

Human-robot interaction enhanced by a Large Language Model (LLM): analysis of the acceptability and usability of a social robot in a geriatric care institution.

Lauriane Broca Living Lab, Sébastien Dacunha, Xavier Alameda-Pineda, Daniel Hernández García, Sharon Gannot, Florian Gras, Nancie Gunson, Séverin Lemaignan, Michal Polic, Pinchas Tandaitnik, Francesco Tonini, Anne-Sophie Rigaud, Maribel Pino

Submitted to: JMIR Human Factors
on: April 29, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 27

 Figures 28

 Figure 1..... 29

 Figure 2..... 30

 Figure 3..... 31

 Figure 4..... 32

Human-robot interaction enhanced by a Large Language Model (LLM): analysis of the acceptability and usability of a social robot in a geriatric care institution.

Lauriane Broca Living Lab^{1, 2, 3}; Sébastien Dacunha^{1, 2, 3}; Xavier Alameda-Pineda⁴; Daniel Hernández García⁵; Sharon Gannot⁶; Florian Gras⁷; Nancie Gunson⁵; Séverin Lemaignan⁸; Michal Polic⁹; Pinchas Tandaitnik⁶; Francesco Tonini¹⁰; Anne-Sophie Rigaud^{1, 2, 3}; Maribel Pino^{1, 2, 3}

¹ INSERM Optimisation thérapeutique en pharmacologie OTEN U1144 Université Paris Cité Paris FR

² AP-HP, Hôpital Broca Centre Mémoire de Ressources et Recherches Île-de-France-Broca Service gériatrie 1&2 75013 Paris FR

³ Broca Living Lab, Hôpital Broca Paris FR

⁴ Centre Inria de l'Université Grenoble Alpes Grenoble FR

⁵ Interaction Lab, Mathematical and Computer Sciences Heriot-Watt University Edinburgh GB

⁶ Faculty of Engineering Bar-Ilan University Bar Ilan IL

⁷ ERM Automatismes Carpentras FR

⁸ PAL Robotics Barcelona ES

⁹ CIIRC - Czech Institute of Informatics, Robotics and Cybernetics Czech Technical University in Prague Prague CZ

¹⁰ Department of Information Engineering and Computer Science University of Trento Trento IT

Corresponding Author:

Lauriane Broca Living Lab

INSERM Optimisation thérapeutique en pharmacologie OTEN U1144

Université Paris Cité

Faculté de Pharmacie de Paris

Paris

FR

Abstract

Background: Socially assistive robots offer promising opportunities to support older adults in healthcare settings by enhancing communication, reducing loneliness, and promoting emotional well-being. Beyond direct patient interaction, socially assistive robots may also assist healthcare professionals by facilitating information delivery, relieving workload, and improving the organization of care activities. However, ensuring that these systems are both acceptable and usable by older adults and healthcare staff remains a critical challenge, particularly in dynamic hospital environments.

Objective: This study aims to evaluate the acceptability and usability of a social robot in a geriatric day care hospital among patients and their informal caregivers and to identify facilitators and barriers to the adoption of social robots in healthcare settings.

Methods: Tests were conducted between May 2023 and July 2024, involving 97 participants, including patients (n=65) and informal caregivers (n=32), across three experiments, called waves. The ARI robot, developed by PAL Robotics, was used in this study. Participants were invited to interact spontaneously with the robot as long as they wished, after which they completed usability and acceptability standardized questionnaires. They were administered orally and complemented by open-comments.

Results: Quantitative analysis showed a significant increase in usability and acceptability scores across the three experimental waves. and qualitative analysis revealed similar improvements and participants reported greater comfort and satisfaction, particularly after the integration of a large language model.

Conclusions: Successive upgrades of the social assistive robot, driven by large-language-model integration, led to significant increases in acceptability and usability scores across the three experimental waves, both in patients and caregivers. Introducing the LLM enabled the robot to engage in dialogues that were more natural, coherent, and context-sensitive. These findings show the value of an iterative, user-centred refinement process when integrating social robots in geriatric care. 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” (IRB: 2021/20) and was compliant with the General Data Protection Regulation (GDPR) (DPO: 20210114153645, AP-HP register).

(JMIR Preprints 29/04/2025:76496)

DOI: <https://doi.org/10.2196/preprints.76496>

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 visible.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/76496>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/76496>

Original Manuscript

Original Paper

Human-robot interaction enhanced by a Large Language Model (LLM): analysis of the acceptability and usability of a social robot in a geriatric care institution.

Lauriane Blavette^{1,2,3,*}, Sébastien Dacunha^{1,2,3}, Xavier Alameda-Pineda⁴, Daniel Hernández García⁵, Sharon Gannot⁶, Florian Gras⁷, Nancie Gunson⁵, Séverin Lemaignan⁸, Michal Polic⁹, Pinchas Tandeitnik⁶, Francesco Tonini¹⁰, Anne-Sophie Rigaud^{1,2,3}, Maribel Pino^{1,2,3}

¹ Université Paris-Cité, INSERM Optimisation thérapeutique en pharmacologie OTEN U1144, 75006 Paris, France

² Service gériatrie 1&2, Centre Mémoire de Ressources et Recherches Île-de-France-Broca, AP-HP, Hôpital Broca, F-75013 Paris, France.

³ Broca Living Lab, Hôpital Broca, Paris, France

⁴ Inria Grenoble & Univ. Grenoble Alpes, France

⁵ Interaction Lab, Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, United Kingdom.

⁶ Faculty of Engineering, Bar Ilan University, Israel

⁷ ERM Automatismes, Carpentras, France

⁸ PAL Robotics, Barcelona, Spain

⁹ CIIRC - Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Czech Republic

¹⁰ Department of Information Engineering and Computer Science, University of Trento, Italy

*Corresponding author: lauriane.blavette@aphp.fr

Abstract

Background: Socially assistive robots offer promising opportunities to support older adults in healthcare settings by enhancing communication, reducing loneliness, and promoting emotional well-being. Beyond direct patient interaction, socially assistive robots may also assist healthcare professionals by facilitating information delivery, relieving workload, and improving the organization of care activities. However, ensuring that these systems are both acceptable and usable by older adults and healthcare staff remains a critical challenge, particularly in dynamic hospital environments.

Objective: This study aims to evaluate the acceptability and usability of a social robot in a geriatric day care hospital among patients and their informal caregivers and to identify facilitators and barriers to the adoption of social robots in healthcare settings.

Methods: Tests were conducted between May 2023 and July 2024, involving 97 participants, including patients (n=65) and informal caregivers (n=32), across three experiments, called waves. The ARI robot, developed by PAL Robotics, was used in this study. Participants were invited to interact spontaneously with the robot as long as they wished, after which they completed usability and acceptability standardized questionnaires. They were administered orally and complemented by open-comments.

Results: Quantitative analysis showed a significant increase in usability and acceptability scores across the three experimental waves. and qualitative analysis revealed similar improvements and

participants reported greater comfort and satisfaction, particularly after the integration of a large language model.

Conclusions: Successive upgrades of the social assistive robot, driven by large-language-model integration, led to significant increases in acceptability and usability scores across the three experimental waves, both in patients and caregivers. Introducing the LLM enabled the robot to engage in dialogues that were more natural, coherent, and context-sensitive. These findings show the value of an iterative, user-centred refinement process when integrating social robots in geriatric care.

Keywords: Socially assistive robots; Hospital environment; Gerontology; Older adults; Informal caregivers; Acceptability; Usability; Large Language Model; Human-robot interaction.

Introduction

The rapid aging of the global population is one of the major demographic challenges of the 21st century. According to United Nations projections, the number of people aged 65 and over worldwide is expected to increase from 727 million in 2020 to over 1.5 billion by 2050 [1]. This demographic evolution is accompanied by a growing need for long-term care and increased use of specialized services, particularly in geriatric institutions [2]. Consequently, geriatric institutions will face significant challenges in managing a rising number of residents, many of whom present multiple chronic conditions that require comprehensive and personalized care [3]. In response to this societal challenge, the adoption of innovative technologies and, in particular, socially assistive robots (SARs), is increasingly considered a complementary solution to improve the well-being and autonomy of older adults (OAs) [4,5]. SARs are robotic entities designed to interact with humans in socially and emotionally engaging ways. According to Dautenhahn (2007) [6], a SAR is “a robot that can interact with humans in a social context, while possessing communication and learning capabilities that mimic, to some extent, human behaviors”. These robots integrate artificial intelligence (AI) and natural language processing to a certain extent, enabling them to interact more naturally with users, leading to enhanced engagement and adaptation to different social contexts [7]. Within geriatric institutions, SARs have the potential to alleviate caregiver workloads, provide cognitive and social stimulation for patients, and assist with daily tasks [8]. However, for these innovations to be successfully implemented and fully beneficial, two critical ergonomic dimensions must be addressed: their acceptability and their usability in human-robot interaction (HRI).

The effectiveness of SARs in geriatric institutions largely depends on the quality of HRI [9,5]. Acceptability refers to how willing users - including patients, their informal caregivers, and healthcare professionals - are to adopt these new tools. This willingness is shaped by their beliefs, needs, and trust in the system [10,11]. Usability, on the other hand, refers to the ergonomic properties of the technology that facilitate a smooth, safe and efficient interaction - a key factor for promoting long-term engagement. Technologies that are easy to use, intuitive and aligned with users' abilities are more likely to foster sustained adoption over time [12,13].

Despite their potential, SARs still face significant limitations regarding the quality of human-robot interaction and user engagement with the technology. Studies highlight that SARs often struggle to convey emotions effectively, understand context, and predict users' behaviors accurately [14-16]. These shortcomings can lead to rigid and unnatural communication, hindering the robot's ability to interpret human intentions and emotions [17]. Additionally, SARs often lack adaptability to individual preferences, which can result in a gradual decline in user's interest and engagement [18]. These challenges highlight the importance of designing HRIs that are intuitive, customizable and aligned with the specific expectations of OAs to enhance both acceptability and successful integration into geriatric institutions.

To overcome these limitations, recent research explored the integration of large language models (LLMs) into SARs to enhance their communication and interaction capabilities [19-21]. LLMs allow for greater conversational flexibility, improved comprehension of complex requests, and enhanced contextual coherence in responses [22,23]. However, these advancements also introduce new challenges, including biases in generated responses, lack of transparency in robot decision-making, and difficulties in real-time interaction management [24,25].

Given the cognitive and physical limitations often present in the OAs, SARs must feature intuitive and user-centered HRIs to maximize their effectiveness [26]. In the context of SARs, acceptability does not only encompass perceived ease of use but also alignment with user needs, values, and preferences [10,27]. Among OAs, factors such as low digital literacy, distrust of technology, and sensory or cognitive impairments can influence acceptance [28,29]. Barriers to adoption may include reluctance regarding the robot's appearance, communication style, or ethical concerns about reduced human interaction [11].

Simultaneously, usability focuses on how efficiently, effectively, and satisfactorily a system can be used to perform expected tasks [13,30]. In geriatric institutions, usability considerations extend to physical ergonomics (e.g., size, weight, tactile interfaces, voice interaction), user safety, and emotional comfort [31]. Evaluating SAR usability through standardized methods, such as the System Usability Scale (SUS), helps identify provide insights about adoption barriers, assess ease of operation, and guide iterative technological improvements [32, 12]. These aspects emphasize the necessity of a user-centered approach tailored to the specific needs of elderly users [26].

To tackle these challenges, empirical research is essential to assess the acceptability and usability of SARs in real-world conditions. This study aims to evaluate the performance, acceptability and usability of a SAR among patients and their informal caregivers in a geriatric day care hospital (DCH) in Paris (France). It also aims to identify the factors that facilitate or hinder SAR adoption in geriatric institutions. The findings will provide deeper insights into the expectations and concerns of OAs and their informal caregivers regarding the use of SAR in healthcare while offering a set of recommendations to guide future technological and ergonomic advancements in this field.

Methods

Participants

Two populations were included in this study: day care hospital patients (OAs) and their informal caregivers.

The *inclusion criteria* for participation in this study were as follows. For OAs: (1) attending one of the consultations at the DCH of Broca, (2) being ≥ 60 years older, (3) having a Mini-Mental State Examination¹ (MMSE) score of over 10, indicating the absence of severe cognitive impairment, (4) not exhibiting symptoms of altered reality, such as delusions or hallucinations and (5) understanding and speaking french fluently. For the informal caregivers: (1) being family members or friends of the patient (primarily spouses or children) and accompanying the patient on the day of the consultation, (2) being ≥ 18 years of age and (3) understanding and speaking french fluently.

Both OAs and informal caregivers had to have expressed consent to their participation in the study. There were no *exclusion criteria* limiting participation on the basis of gender, socio-professional category or ethnic origin.

¹The Mental-State Examination (MMSE) is a commonly used set of questions for screening cognitive function that can be used to indicate the presence of cognitive impairment. Scores of 25-30 out of 30 are considered normal, 21-24 as mild, 10-20 as moderate and below 10 as severe impairment.

Recruitment of participants

Participants recruitment was carried out using the DCH database to assess the inclusion criteria two weeks prior to the study. Then, the researcher contacted the potential participants through a phone call to introduce the study and issued an invitation to participate in the study the day of their next consultation at the DCH. Upon their agreement, an information letter was sent to the participants address and they were asked to sign an informed consent the day of the experiment.

The study was conducted between May 2023 and July 2024. Three waves of data collection were conducted during this period, with a different sample of participants included in each iteration. Participants were recruited from the DCH of the French Memory Clinic at Broca Hospital (Assistance Publique-Hôpitaux de Paris - AP-HP, France).

Study design

The study adopts a mixed research approach, combining qualitative and quantitative methods to provide an in-depth understanding of the acceptability and usability of a SAR in a geriatric institution. Mixed methods are particularly suited to exploring complex phenomena in the health and social sciences, as they enable the integration of complementary perspectives and strengthen the validity of results [33,34]. By combining quantitative rating scales with semi-structured interviews, this approach identifies both general trends and nuances specific to users' perceptions [35,36]. Mixed methods enable data triangulation, essential for understanding factors that influence SAR adoption in real-world contexts [37]. This pragmatic approach, widely recognized in health services evaluation, guarantees a richer, contextualized analysis of the dynamics at play [38,39].

Material

ARI robot

The robot used for the experiments was ARI, developed by PAL Robotics (Spain). The robot is 1.65 m (5 ft 5 in) tall and weighs 50 kg. The robot moves by rolling and is equipped with articulated arms that are not designed for load-bearing. Its interface includes a touch screen for dialogue transcription located on the torso, animated eyes with a gaze-tracking module, luminous ears, and an emergency stop button on the back of the robot. ARI supports autonomous operation for 8 to 12 hours prior to recharging. It is equipped with wired and wireless connectivity capabilities, for flexible integration into a variety of environments. For visualization, ARI is equipped with 3 wide-angle cameras positioned on the head, chest and back, giving it a wide and versatile view of its surroundings. In terms of audio, the robot features four microphones, facilitating voice recognition and the efficient capture of ambient sounds.

For this study, the ARI robot was programmed with a set of modules explicitly developed as part of the European H2020 SPRING project described in Alameda-Pineda et al. (2024) [40].



Figure 1. Photos of ARI robot, front and side (Photo credit: Pal robotics)

Settings

The study was conducted at the geriatric DCH (Broca hospital, AP-HP) located in Paris (France) aimed at providing specialized care and support for OAs, who may have physical or cognitive impairments. The Broca DCH provides a range of geriatric consultations (neurology, oncology, cardiology, behavioral disorders/psychiatry, memory consultations, etc.). DCHs are outpatient facilities that provide healthcare services to individuals who require care or clinical consultations but do not need overnight hospitalization.

Introduction of ARI's capabilities

The ARI robot was deployed as a SAR in the waiting area of a geriatric DCH, with the objective of welcoming and assisting patients and their informal caregivers. Although participants were free to interact spontaneously with the robot, the researcher introduced five illustrative use cases at the beginning of the session to showcase the robot's potential functionalities: (1) greeting and welcoming the user; (2) recalling hygiene and prevention procedures; (3) providing personalized information on the consultation procedure; (4) offering orientation and guidance about hospital services and (5) providing entertainment activities. These examples were intended to guide participants' understanding of the robot's capabilities, without restricting the content of their interactions.

Ethical approval and compliance with data protection regulations

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" (IRB: 2021/20) and was compliant with the General Data Protection Regulation (GDPR) (DPO: 20210114153645, AP-HP register).

Assessment scales

To assess the acceptability and usability of the ARI robot in a DCH, we used two standardized scales.

- (a) *Acceptability*: Assessed using the Acceptability E-scale (AES), french version [41]. Acceptability is defined as the psychological antecedents influencing the behavioral intention to use technology without prior system experience. The AES scale comprises six items rated on a 5-point Likert scale, yielding a total score ranging from 6 to 30. For marketable

products, the scale's acceptability threshold is set at 25.81/30.

- (b) *Usability*: Assessed using the System Usability Scale (SUS) [42,43]. Usability refers to the ease of use and acceptability of a system or product for a specific user group performing defined tasks in a given context. Ease of use influences user performance and satisfaction, while acceptability determines actual usage [44]. The SUS consists of a 10-item scale designed to assess the overall usability of a system, generating an overall score out of 100, where a higher score reflects better usability. For this scale, experimenters are asked to respond to statements on a 5-point likert scale ranging from “strongly disagree” to “strongly agree”. The usability threshold for marketable products is 72/100 for this scale.

For each of the scales, participants were invited to comment on their choices in a discussion with the researcher.

Assessment Procedure

Participants were first received in the DCH waiting area and then guided by the experimenter to a designated room, where the experimental session (~45 minutes) was conducted.

Participants were reminded of the study objectives before signing the informed consent and image release forms. They were then positioned—either standing or seated—in front of the ARI robot and invited to interact with it freely for as long as they wished.

Following the interaction, the experimenter administered the AES and SUS scales. Each item was read aloud, and participants were encouraged to elaborate on their responses, supplementing the quantitative data with qualitative insights. Finally, participants were accompanied back to the waiting area

Data analysis

To ensure a comprehensive understanding of the research topic, both qualitative and quantitative data were collected. Each session was audio-recorded and subsequently transcribed for analysis.

Descriptive statistics (means, standard deviations and percentages) were used to describe the sample characteristics and the scores obtained on the AES and SUS scales.

To assess statistical differences between waves and between groups, several statistical tests were performed. First, Kruskal-Wallis tests were used to compare multiple groups and Mann-Whitney tests were used to compare two scores or two groups. Shapiro-Wilk tests were used for normality assumption.

When significant differences were found, Tukey's post-hoc tests were conducted to further explore the data. A p-value less than 0.05 was considered statistically significant in all analyses.

Qualitative data were analyzed using inductive thematic analysis [45], allowing themes to emerge from the transcripts of the interviews with the researcher.

System evolution for each experimental wave

In order to evaluate the evolution of the ARI robotic system, with the help of feedback from participants, we carried out system updates. Over the evaluation period, the ARI robot received two updates, resulting in 3 waves of experimentation. Table 1 shows the overall differences in system performance over the three test waves and the full list is described in SPRING Deliverable D1.6. [46]

Table 1. System evolution based on participants' feedback

Aspect	Wave 1	Wave 2	Wave 3
--------	--------	--------	--------

	<i>(May - July 2023)</i>	<i>(Sept. - Dec. 2023)</i>	<i>(March - May 2024)</i>
Robustness of robot's responses	Basic diagnostic tools to prevent operational failures; first version of a modular dialogue system.	Improved diagnostic tools accuracy; reduced memory/computational load for vision modules; integration of an extended Large Language Model (LLM) (<i>Alana V2</i>).	Enhanced general system stability; improved speech recognition robustness.
Features display and interactive functionalities	Display of dialogues transcription on the robot screen.	Optimized display output; improved display model for transcribed speech	Conversation start/end display; operator speech control; refined LLM prompt tailored to hospital context.
Perception and tracking	Depth estimation of the environment, person tracking, and facial identification modules.	Integration of a module for associating vocal input to individual users (<i>Ecapa model</i>); optimization of facial recognition processing time.	Addition of head movements enabling the robot to follow participants' gaze; optimization of gaze estimation; improved audio-visual tracking; enhanced speaker identification (<i>Ecapa2 model</i>); improved Voice Activity Detector.

Results

Participants - Socio-demographic data

Descriptive statistics

Over the three experimental waves, one hundred and ten participants (n=110) agreed to take part in the study. Of these, only ninety-seven (n=97) completed the experiment. Incomplete assessments were due to participants being called to their scheduled medical appointments. The results presented below reflect the feedback from these ninety-seven participants (n=97) across the 3 experimental waves (Wave 1=14 participants; Wave 2=43 participants; Wave 3=40 participants). Table 2 shows the socio-demographic data for these participants in each wave.

Table 2. Socio-demographic data for each experimental wave.

Wave n	Profile n (%)	Gender n (%)	Age years ($\pm$ SD)	Education years ($\pm$ SD)	MMSE score ($\pm$ SD)
Wave 1 14	Patients: 11 (79%)	M: 2 (18%); F: 9 (82%)	78.4 $\pm$ 7.1	13.6 $\pm$ 2.1	27.8 $\pm$ 1.9
	IC ² : 3 (21%)	M: 1 (33%); F: 2 (67%)	75.3 $\pm$ 11.6	12 $\pm$ 3	N/A
Wave 2 43	Patients: 30 (70%)	M: 9 (30%); F: 21 (70%)	78.3 $\pm$ 6.6	12.9 $\pm$ 3.5	25.2 $\pm$ 4.5
	IC: 13 (30%)	M: 5 (39%); F: 8 (61%)	63.9 $\pm$ 19.4	15.0 $\pm$ 00	N/A
Wave 3 40	Patients: 24 (60%)	M: 6 (25%); F: 18 (75%)	81.2 $\pm$ 6.4	11.4 $\pm$ 3.6	25.5 $\pm$ 3.9
	IC: 16 (40%)	M: 9 (56%); F: 7 (44%)	67.9 $\pm$ 12.9	13.2 $\pm$ 4.1	N/A
Total 97	Patients: 65 (67%)	M: 17 (26%); F: 48 (74%)	79.4 $\pm$ 6.7	12.5 $\pm$ 3.4	25.7 $\pm$ 4.1
	IC: 32 (33%)	M: 15 (47%); F: 17 (53%)	67 $\pm$ 15.7	13.8 $\pm$ 3.2	N/A

Socio-demographic data analysis using non-parametric tests (justified by significant Shapiro-Wilk tests indicating non-normal distributions: $P < .001$ for socio-educational level, $P = .008$ for MMSE scores, and $p \leq 0.003$ for age across waves 2 and 3) revealed no significant difference in terms of age across the three waves ($\chi^2(2) = .388$, $P = .824$), no significant difference in socio-educational level ($\chi^2(2) = 3.330$, $P = .189$), suggesting a relatively homogeneous distribution of these characteristics among the groups. Furthermore, no significant difference was observed regarding the MMSE scores of patients across waves ($\chi^2(2) = 4.064$, $P = .131$).

A correlation analysis was carried out between the AES and the SUS scores in order to explore the consistency of participant's responses across the two scales. The results reveal a strong positive correlation between the two variables ($r = 0.664$, $P < .001$), indicating that higher acceptability scores were associated with higher usability scores.

² IC = informal caregivers



Figure 2. Illustration of experiments

Acceptability E-Scale (AES)

AES results over the three waves of observation showed a positive progression between waves: 15.4/30 (SD=5.81) for Wave 1, 20.9/30 (SD=5.25) for Wave 2, and 22.5/30 (SD=4.23) for wave 3, indicating a steady increase over time. Using the Kruskal-Wallis test, the mean scores for AES show a significant increase in scores between waves ($\chi^2(2)=13.4$; $P<.001$). The Mann-Whitney test revealed no significant differences in AES scores between male and female participants ($U = 960$, $p = .705$), nor between patients and informal caregivers ($U = 905$, $p = .567$).

Analysis of the AES scores across the three experimental waves showed a significant difference between Wave 1 and Wave 2 ($P<.01$) as well as between Wave 1 and Wave 3 ($P<.001$). Additionally, no significant difference was observed between Wave 2 and Wave 3.

Based on these findings, an item-by-item analysis was conducted to identify specific differences across the themes related to system acceptability. Tukey post hoc tests revealed significant differences in four of the six items: robot usability, robot's perceived usefulness in the hospital setting, robot's response time, and overall satisfaction with the robot.

Regarding satisfaction with the robot, a significant mean difference of -0.956 was observed between Waves 1 and 2 ($p = .018$), and a difference of -1.336 between Waves 1 and 3 ($p < .001$). With respect to the robot's perceived usefulness in the hospital, a significant mean difference of -1.289 was found between Waves 1 and 3 ($p = .004$). Regarding the robot's response time showed a significant difference of -1.86 between Waves 1 and 2, and a significant difference of -2.079 between Waves 1 and 3 ($P < .001$). Finally, on overall robot satisfaction, a mean difference of -0.815 was observed between Waves 1 and 2 ($P = .042$), as well as a mean difference of -1.246 between Waves 1 and 3 ($P < .001$).

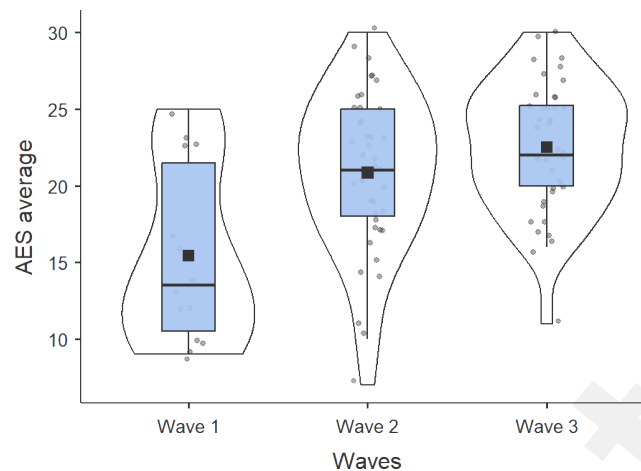


Figure 3. AES scores across the three experimental waves

System Usability Scale (SUS)

SUS results over the three waves of observation showed a positive progression between waves: 47.9/100 (SD=24.18) for Wave 1, 57.4/100 (SD=22.46) for Wave 2, and 69.3/100 (SD=16.03) for Wave 3, indicating a steady increase over time. Using the Kruskal-Wallis test, the mean scores for SUS show a significant increase in scores between waves ($\chi^2(2)=11.4$; $P<.001$). Mann-Whitney test revealed that no significant differences were observed in SUS between the results for men and women ($U=848$, $P=.141$) and for patients and those of informal caregivers ($U=855$, $P=.189$).

Analysis of SUS scores across the three experimental waves revealed significant differences between Wave 1 and Wave 3 ($p < .01$), and between Wave 2 and Wave 3 ($p < .05$). No significant difference was found between Wave 1 and Wave 2 ($p > .05$).

Based on these findings, an item-by-item analysis was conducted to examine specific differences in responses related to system usability. Tukey post hoc tests revealed significant differences in five of the ten items: conversation complexity, assistance, inconsistency, ease of use, and user confidence. On the complexity of conversations, a significant difference of -1.089 was observed between Waves 1 and Wave 3 ($P=.014$). Regarding the assistance required to interact with the robot, a significant difference of -1.182 was observed between Waves 1 and 3 ($P=.023$). Regarding the design of the robot's functionality, a significant difference of -1.039 was observed between Waves 1 and 3 ($P=.005$) as well as a mean difference of -0.569 between Waves 2 and 3 ($P=.039$). Robot inconsistencies showed a significant difference of -1.28 between Waves 1 and 2 ($P=.003$) and of -1.632 between Waves 1 and 3 ($P<.001$). Finally, for confidence using the robot, a significant difference -1.050 was observed between Waves 1 and 3 ($P=.014$).

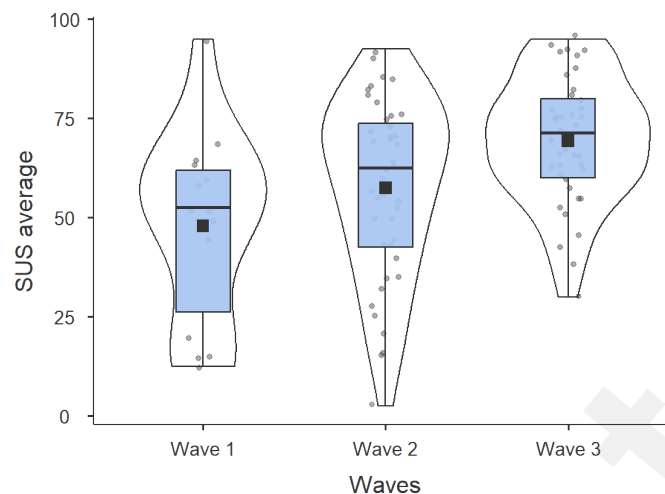


Figure 4. SUS scores across the three experimental waves

Qualitative data

Thematic qualitative analysis

A thematic qualitative analysis of participants' feedback was conducted to explore perceptions of the ARI robot's acceptability and usability. Verbatim responses from patients and informal caregivers were reviewed and categorized into five themes representing key aspects of user experience: (1) ease of use, (2) comprehension of responses, (3) perceived usefulness of the robot in a hospital setting, (4) quality of interaction, and (5) overall system satisfaction. These themes correspond to established dimensions of acceptability (e.g., perceived usefulness, satisfaction) and usability (e.g., ease of use, comprehension, efficiency). Each theme is illustrated with representative quotes translated from French to English.

Ease of use

The robot's ease of use—an essential component of acceptability—was progressively enhanced throughout the study. While its physical ergonomics remained constant, updates to the dialogue system facilitated more fluid and coherent interactions. These improvements contributed to users' perceptions of the robot as simpler and more intuitive to operate. Several participants pointed out that the novelty of the robot could, at first, generate a certain amount of reluctance, suggesting that perceived ease of use also depended on experience acquired through interaction. For example, in Wave 1, some found the voice interface practical: *"The voice is helpful because I can't see well [referring to the transcript on the robot's screen]. You just have to ask it a question."* (IC-Wave1), while others found it unintuitive: *"Not at all [easy to use]!"* (P-Wave1). However, some participants pointed out the limit of novelty when faced with a robot: *"I mean, we're not used to it. When you're used to it, it's easy, but not the first time. It's a bit confusing. You have to get used to it."* (P-Wave1). As waves progressed, more participants acknowledged the simplicity of interacting with the robot: *"It's easy, you just need to ask questions, and if the question is relevant, it answers."* (P-Wave2). However, some participants reported challenges in formulating their questions appropriately: *"I had to force myself to phrase my questions differently from how I would naturally say them."* (IC-Wave3).

Understanding the robot's response

Although voice synthesis remained consistent, updates to the dialogue system—particularly the

integration of a Large Language Model (LLM)—facilitated clearer and more coherent exchanges. While participants generally found the language comprehensible, some reported mismatches between their questions and the robot's responses, suggesting that perceived comprehension remained sensitive to the contextual specificity of the interactions. In Wave 1, participants' opinions ranged from finding the robot's speech fully understandable—*"They were completely understandable and easy."* (P-Wave1)—to reporting inconsistencies in comprehension. In Wave 2, many agreed that the robot spoke clearly: *"It was perfectly understandable."* (P-Wave2), although some responses were off-topic: *"Its explanations are clear, but some responses were off the mark."* (IC-Wave2). In Wave 3, while participants generally found the language clear: *"I found it very clear and precise."* (P-Wave3), some noted occasional mismatches between their questions and the robot's responses: *"Not entirely understandable, because it made me ask the same question multiple times."* (P-Wave 3).

Perceived usefulness of the robot in hospital settings

Perceptions of the robot's usefulness evolved over the three waves. While some participants found it useful for providing basic information, others expected a more interactive experience. Over time, the focus shifted to the accuracy of the robot's responses, with some seeing it as a "glimpse into the future", while others emphasized its current limitations. In Wave 1, some found it practical for basic information: *"Just telling me [the location needed] it's next door was enough for me."* (IC-Wave1), while others expected a more interactive experience: *"If it [the robot] had come with me [to show me the path], it would have been better."* (P-Wave1). In Wave 2, usefulness was conditional on the accuracy of responses: *"Plenty useful, but only if it understood the question and answered it correctly."* (IC-Wave2), *"It did the job, but a good signpost could do it too."* (IC-Wave2). By Wave 3, opinions remained divided, with some participants perceiving the robot as indicative of future technological developments: *"It's an opening into something futuristic. So, it's useful to me."* (IC-Wave3), while for others opinions remained mixed with respect to the robot. *"Well, it annoyed me quite a bit [...] Every time I ask it a question, it tells me to go to the reception."* (IC-Wave3). However, some participants highlighted its efficiency: *"It answered all those questions 100%."* (IC-Wave3), and its consistency compared to human interactions: *"Even if some people might find a human more pleasant, humans can get tired or irritated—whereas ARI [the robot] never does."* (IC-Wave3).

Quality of interaction

Throughout the study, updates to the robot enhanced the fluidity of interactions and the quality of user feedback—two factors critical to its acceptability. These improvements, particularly in response speed, interruption management, and response richness, were reflected in participants' feedback. During Wave 1, the robot demonstrated inconsistent response patterns, as reported by participants *"I asked a question, but it [the robot] responded with something completely unrelated—it was totally incoherent."* (P-Wave1), and some participants also perceived that the robot had a limited range of programmed responses: *"I have the impression that it is programmed only to answer certain questions."* (P-Wave1). Additionally, participants noted that the robot often repeated suggestions unrelated to their queries: *"But it does offer to go to lunch a lot."* (P-Wave1). In Wave 3, some participants noted that the robot responded promptly. *"You ask the question, and it answers almost immediately."* (IC-Wave3), whereas others reported challenges in interrupting the robot while it was delivering a response. *"You can't interrupt it, and that bothered me a bit."* (IC-Wave3). Furthermore, some participants expressed a preference for the robot's responses over those provided by human receptionists: *"I prefer your robot's answers to those of the ladies at the reception."* (P-Wave3), *"There are humans who aren't as direct and precise [...] I'm surprised."* (P-Wave3).

General satisfaction

Overall satisfaction with the experimentation of the robot reflected growing acceptance in terms of both functional effectiveness and ease of use, with clear progression observed from the first to the final wave. In Wave 1, some participants appreciated the robot's novelty: *"I think it is really nice. I think it's fun and funny."* (IC-Wave1), *"I like it—it's nice. Really nice. It's great to see a robot. But a human is still nicer."* (IC-Wave1). Conversely, some participants expressed dissatisfaction: *"I'm sorry if I sound upset, but I didn't feel any satisfaction."* (P-Wave1), *"I didn't get any satisfaction—the robot didn't help me at all. It's actually a bit scary."* (P-Wave1). In Wave 2, reactions remained mixed; some participants found the robot engaging and entertaining. *"I find it amazing, it's super."* (P-Wave2), *"It's impressive, it's the first time it's happened to me, I didn't know robots at all before today and it's impressive to see a machine that you ask questions to, that answers you."* (P-Wave2), while others highlighted its limitations: *"I'm sure the robot has potential, but at the moment, it's not performing well."* (IC-Wave2). In Wave 3, participant feedback was generally more favorable. Participants described the robot as *"very interesting, very instructive because it's very knowledgeable. It answered me perfectly well."* (P-Wave3) and declared that *"it's funny because you don't expect to find something like this, and it does have quite a lot of answers."* (IC-Wave3). Others expressed an overall appreciation: *"Oh, I like it a lot. I think it's nice, it has beautiful eyes."* (P-Wave3), *"I like this thing."* (P-Wave3).

Discussion

Synthesis of results

Quantitative data indicated a positive progression in the acceptability and usability of the ARI robot, corroborating the qualitative findings. Participants reported improved ease of use, comprehension, and perceived usefulness across experimental waves, as reflected in the significant increases in AES and SUS scores. Acceptability scores increased from 15.4/30 in Wave 1 to 22.5/30 in Wave 3 ($P < .001$), while usability scores rose from 47.9/100 to 69.3/100 ($P = .003$). Verbatim feedback reflected a growing appreciation for the robot's functionality and ease of interaction, aligning with observed improvements in satisfaction and perceived usefulness. These improvements are largely attributable to technical upgrades—most notably the integration of a large language model (LLM)—which enhanced the naturalness, contextual relevance, and consistency of interactions. Collectively, these developments contributed to a more fluid, engaging, and user-centered experience with the robot.

Although the ARI robot is commercially available, the conversational system developed in this study remains experimental. Final scores did not meet established thresholds for acceptability (AES > 25.81) and usability (SUS > 72), underscoring the need for further refinement prior to large-scale hospital implementation. Nonetheless, the upward trajectory of these scores is encouraging, suggesting strong potential for future clinical deployment as performance approaches benchmark criteria.

Limitations

This study is subject to several limitations concerning both the methodological framework and the current technical constraints of the robotic system. First, the recruitment strategy may have introduced a self-selection bias: participants were individuals who had already expressed willingness to engage in a robot-assisted intervention, potentially excluding those with skepticism or discomfort toward human-robot interaction. This limitation may have led to an overestimation of acceptability

and usability metrics, as the sample may not fully represent the broader population of older adults or healthcare professionals. Second, the study was conducted exclusively within a single site—Broca Hospital in Paris—which, while providing a well-controlled and contextually rich environment for experimentation, restricts the external validity of the findings. The specific institutional culture, infrastructure, and familiarity with innovation at this site may not be generalizable to other geriatric care settings, particularly those in regions with different sociocultural, organizational, or technological readiness profiles. Future research should address these limitations through multi-site studies and more heterogeneous participant samples to better capture variability in attitudes and interaction outcomes across diverse healthcare environments.

Another methodological limitation concerns the environmental conditions in which the experiments were conducted. Interactions took place in a controlled, relatively quiet room, which does not fully replicate the dynamic, often noisy, and unpredictable environment of a typical hospital waiting area. This semi-naturalistic setting may have influenced participant behavior and engagement, potentially leading to an overestimation of the robot's perceived usability and acceptability. The artificial calm may have reduced distractions and facilitated smoother interactions, conditions not typically encountered in real-world hospital contexts. Furthermore, participants interacted with the robot on a single occasion. As such, the reported acceptability scores primarily reflect first impressions and do not account for how perceptions might evolve over time or with repeated exposure. Longitudinal studies involving multiple interactions across varied real-world contexts are needed to assess sustained engagement, evolving user trust, and the long-term integration potential of socially assistive robots in clinical environments.

A final limitation pertains to the current technical capabilities of the robot, specifically the absence of multilingual support in its conversational system. Given the multicultural and linguistically diverse nature of many healthcare environments, this constraint may have limited the inclusivity and accessibility of the robot for non-French-speaking users. The inability to communicate in multiple languages may have excluded potential participants or diminished the quality of interaction for those with limited proficiency in the default language. Integrating robust multilingual functionalities would not only broaden the range of users but also enhance the robot's usability and relevance across diverse care settings. As healthcare institutions increasingly serve heterogeneous populations, this enhancement is essential for ensuring equitable and culturally competent deployment of socially assistive robotics [26].

Comparison with prior work

The integration of a LLM between Wave 1 and Wave 2 led to a marked improvement in the robot's conversational fluency and contextual relevance. This enhancement aligns with existing literature suggesting that AI-powered SARs can significantly increase user engagement through more natural and adaptive dialogue capabilities [19,21]. Additional technical upgrades—including the incorporation of a gaze tracking module and improvements in response stability—further contributed to a more seamless and intuitive interaction experience. These modifications reflect critical dimensions of human-robot interaction, as prior research has highlighted the role of gaze behavior and system responsiveness in fostering user engagement, building trust, and supporting sustained interaction [47]. Collectively, these enhancements underscore the importance of multimodal and context-aware communication features in the design of socially assistive technologies for healthcare settings.

Despite these advancements, challenges remained concerning the accuracy and adaptability of the robot's responses. These limitations are consistent with known constraints of LLMs, particularly in processing ambiguous or highly context-dependent inputs [48]. Such issues can hinder the coherence

and appropriateness of interactions, especially in complex or emotionally nuanced scenarios. To address these shortcomings, future iterations should incorporate multimodal communication strategies—including gesture recognition, facial expression analysis, and emotion detection—to enrich the robot's social presence and improve the overall quality of human-robot interaction [49,16]. These enhancements are critical for fostering more natural, empathetic, and contextually appropriate exchanges in healthcare environments.

A central factor influencing ease of use was the robot's voice-based interaction modality. This feature was particularly appreciated by OAs who are less familiar or comfortable with tactile interfaces [50]. Our findings reinforce those of Jakob et al. (2025) [51], as participants consistently expressed a preference for spoken communication over touchscreen-based interaction. Furthermore, the integration of on-screen dialogue transcription enhanced accessibility by allowing users to visually track the exchange and adjust their speech in cases of miscommunication. This design element aligns with the results of Chen et al. (2023) [52], who demonstrated that real-time transcript visualization supports improved engagement and facilitates error recovery among OAs. Together, these findings highlight the value of combining auditory and visual modalities to promote usability and inclusivity in socially assistive robotic systems.

User confidence in interacting with the robot appeared closely tied to the fluidity and coherence of its responses. Participants who experienced consistent and contextually appropriate interactions reported greater trust in the system, while those who encountered erratic or incoherent outputs expressed ongoing skepticism. These findings underscore the critical role of transparency and predictability in AI-driven behaviors as key determinants of user trust, reinforcing prior research in human-robot interaction [15].

Despite overall improvements, instances of inaccurate responses or misinterpretations by the system occasionally led to user frustration. Notably, participants did not consistently attribute these breakdowns to the robot's limitations; instead, many questioned the clarity of their own input. This pattern reflects a well-documented phenomenon in human-robot interaction literature—automation bias—where users tend to overestimate the capabilities of intelligent systems and assume personal responsibility for communication failures [53,54]. Such findings underscore the importance of designing transparent and error-tolerant interfaces that help users appropriately calibrate their expectations and maintain trust during interactions.

In addition to its practical utility—particularly in delivering information such as wayfinding—participants expressed concerns about the robot's potential impact on human interaction within healthcare settings. While some valued the efficiency and accessibility of robotic assistance, others feared it might diminish opportunities for meaningful human connection, particularly in contexts requiring empathy, reassurance, or emotional support. These concerns echo broader ethical debates surrounding the integration of SARs in clinical environments, emphasizing the importance of safeguarding the relational aspects of care. They highlight the ethical imperative to balance technological innovation with human-centered values, ensuring that robots are used to complement rather than substitute the compassionate presence of healthcare professionals [28,5].

Another important consideration is the level of user familiarity with robotic systems. For the majority of participants, this study marked their first direct interaction with a SAR, and many indicated the need for a period of adjustment. This finding aligns with existing research suggesting that repeated exposure to robotic technologies fosters greater comfort, trust, and acceptance (Mishra et al., 2021). These results point to the importance of integrating structured onboarding strategies—such as guided tutorials or gradual exposure—in future implementations. Such measures could

support user adaptation, reduce initial hesitation, and enhance long-term acceptability of SARs in healthcare settings.

Recommendations

Drawing on the empirical findings of this study, the following recommendations are intended to inform the design and implementation of SARs tailored to the specific interactional needs of OAs in geriatric care settings. These evidence-based insights aim to guide researchers and industry stakeholders in developing socially assistive technologies that are both user-centered and contextually appropriate within healthcare environments.

Improve natural language processing, speech recognition, and multilingual support

To enhance the quality of human-robot interaction in geriatric care, it is recommended that future socially assistive robots (SARs) integrate more advanced natural language processing capabilities to ensure greater coherence and contextual relevance in responses. Speech recognition systems should be optimized to account for prosodic variation, regional accents, and articulation differences, particularly those common among older adults. In addition, the incorporation of multilingual functionality is essential to accommodate the linguistic diversity of users in healthcare settings and to promote equitable access to robotic services.

Develop clear onboarding protocols and autonomous robot self-presentation

To facilitate user engagement and support initial interaction, it is recommended that future deployments of SARs include a structured onboarding protocol. This may involve brief demonstrations or the distribution of printed materials (e.g., leaflets with example questions) to familiarize users with the robot's capabilities. However, recognizing that real-world hospital settings may not consistently offer human assistance at the point of interaction, SARs should also be equipped with a robust self-introduction feature. This feature should autonomously guide users through the robot's purpose, functionalities, and appropriate use cases—clarifying when and how to engage, and what types of questions it can address. Such improvements are essential for promoting user autonomy, reducing uncertainty, and improving the perceived usefulness of SARs in healthcare environments.

Ensure compliance with privacy, ethical, and regulatory standards

It is essential that the development and deployment of SARs in healthcare settings align with existing ethical guidelines and data protection regulations. Particular attention must be given to the collection, processing, and storage of sensitive health-related information. Developers and implementers should establish clear protocols for obtaining informed consent, ensuring transparency regarding data usage, and enabling users to understand the scope and limitations of data handling. Additionally, SARs should be designed to support confidential interactions—especially when used with vulnerable populations such as older adults—by incorporating features that safeguard user privacy and uphold professional standards of care. Early integration of ethical and legal considerations will be critical to building trust and ensuring responsible implementation.

Facilitate operational integration through comprehensive staff training

The successful adoption of socially assistive robots (SARs) in healthcare settings relies heavily on their effective integration into existing clinical workflows and team dynamics. It is therefore recommended that structured training programs be implemented for caregivers and healthcare staff who will interact with or support the use of SARs. Training should address both technical competencies—such as routine maintenance, cleaning protocols, storage, charging procedures, and responses to malfunctions—and usability-related aspects, including how to support end-users and

integrate the robot into care routines. This dual approach, combining technical preparedness with cultural adaptation, will help ensure that staff are confident in the robot's role, thereby promoting sustained acceptability and long-term integration within hospital environments.

Conclusions

This study explored the progressive integration and evaluation of a SAR within a geriatric hospital setting, with the aim of assessing its acceptability and usability among OAs and healthcare professionals. Through a mixed-methods approach conducted over three experimental waves, the research sought to understand how iterative technical enhancements—particularly the integration of a LLM—could influence user perceptions and interaction quality.

The results indicate a clear improvement in both acceptability and usability over time, as reflected by significant increases in acceptability and usability scores. The incorporation of the LLM represented a critical milestone, significantly enhancing the robot's ability to engage in more natural, coherent, and contextually appropriate conversations. This advancement contributed to smoother user interactions, improved response accuracy, and a higher perception of the robot's usefulness—factors that were consistently echoed in the qualitative data, where participants described later-stage interactions as more intuitive, fluid, and engaging compared to earlier phases.

Looking ahead, future work should focus on further leveraging the capabilities of LLMs to enable adaptive, personalized interactions that respond to individual users' communication styles, memory, and emotional cues. Additionally, the integration of multimodal communication features—such as gaze tracking, gesture and facial expression recognition, and targeted dialogue strategies—will be essential to enhancing the robot's relevance, trustworthiness, and effectiveness in complex healthcare environments.

Acknowledgements

The authors confirm contribution to the paper as follows: study conception and design: L. Blavette, S. Dacunha, X. Alameda-Pineda, D. Hernández García, S. Gannot, F. Gras, N. Gunson, S. Lemaignan, M. Polic, P. Tandeitnik, F. Tonini, A-S. Rigaud, M. Pino; data collection: L. Blavette; analysis and interpretation of results: L. Blavette, S. Dacunha; draft manuscript preparation: L. Blavette, S. Dacunha, M. Pino, A-S. Rigaud. 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 [55].

The authors would like to sincerely thank all the participants who took part in this study, as well as all the healthcare 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

IC: informal caregiver(s)

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 anne-sophie.rigaud@aphp.fr.

References

1. United Nations Department of Economic and Social Affairs, Population Division. World population ageing 2020 highlights. New York: United Nations; 2020. Accessed April 16, 2025. <https://www.un.org/development/desa/pd/news/world-population-ageing-2020-highlights>
2. World Health Organization. Ageing and health. 2023. Accessed April 16, 2025. <https://www.who.int/news-room/fact-sheets/detail/ageing-and-health>
3. OECD. Health at a glance: Europe 2020. Paris; 2020. Accessed April 16, 2025. https://www.oecd.org/en/publications/health-at-a-glance-europe-2020_82129230-en.html
4. Chu L, et al. Identifying features that enhance older adults' acceptance of robots: a mixed methods study. *Gerontology*; 2019;65(4):441–450. doi:10.1159/000494881
5. Naneva S, et al. A systematic review of attitudes, anxiety, acceptance, and trust towards social robots. *Int J Soc Robot*; 2020;12(6):1179–1201. doi:10.1007/s12369-020-00659-4
6. Dautenhahn K. Socially intelligent robots: dimensions of human–robot interaction. *Philos Trans R Soc Lond B Biol Sci*; 2007;362(1480):679–704. doi:10.1098/rstb.2006.2004
7. Mataric M. Socially assistive robotics: human-robot interaction methods for creating robots that care. In: *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction*; 2014 Mar 3–6; Bielefeld, Germany. New York: ACM; 2014. p. 333. doi:10.1109/HRI.2014.7020000
8. Saadatzi MN, et al. Acceptability of using a robotic nursing assistant in health care environments: experimental pilot study. *J Med Internet Res*; 2020;22(11):e17509. doi:10.2196/17509
9. David D, et al. The acceptability of social robots: a scoping review of the recent literature. *Comput Human Behav*; 2022;137:107419. doi:10.1016/j.chb.2022.107419
10. de Jong C, et al. Intentional acceptance of social robots: development and validation of a self-report measure for children. *Int J Hum Comput Stud*; 2020;139:102426. doi:10.1016/j.ijhcs.2020.102426
11. Mishra N, et al. Exploring potential and acceptance of socially intelligent robot. In: Thalmann NM, Zhang JJ, Ramanathan M, Thalmann D, editors. *Intelligent scene modeling and human-computer interaction*. Cham: Springer International Publishing; 2021. p. 259–282. doi:10.1007/978-3-030-71002-6_15
12. Martín Rico F, et al. An acceptance test for assistive robots. *Sensors*; 2020;20(14):3912. doi:10.3390/s20143912
13. Holden RJ. A simplified system usability scale (SUS) for cognitively impaired and older adults. *Proc Int Symp Hum Factors Ergon Health Care*; 2020;9(1):180–182. doi:10.1177/2327857920091021
14. Bishop L, et al. Social robots: the influence of human and robot characteristics on acceptance. *Paladyn J Behav Robot*; 2019;10(1):346–358. doi:10.1515/pjbr-2019-0028
15. de Graaf MMA, Ben Allouch S. Exploring influencing variables for the acceptance of social robots. *Robot Auton Syst*; 2013;61(12):1476–1486. doi:10.1016/j.robot.2013.07.007
16. Eyssel F, et al. Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. In: *Proceedings of the 6th International Conference on Human-Robot Interaction*; 2011 Mar 6–9; Lausanne, Switzerland. New York:

- ACM; 2011. p. 61–68. doi:10.1145/1957656.1957673
17. Heerink M, et al. The influence of a robot's social abilities on acceptance by elderly users. In: Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication; 2006 Sep 6–8; Hatfield, UK. Piscataway, NJ: IEEE; 2006. p. 521–526. doi:10.1109/ROMAN.2006.314442
 18. Whelan S, et al. Factors affecting the acceptability of social robots by older adults including people with dementia or cognitive impairment: a literature review. *Int J Soc Robot*; 2018;10(5):643–668. doi:10.1007/s12369-018-0471-x
 19. Hanschmann L, et al. A LLM-based social robot for human-like sales conversations. In: Følstad A, Araujo T, Papadopoulos S, Law ELC, Luger E, Goodwin M, Hobert S, Brandtzaeg PB, editors. *Chatbot research and design*. Cham: Springer Nature Switzerland; 2024. p. 61–76. doi:10.1007/978-3-031-54975-5_4
 20. Atuhurra J. Leveraging large language models in human-robot interaction: a critical analysis of potential and pitfalls. *arXiv*; 2024. doi:10.48550/arXiv.2405.00693
 21. Kim CY, et al. Understanding large-language model (LLM)-powered human-robot interaction. In: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction; 2024 Mar 11–14; Boulder, CO. New York: ACM; 2024. p. 371–380. doi:10.1145/3610977.3634966
 22. Shibi S, Zaidi S. STEM approach to enhance robot-human interaction through AI large language models and reinforcement learning. In: Proceedings of the 2024 IEEE Integrated STEM Education Conference (ISEC); 2024 Mar 9; Princeton, NJ. Piscataway, NJ: IEEE; 2024. p. 1–2. doi:10.1109/ISEC61299.2024.10665163
 23. Sobrín-Hidalgo D, et al. Explaining autonomy: enhancing human-robot interaction through explanation generation with large language models. *arXiv*; 2024. doi:10.48550/arXiv.2402.04206
 24. Zhang B, Soh H. Large language models as zero-shot human models for human-robot interaction. In: Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2023 Oct 1–5; Detroit, MI. Piscataway, NJ: IEEE; 2023. p. 7961–7968. doi:10.1109/IROS55552.2023.10341488
 25. Zhang C, et al. Large language models for human–robot interaction: a review. *Biomim Intell Robot*; 2023;3(4):100131. doi:10.1016/j.birob.2023.100131
 26. Mahmoudi Asl A, et al. Evaluating the user experience and usability of the MINI robot for elderly adults with mild dementia and mild cognitive impairment: insights and recommendations. *Sensors*; 2024;24(22):7180. doi:10.3390/s24227180
 27. Liberman-Pincu E, Oron-Gilad T. Exploring the effect of mass customization on user acceptance of socially assistive robots (SARs). In: Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI); 2022 Mar 7–10; Sapporo, Japan. New York: ACM; 2022. p. 880–884. doi:10.1109/HRI53351.2022.9889466
 28. Beuscher LM, Fan J, et al. Socially assistive robots: measuring older adults' perceptions. *J Gerontol Nurs*; 2017;43(12):35–43. doi:10.3928/00989134-20170707-04
 29. Becchimanzi C, et al. Acceptability of assistive robotics by older adults: results from a human-centred qualitative study. In: Proceedings of the AHFE 2022 International Conference on Human Factors in Accessibility and Assistive Technology; 2022 Jul 24–28; New York, NY. Cham: Springer; 2022. p. 37. doi:10.54941/ahfe1001637
 30. Bevilacqua R, et al. Robot-Era project: preliminary results on the system usability. In: Marcus A, editor. *Design, user experience, and usability: interactive experience design*. Cham: Springer International Publishing; 2015. p. 553–561. doi:10.1007/978-3-319-20889-3_51
 31. Olde Keizer RACM, et al. Using socially assistive robots for monitoring and preventing frailty among older adults: a study on usability and user experience challenges. *Health Technol*; 2019;9(4):595–605. doi:10.1007/s12553-019-00320-9

32. Olatunji S, et al. Usability testing for the operation of a mobile robotic telepresence system by older adults. *Proc Hum Factors Ergon Soc Annu Meet*; 2020;64(1):1284–1288. doi:10.1177/1071181320641284
33. Almeida F. Strategies to perform a mixed methods study. Zenodo; 2018. doi:10.5281/zenodo.1406214
34. Pluye P, Hong QN. Combining the power of stories and the power of numbers: mixed methods research and mixed studies reviews. *Annu Rev Public Health*; 2014;35:29–45. doi:10.1146/annurev-publhealth-032013-182440
35. Curry LA, et al. Mixed methods in biomedical and health services research. *Circ Cardiovasc Qual Outcomes*; 2013;6(1):119–123. doi:10.1161/CIRCOUTCOMES.112.967885
36. Scammon DL, et al. Connecting the dots and merging meaning: using mixed methods to study primary care delivery transformation. *Health Serv Res*; 2013;48(6pt2):2181–2207. doi:10.1111/1475-6773.12114
37. Shorten A, Smith J. Mixed methods research: expanding the evidence base. *Evid Based Nurs*; 2017;20(3):74–75. doi:10.1136/eb-2017-102699
38. Terrell S. Mixed-methods research methodologies. *Qual Rep*; 2012;17:254–265. doi:10.46743/2160-3715/2012.1819
39. McKim CA. The value of mixed methods research: a mixed methods study. *J Mixed Methods Res*; 2017;11(2):202–222. doi:10.1177/1558689815607096
40. Alameda-Pineda X, et al. Socially pertinent robots in gerontological healthcare. *arXiv*; 2024. doi:10.48550/arXiv.2404.07560
41. Micoulaud-Franchi JA, et al. Validation of the French version of the Acceptability E-scale (AES) for mental e-health systems. *Psychiatry Res*; 2016;237:196–200. doi:10.1016/j.psychres.2016.01.043
42. Bangor A, et al. An empirical evaluation of the system usability scale. *Int J Hum Comput Interact*; 2008;24(6):574–594. doi:10.1080/10447310802205776
43. Brooke J. SUS: a quick and dirty usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. *Usability evaluation in industry*. London: Taylor & Francis; 1996. p. 189–194. ISBN:0748404606
44. Shackel B. What is usability? In: *Proceedings of the 4th International Conference on Human-Computer Interaction*; 1991; Stuttgart, Germany. Amsterdam: Elsevier; 1991. p. 3–15. doi:10.1016/B978-0-444-88464-4.50003-4
45. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*; 2006;3(2):77–101. doi:10.1191/1478088706qp063oa
46. SPRING Consortium. SPRING D1.6 – User Feedback from the Final Validation in Relevant Environments; 2024. 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
47. Admoni H, Scassellati B. Social eye gaze in human-robot interaction: a review. *J Hum Robot Interact*; 2017;6(1):25–63. doi:10.5898/JHRI.6.1.Admoni
48. Ramrakhya R, et al. Grounding multimodal LLMs to embodied agents that ask for help with reinforcement learning. *arXiv*; 2025. doi:10.48550/arXiv.2504.00907
49. Dautenhahn K, et al. KASPAR – a minimally expressive humanoid robot for human–robot interaction research. In: *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*; 2013 Aug 26–29; Gyeongju, Korea. Piscataway, NJ: IEEE; 2013. p. 931–936. <https://www.researchgate.net/publication/255601703>
50. Granata C et al.. Robot services for elderly with cognitive impairment: testing usability of graphical user interfaces. *Technol Health Care*; 2013;21(3):217–231. doi:10.3233/THC-130718
51. Jakob D, et al. Adapting voice assistant technology for older adults: a comprehensive study on usability, learning patterns, and acceptance. *Digit*; 2025;5(1):1.

doi:10.3390/digital5010004

52. Chen C, et al. Screen or no screen? Lessons learnt from a real-world deployment study of using voice assistants with and without touchscreen for older adults. In: Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility; 2023; New York, NY. New York: ACM; 2023. p. 1–21. doi:10.1145/3597638.3608378
53. Jones-Jang SM, Park YJ. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. J Comput Mediat Commun; 2022;28(1):zmac029. doi:10.1093/jcmc/zmac029
54. Rédaction L. La relation empathique homme/robot: normale ou pathologique? Implications philosophiques; 2021. Accessed April 16, 2025. <https://www.implications-philosophiques.org/la-relation-empathique-homme-robot-normale-ou-pathologique>
55. SPRING: Socially pertinent robots in gerontological healthcare; 2025. <https://spring-h2020.eu>

Supplementary Files

Figures

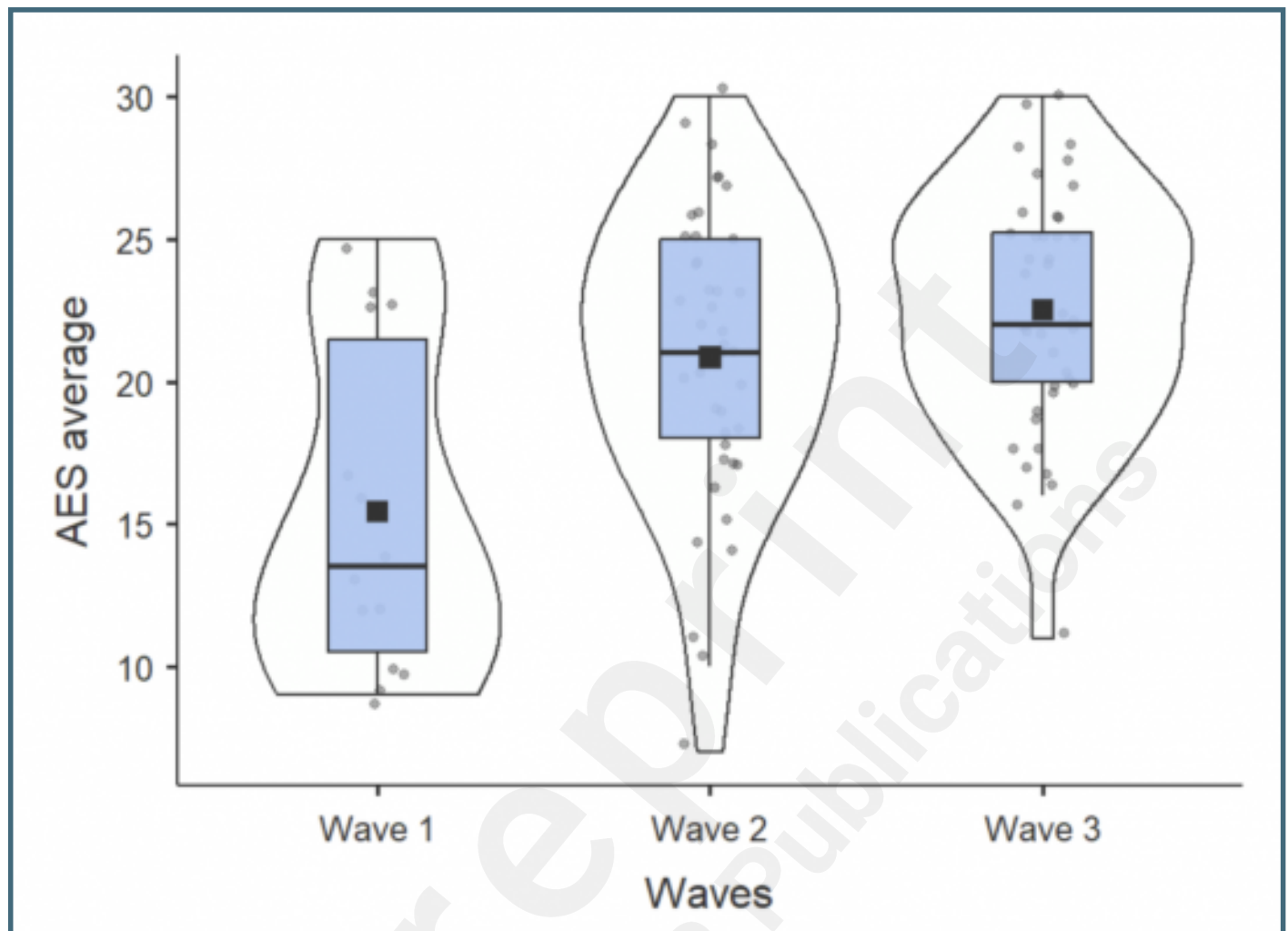
Photos of ARI robot, front and side (Photo credit: Pal robotics).



Illustration of experiments.



AES scores across the three experimental waves.



SUS scores across the three experimental waves.

