlu B. Understanding Large-Language Model (LLM)-powered Human-Robot Interaction. Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction New York, NY, USA: Association for Computing Machinery; 2024. p. 371–380. doi: 10.1145/3610977.3634966
55. Lima M, Su T, Jouaiti M, Wairagkar M, Malhotra P, Soreq E, Barnaghi P, Vaidyanathan R. Discovering Behavioral Patterns Using Conversational Technology for In-Home Health and Well-Being Monitoring. IEEE Internet of Things Journal 2023 Nov 1;PP:1–1. doi: 10.1109/JIOT.2023.3290833
56. Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults | Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. Available from: https://dl.acm.org/doi/10.1145/3659625 [accessed Aug 12, 2025]
57. Lima M, O'Connell A, Zhou F, Nagahara A, Hulyalkar A, Deshpande A, Thomason J, Vaidyanathan R, Matarić M. Promoting Cognitive Health in Elder Care with Large Language Model-Powered Socially Assistive Robots. 2025. p. 22. doi: 10.1145/3706598.3713582
58. Marge M, Espy-Wilson C, Ward NG, Alwan A, Artzi Y, Bansal M, Blankenship G, Chai J, Daumé H, Dey D, Harper M, Howard T, Kennington C, Kruijff-Korbayová I, Manocha D, Matuszek C, Mead R, Mooney R, Moore RK, Ostendorf M, Pon-Barry H, Rudnicky AI, Scheutz

M, Amant RSt, Sun T, Tellex S, Traum D, Yu Z. Spoken language interaction with robots: Recommendations for future research. Computer Speech & Language 2022 Jan 1;71:101255. doi: 10.1016/j.csl.2021.101255

59. Diaz M, Johnson I, Lazar A, Piper AM, Gergle D. Addressing Age-Related Bias in Sentiment Analysis. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems New York, NY, USA: Association for Computing Machinery; 2018. p. 1–14. doi: 10.1145/3173574.3173986

60. Chu CH, Nyrup R, Leslie K, Shi J, Bianchi A, Lyn A, McNicholl M, Khan S, Rahimi S, Grenier A. Digital Ageism: Challenges and Opportunities in Artificial Intelligence for Older Adults. Gerontologist 2022 Jan 20;62(7):947–955. PMID:35048111

61. Kobayashi M, Kosugi A, Takagi H, Nemoto M, Nemoto K, Arai T, Yamada Y. Effects of Age-Related Cognitive Decline on Elderly User Interactions with Voice-Based Dialogue Systems. In: Lamas D, Loizides F, Nacke L, Petrie H, Winckler M, Zaphiris P, editors. Human-Computer Interaction – INTERACT 2019 Cham: Springer International Publishing; 2019. p. 53–74. doi: 10.1007/978-3-030-29390-1_4

62. Mahmood A, Cao S, Stiber M, Antony VN, Huang C-M. Voice Assistants for Health Self-Management: Designing for and with Older Adults. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems New York, NY, USA: Association for Computing Machinery; 2025. p. 1–22. doi: 10.1145/3706598.3713839

63. Mahmood A, Wang J, Huang C-M. Situated Understanding of Errors in Older Adults' Interactions with Voice Assistants: A Month-Long, In-Home Study. arXiv; 2024. doi: 10.48550/arXiv.2403.02421

64. Liu Y-C, Chen C-H, Lin Y-S, Chen H-Y, Irianti D, Jen T-N, Yeh J-Y, Chiu SY-H. Design and Usability Evaluation of Mobile Voice-Added Food Reporting for Elderly People: Randomized Controlled Trial. JMIR mHealth and uHealth 2020 Sep 28;8(9):e20317. doi: 10.2196/20317

65. A Systematic Cross-Corpus Analysis of Human Reactions to Robot Conversational Failures | Proceedings of the 2021 International Conference on Multimodal Interaction. Available from: https://dl.acm.org/doi/10.1145/3462244.3479887 [accessed Aug 13, 2025]

66. Liu S, Parreira MT, Ju W. "I'm Done": Describing Human Reactions to Successive Robot Failure. 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI) 2025. p. 1458–1462. doi: 10.1109/HRI61500.2025.10974098

67. Naneva S, Sarda Gou M, Webb TL, Prescott TJ. A Systematic Review of Attitudes, Anxiety, Acceptance, and Trust Towards Social Robots. Int J of Soc Robotics 2020 Dec 1;12(6):1179–1201. doi: 10.1007/s12369-020-00659-4

68. Lin Y-C, Fan J, Tate JA, Sarkar N, Mion LC. Use of Robots to Encourage Social Engagement between Older Adults. Geriatr Nurs 2022;43:97–103. PMID:34847509

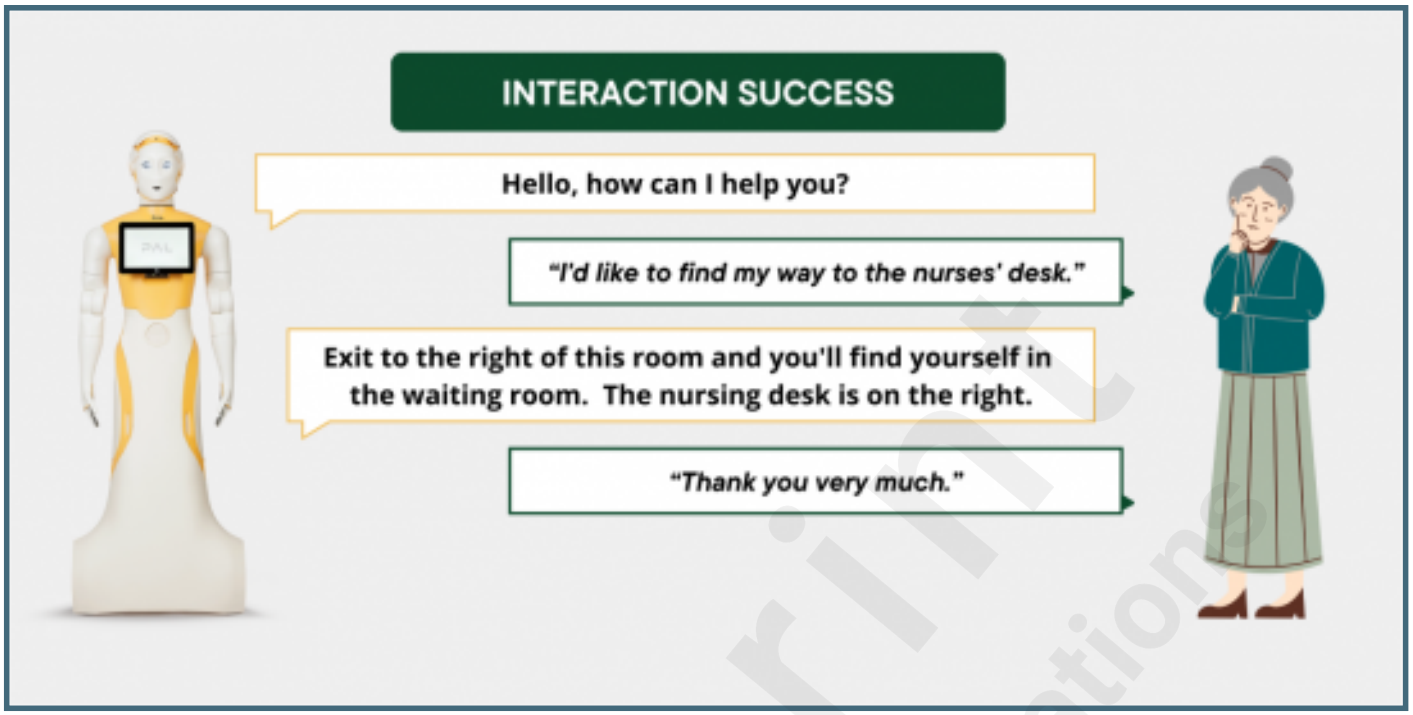
69. Home - SPRING: Socially Pertinent Robots in Gerontological Healthcare. Available from: https://spring-h2020.eu/ [accessed Aug 13, 2025]

Supplementary Files

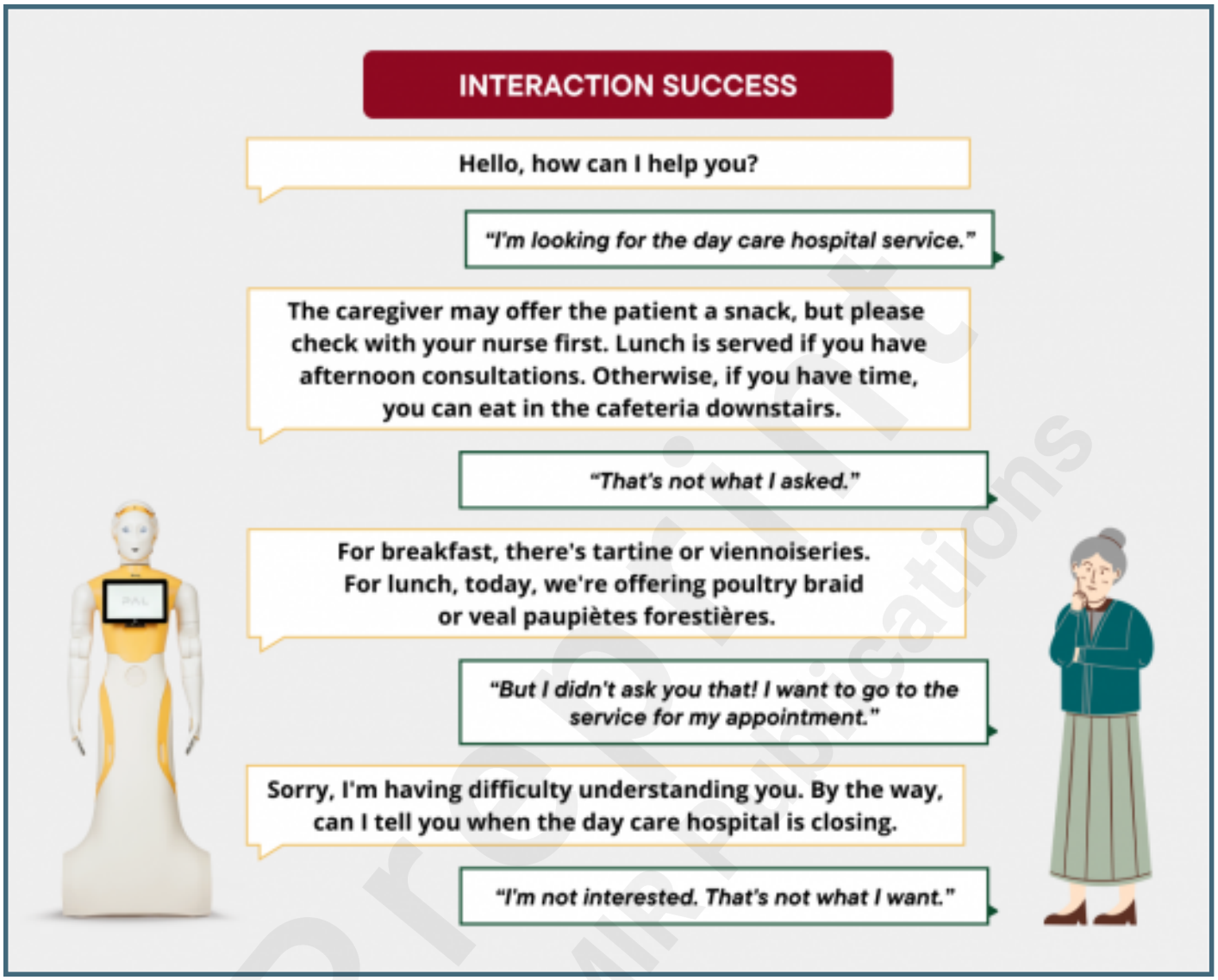
Figures

Front and side views of the ARI robot (Photo credit: PAL Robotics.

Example of a successful interaction between the ARI robot and a participant.

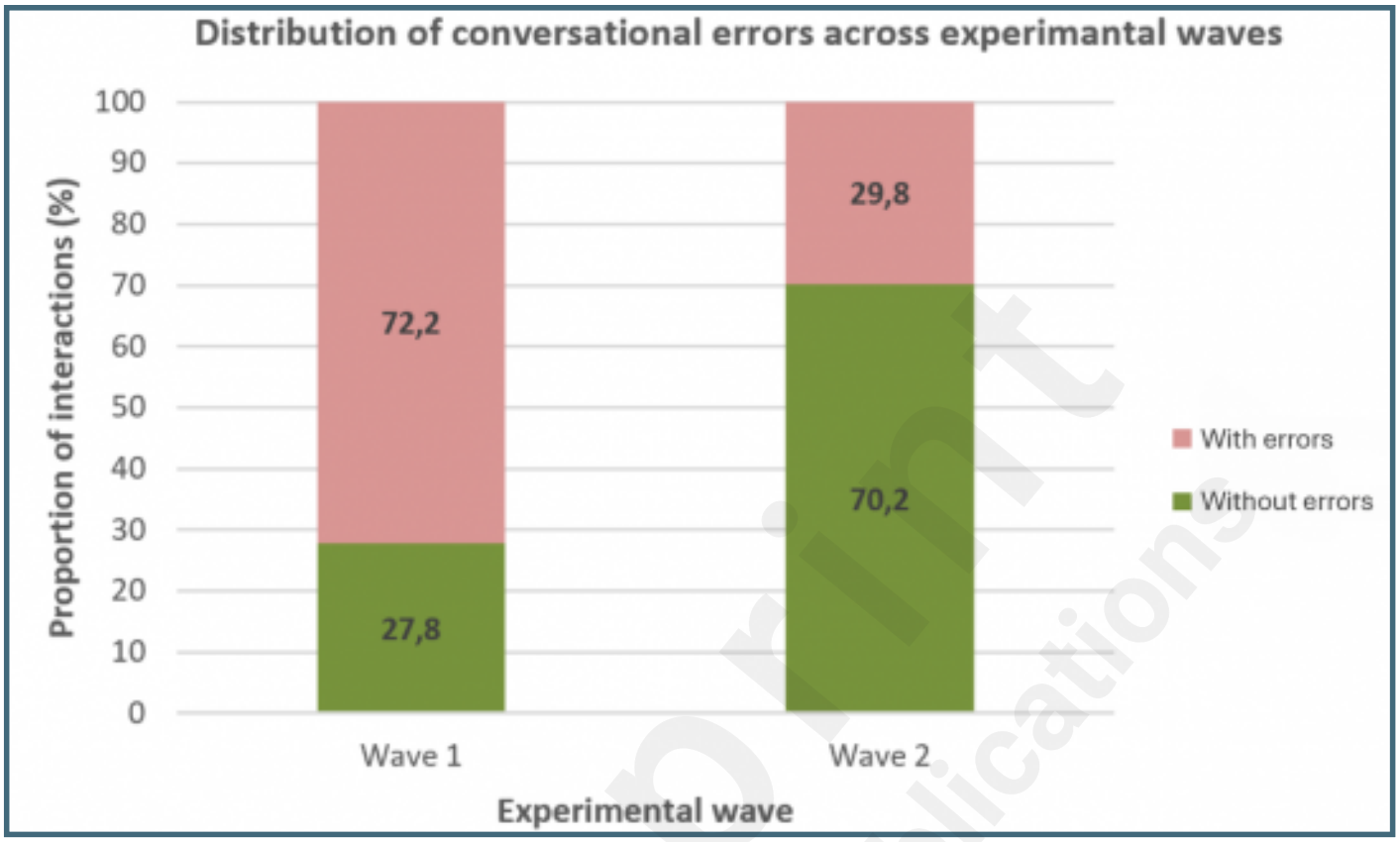


Example of a failed interaction between the ARI robot and a participant.

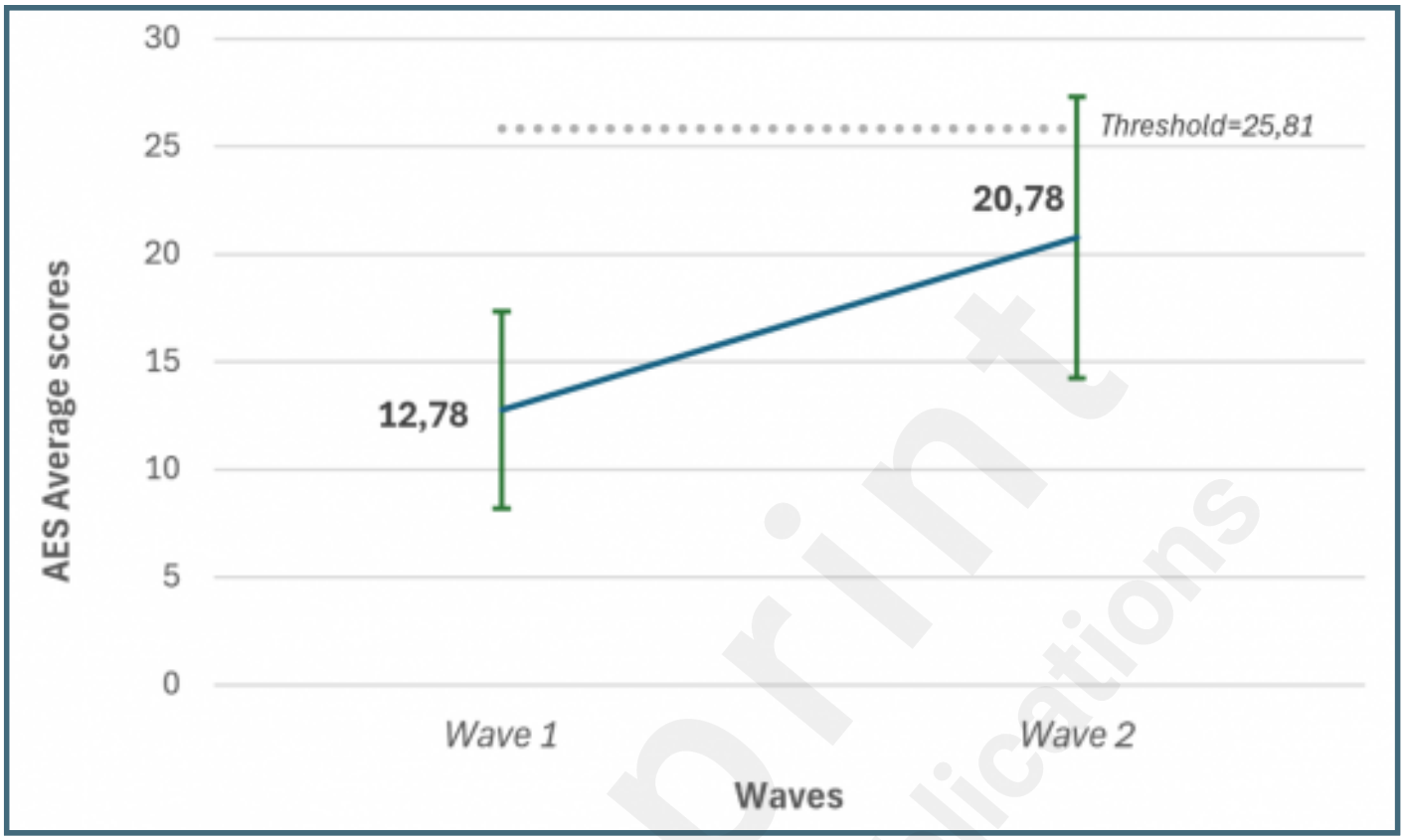


Experimental Context: Participant interacting with the SAR.

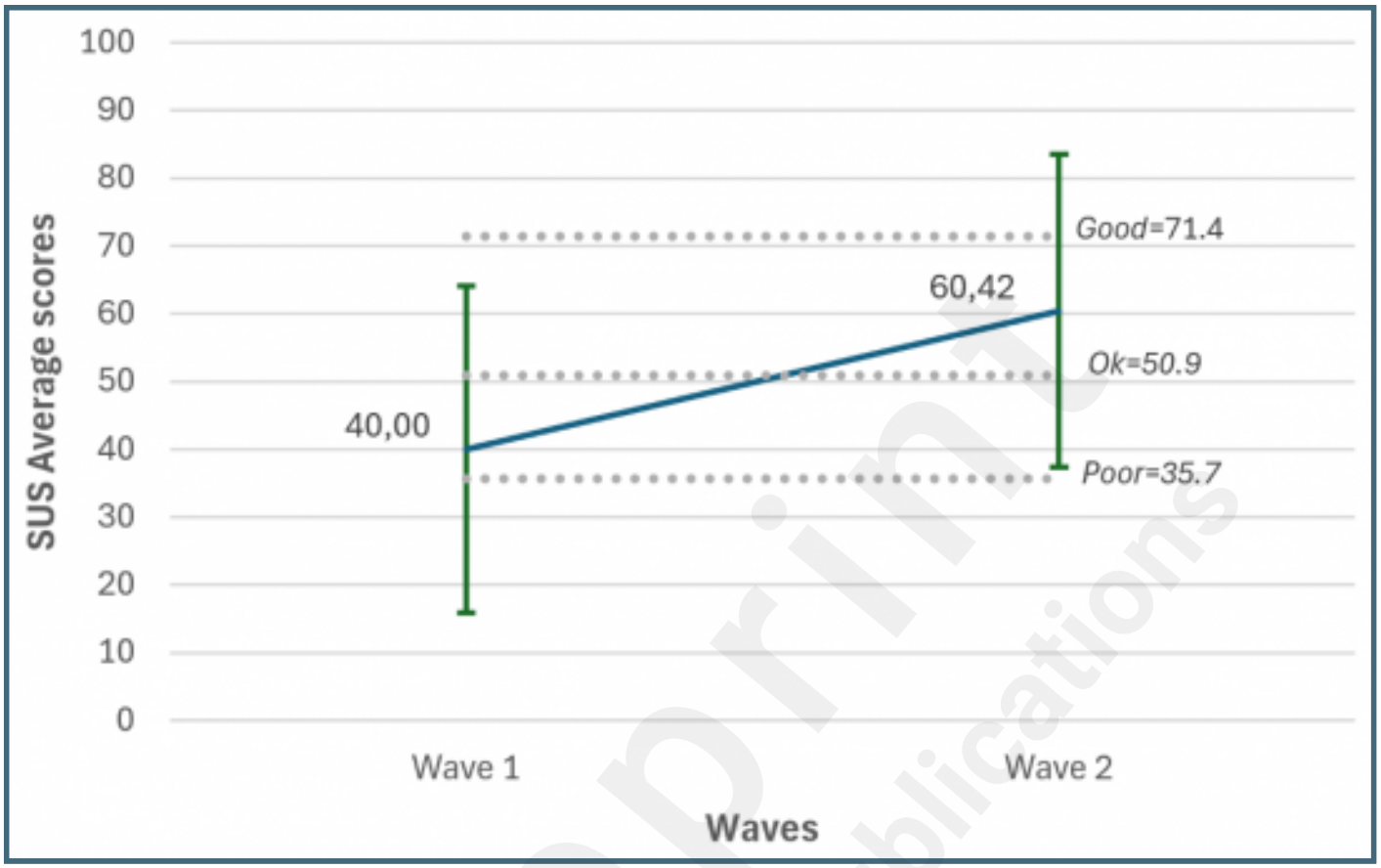
Proportion of error-free and problematic interactions by experimental wave.



AES scores across the two experimental waves .



SUS scores across the two experimental waves.



Multimedia Appendixes

Evaluation scale “AES” & “SUS” (Translated from French).
URL: http://asset.jmir.pub/assets/cf41eb9b7d3ea4e6c31fdf363d1dd562.docx

Correlations among behavioral, successful and subjective measures.
URL: http://asset.jmir.pub/assets/03fd2848a3ba08596103d612cae36837.docx

TOC/Feature image for homepages

Interaction Human Robot.