
Conversational Robots to Support Well-being and Home Safety in Dementia Care

by

Maria Raposo de Lima

A thesis submitted in fulfilment of the requirements for the degree of *Philosophiae Doctor*

Department of Mechanical Engineering
Imperial College London

October 2024

Department of Mechanical Engineering,
Imperial College London,
Exhibition Rd,
South Kensington,
London SW7 2BX

Maria R. Lima © 2024

This dissertation is dedicated to my Family, my safe harbour every day, whose unconditional love, incomparable support, and infinite encouragement have made it all worthwhile. To my Mom and Dad, my constant source of inspiration, perseverance, creativity, positive attitude, and strength. To my sisters Madalena and Marta, my two special guiding lights, for their presence, kindness, and support in all my highs and lows. To my grandparents Avó Zé, Avô Vasco, Avó Maria José and Avô Mário, who anchored me with the roots and values that shape who I am, and believed in me at every step of this journey. Muito obrigada!

Declaration

I declare that the content presented in this dissertation is original, except where specific reference is made to the work of others.

Maria Raposo de Lima

3rd October 2024

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

What a journey! I am profoundly thankful for the unwavering support I received throughout the past four PhD years that has made this dissertation possible, helped me overcome every challenge along the way, and allowed me to enjoy the ride to get where I am today.

To my advisor Ravi Vaidyanathan, for the guidance, patience and continuing support, for the invaluable and fun discussions, and for opening doors to multiple research opportunities that led to this dissertation and greatly contributed to my personal growth.

To Payam Barnaghi, for the 'this is brilliant' constant encouragement, the ideas discussed that shaped much of the work accomplished and challenged me to think more critically about the potential of this research, and for always having smart answers to my questions and PhD worries.

To Ferdinando Rodriguez y Baena, for always having his door open to share invaluable career advice, for the support in all stages of my PhD and the 'good morning super achiever' motivation.

To Maja Matarić, for the invaluable mentorship, the weekly outdoor meetings during my visit that were equally fruitful and inspiring, for endless ideas and rounds of human-robot interaction design, and the advice on pursuing purpose in the work we choose to do.

I am thankful to the Imperial Mechanical Engineering Department. Getting to work in a vibrant research community where I feel inspired, supported, and heard, really makes a difference. My thanks go to the UK DRI CR&T team for very fruitful and respectful collaborations: Sarah Daniels, Eyal Soreq, Sophie Horrocks, Matthew Harrison, and Alexander Capstick. To Sridhar Vaitheswaran and the SCARF India team, for the warm welcome and support during my research visit in Chennai. I also thank the USC Interaction Lab for the fun and collaborative culture during my research visit.

I am thankful for the financial support provided by the Imperial President's PhD Scholarship scheme, the UK Institution of Mechanical Engineers, and the Lloyd's Register Foundation through the Dr Ashraf Ben El-Shanawany Memorial Prize. Together with my advisor's support, this enabled me to pursue the summer research visit at USC in Los Angeles, CA.

To my great friends, for the laughter and encouragement that pushed me through the most challenging times during this journey.

Lastly, my heartfelt thanks to my partner in life Miguel, for the love, memorable adventures, and the happiest London life that made every moment more worthwhile.

Abstract

As the global population ages and dementia prevalence increases, there is a pressing need for accessible and scalable technologies to support well-being and safety at home. Timely detection of cognitive impairment remains a major unmet need. Conversational robots hold potential for home assistance through verbal interactions. However, real-world deployments face challenges in long-term engagement, personalised care, and clinical utility. This dissertation designed and evaluated conversational robots to support well-being, enhance home safety, and monitor cognitive health.

In a year-long deployment of conversational technology in homes of people affected by dementia (N=14), this work demonstrated the feasibility of mapping behaviour by analysing activity sequences, highlighting the potential of conversational AI to provide proactive verbal support in response to safety risks.

This research introduced a machine learning approach and evaluated spoken language as a biomarker for automated assessment of cognitive health. In screening cognitive impairment, the best-performing model, trained on 100 interpretable linguistic features extracted from a benchmark DementiaBank dataset (N=291), achieved a mean sensitivity of 69.4% and specificity of 83.3%. On pilot data collected in-residence (N=22), this model achieved a mean sensitivity of 70% and specificity of 52.5%. In predicting the severity of cognitive decline via Mini-Mental State Examination (MMSE) scores, results showed a mean absolute MMSE error of 3.7 and comparable performance of 3.3 on real-world data, demonstrating generalisability. Risk stratification and feature importance analysis enhanced the clinical utility of model predictions and enabled the identification of higher-risk individuals for timely interventions.

This dissertation involved stakeholders with lived experience of ageing and dementia (N=111) in real-world evaluation studies conducted in the UK, India, and the USA, assessing technology acceptance, user engagement, and feasibility of home use. Additionally, this work designed a large language model-powered socially assistive robot and validated a new approach to in-residence cognitive support through human-robot verbal dialogue.

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Objectives	6
1.3	Contributions	7
1.4	Publications	9
1.5	Dissertation Outline	11
2	Related Work	12
2.1	Socially Assistive Robots in Older Age and Dementia Care	13
2.2	Conversational AI for Health and Well-being Support	19
2.2.1	Communication and Adoption Challenges	19
2.2.2	Opportunities and Risks of LLMs	20
2.3	Machine Learning for Cognitive Health Assessment	22
2.4	Research Challenges	23
3	Behaviour Monitoring with Conversational Technology at Home	24
3.1	Overview	25
3.2	Background	26
3.2.1	Home Activity Monitoring	26
3.2.2	Similarity Methods	26
3.2.3	Topic Modelling	27
3.3	Methods	28
3.3.1	Research Questions	28
3.3.2	Participants and Setting	29
3.3.3	Longitudinal Deployment	29
3.3.4	Similarity of Activity Sequences	31
3.3.5	Topic Modelling of Interaction Data	34
3.4	Results	35
3.4.1	Long-Term Engagement at Home	35
3.4.2	Mapping Behaviour from Activity Sequences	38
3.4.3	Use of Conversational Technology After Clinical Events	43
3.5	Discussion	44
3.6	Summary	47
4	Predictive Modelling of Cognitive Health from Spoken Language	48

4.1	Overview	49
4.2	Background	50
4.2.1	Cognitive Health Screening	50
4.2.2	Voice Biomarkers	50
4.3	Methods	51
4.3.1	Research Questions	51
4.3.2	Speech Datasets	51
4.3.3	Acoustic and Linguistic Features	52
4.3.4	Predictive Modelling	54
4.3.5	Risk Scores	58
4.4	Results	59
4.4.1	Model Performance in ADRD Detection	59
4.4.2	Correlation with Cognitive Scores	62
4.4.3	Risk Analysis	63
4.4.4	Linguistic Feature Importance	64
4.4.5	Model Performance in Severity Prediction	66
4.4.6	Real-World Pilot Study	68
4.5	Discussion	69
4.6	Summary	73
5	Robot Evaluation and User Engagement in Real-World Settings	74
5.1	Study 1:	
	The Role of Conversational AI in Dementia Care at Home	75
5.1.1	Background	75
5.1.2	Methods	75
5.1.2.1	Research Questions	76
5.1.2.2	Participants	76
5.1.2.3	Participatory Workshops	76
5.1.3	Results	78
5.1.3.1	Perceived Utility for Home Support	78
5.1.3.2	Robot Design Preferences	80
5.1.3.3	Perceived Barriers to Adoption	82
5.1.4	Discussion	84
5.2	Study 2:	
	Cultural Feasibility in India	87
5.2.1	Background	87
5.2.2	Methods	87
5.2.2.1	Research Questions	88
5.2.2.2	Participants and Setting	88

5.2.2.3	Robotic Platforms	89
5.2.2.4	Outcome Measures	90
5.2.3	Results	92
5.2.3.1	Human-Robot Engagement	92
5.2.3.2	Robot Acceptance	94
5.2.3.3	Feasibility and Perceived Utility in India	95
5.2.4	Discussion	98
5.3	Study 3:	
	Promoting Cognitive Health with LLM-Powered Socially Assistive Robots . .	101
5.3.1	Background	101
5.3.2	Methods	103
5.3.2.1	Research Questions	103
5.3.2.2	Conversational SAR	104
5.3.2.3	Interaction Design	105
5.3.2.4	User Study in Retirement Homes	108
5.3.2.5	Linguistic Analysis of Cognitive Tasks	111
5.3.3	Results	112
5.3.3.1	Task Performance	112
5.3.3.2	Verbal Engagement	114
5.3.3.3	Perceived Usability and Robot Acceptance	115
5.3.3.4	User Experience in Robot-Administered Sessions	116
5.3.4	Discussion	120
5.4	Summary	127
6	Conclusion	128
6.1	Dissertation Summary	129
6.2	Limitations	134
6.3	Future Directions	136
6.4	Closing Remarks	138
	References	139
	Appendix	157

List of Figures

1.1	Clinical pathway for ADRD detection, diagnosis and care.	5
2.1	SAR platforms used for health and well-being support in elder care.	14
2.2	Design trends of SARs to support target populations.	15
3.1	Technology used for in-home data collection.	30
3.2	Illustration of the behaviour analysis pipeline.	32
3.3	Topic modelling.	34
3.4	Longitudinal Alexa interaction data.	36
3.5	Engagement with Alexa during and after the novelty period.	37
3.6	Topic distribution.	38
3.7	Probability of behavioural events before or after Alexa use.	39
3.8	Similarity matrices of in-home activity sequences.	40
4.1	Feature extraction from speech and language.	53
4.2	Proposed ML pipeline for cognitive health assessment.	54
4.3	Model calibration and reliability plots.	58
4.4	Confusion matrices for the best-performing models in ADRD detection.	60
4.5	Predicted positive cases per cognitive group.	61
4.6	Distribution of model predictions for MMSE groups.	62
4.7	Risk level distribution by MMSE scores.	63
4.8	SHAP results.	64
4.9	Feature importance results for individual predictions.	65
4.10	Model performance in individual MMSE score prediction.	66
4.11	Model performance in MMSE prediction across cognitive groups.	67
5.1	Activities, material and demos used in the participatory workshops.	78
5.2	Stakeholder preferences for conversational AI features and support functions.	83
5.3	Affective robotic platform tested in India.	89
5.4	Experimental setup in the participatory study in India.	91
5.5	Emotional, verbal and visual engagement with the three conversational robots.	94
5.6	Perceived sociability, ease of use and trust for each robot evaluated.	95
5.7	System implementation of an LLM-powered SAR administering cognitive tasks.	104
5.8	Experimental setup during a picture description task.	105
5.9	Study design with five weekly sessions of interactive cognitive tasks.	109
5.10	Task performance comparison between the first and last study sessions.	113

5.11 Verbal social engagement in study sessions. 115

5.12 Results from acceptance questionnaires. 116

A4.1 Cookie Theft picture. 169

A4.2 Picnic Scene picture. 169

List of Tables

2.1	Conversational SARs evaluated in clinical applications.	16
3.1	Participant cohort and data acquisition.	29
3.2	Behavioural events captured by IoT technologies.	31
3.3	Case study 1 – probability of behavioural events for each cluster identified.	41
3.4	Case study 2 – probability of behavioural events for each cluster identified.	41
3.5	Case study 3 – probability of behavioural events for each cluster identified.	42
4.1	Demographic characteristics of the study cohorts.	52
4.2	DementiaBank datasets.	52
4.3	NLP-based linguistic features extracted for predictive modelling.	55
4.4	ML models and parameters evaluated.	56
4.5	Results of the feature sets tested for ADRD detection in binary classification.	60
4.6	Model performance in binary classification using interpretable linguistic features.	60
4.7	Model performance on demographic group splits.	61
4.8	Model performance on MMSE group splits.	62
4.9	MMSE prediction results for the 10-fold CV.	66
4.10	Severity prediction results using the best-performing RFR-NLP model.	66
4.11	Participant word transcription proportion by cognitive group.	67
5.1	Scenarios and prompts used in participatory workshops.	77
5.2	Results of the OME engagement metrics for the three robots evaluated.	93
5.3	Results from robot acceptance questionnaires.	95
5.4	Prompt used in LLM-powered robot-administered cognitive tasks.	107
5.5	Prompt designed for the LLM-driven moderator.	107
5.6	Demographics and number of animals named in the semantic fluency task.	114
5.7	Perceived ability to focus and promote cognitive health.	117
5.8	HRI experience, perceptions of the robot, and suggested improvements.	119
5.9	Perceived usefulness and interest in continued use.	120
A1.1	Summary of key HRI studies reviewed.	159
A2.1	Comparison analysis of clustering methods.	164
A2.2	Distribution of topic triggers by time of day and novelty period.	164
A2.3	Cluster distributions within the selected time period.	165
A2.4	Cluster distributions outside the selected time period.	165
A3.1	OME metrics used to assess engagement during HRI sessions.	166

A3.2 Engagement metrics assessed from video analysis. 167

A3.3 Items used in the robot acceptance questionnaire. 168

A3.4 OME results in robot-assisted cognitive engagement. 168

A5.1 LLM-generated response parameters. 170

A5.2 Groups used in LLM-generated hints during verbal cognitive tasks. 171

A5.3 Statistics results for linguistic analysis during the picture description task. 172

A5.4 Acceptance scores for robot-administered sessions. 173

A5.5 Spearman's correlation between robot acceptance ratings and BFI personality traits. 174

Acronyms

ADAS-Cog Alzheimer's Disease Assessment Scale-Cognitive Subscale

AD Alzheimer's disease

ADLs Activities of daily living

ADRD Alzheimer's disease and related dementias

AI Artificial intelligence

ASR Automatic speech recognition

AWS Amazon Web Services

BDAE Boston Diagnostic Aphasia Examination

BERT Bidirectional encoder representations from transformers

BFI Big Five Inventory

CI Confidence interval

CN Cognitively normal

CLES Common-language effect size

CV Cross-validation

DRI-CR&T Dementia Research Institute Care Research and Technology Centre

eGeMAPS Extended Geneva Minimalistic Acoustic Parameter Set

FDR False discovery rate

GPT Generative pre-trained transformer

HRI Human-robot interaction

IoT Internet of things

IRB Institutional Review Board

LIWC Linguistic Inquiry and Word Count

LLMs Large language models

LMICs Low- and middle-income countries

LR Logistic regression

MAE Mean absolute error

MCI Mild cognitive impairment

ML Machine learning

MLP Multilayer Perceptron

MMSE Mini-Mental State Examination

MoCA Montreal Cognitive Assessment

MRI Magnetic resonance imaging

NB Naive Bayes

NICE National Institute for Health and Care Excellence

NLP Natural language processing

NPT Neuropsychological tests

OME Observational Measure of Engagement

OpenSMILE open-source Speech and Music Interpretation by Large-space Extraction

PET Positron emission tomography

PIR Passive infrared

PLWD People living with dementia

PPIE Patient and public involvement and engagement

RAG Retrieval Augmented Generation

RF Random Forest

RFR Random Forest Regressor

RMSE Root mean square error

ROC-AUC Receiver operating characteristic area under the curve

RR Ridge regression

SARs Socially assistive robots

SCARF Schizophrenia Research Foundation

SD Standard deviation

SHAP SHapley Additive exPlanations

STT Speech-to-text

SUS System Usability Scale

SVM Support Vector Machine

SVR Support Vector Regression

TF-IDF Term frequency-inverse document frequency

TTS Text-to-speech

WoZ Wizard of Oz

XGBoost Extreme Gradient Boosting Decision Tree

Chapter 1

Introduction

This chapter describes the problem and context of this dissertation. First, this chapter discusses the unmet need for accessible, scalable and cost-effective solutions to support ageing and dementia care at home. Next, it discusses how robotics and conversational artificial intelligence (AI) can enhance well-being, support home safety, and monitor cognitive health. Additionally, this chapter summarises the objectives and main contributions of this work. The chapter concludes with the outline of this dissertation.

1.1 Motivation

A. The Unmet Need

Every country in the world is experiencing an increase in both the number and proportion of people aged 65 years or older. People are living longer, and more are older than ever before. This global trend is driven by declining fertility rates and increasing life expectancy in most countries [252]. This demographic shift in the distribution of the world's population towards older ages – known as population ageing [269] – is set to increase from 10% in 2022 to 16% in 2050¹. Additionally, the number of older adults aged 65 or older globally is set to double during the same period [253]. As a result, the prevalence of dementia is rapidly increasing, with 10 million new cases identified each year [268].

Dementia is one of the major causes of disability and dependency among older people and the seventh leading cause of death globally [268]. Therefore, its early detection and care have become a public health priority [184]. Although ageing amplifies the risk of cognitive impairment, dementia is not a part of normal ageing. Dementia is a progressive neurodegenerative disorder that affects cognitive abilities, including memory, reasoning, and communication skills, leading to a gradual decline in independence, social engagement, and overall quality of life. Still with no cure, dementia most commonly manifests as Alzheimer's disease (AD).

Mild cognitive impairment (MCI) is the intermediate state between normal ageing and neurodegeneration in individuals who maintain the ability to independently perform activities of daily living (ADLs). While MCI is a risk factor for dementia, many individuals remain stable or revert to normal cognition, with 10% to 15% progressing to dementia each year [7]. Today, dementia affects over 55 million people. Ageing projections suggest this will nearly triple by 2050 [178]. In the UK, 25% of hospital beds are occupied by people living with dementia (PLWD) [10]. Half of PLWD will experience at least one hospital admission each year, with 20% of these admissions due to preventable causes such as falls, neuropsychiatric problems or infections [216]. The rise in dementia cases is particularly pronounced in low- and middle-income countries (LMICs), where two thirds of PLWD are projected to reside by 2050 [8]. India is a striking example, with its vast population and high incidence rates, limited resources, poor clinician accessibility, and a staggering 90% gap in diagnosis and care [62].

The growing demand for physical and cognitive support among older adults is outpacing the global supply of long-term care workers [252]. Societies lack the resources to provide residential care for everyone in need [8]. With limited resources for home care services and no immediate cure in sight, the global socioeconomic burden on healthcare systems is only expected to become more critical. This, in turn, places an enormous psychological and financial burden on families and carers who support PLWD in daily living [29].

¹This is measured as the proportion of the global population aged 65 years or older [252].

The dramatic demographic shift toward an ageing population underscores the pressing need for accessible, scalable, and cost-effective solutions that can: 1) provide support and health monitoring at home to enhance well-being and safety; 2) promote cognitive health; and 3) detect early signs of cognitive deterioration, ultimately enabling more personalised care and timely interventions.

B. The Promise of Robotics and Conversational AI

Socially assistive robots (SARs) and conversational AI are recognised for their potential to support ageing and dementia care [87, 170]. SARs interact through social and emotional cues, combining physical embodiment, gestures, facial expressions, and speech. Their embodiment and social presence have been argued to promote user engagement in human-robot interaction (HRI), enjoyment, and task performance compared to disembodied or virtual agents [59, 257, 261]. For instance, previous research has shown that older people were more engaged in verbal dialogues with a physically embodied robot than with a virtual one [179].

Extensive past research has demonstrated SARs' utility in providing home assistance, including medication reminders [28], cognitive stimulation [242], remote therapy [150], companionship [2], and daily living support [217], while reducing the carer burden [39]. Voice interaction, in particular, offers an intuitive, hands-free, and natural user interface, promoting usability among older adults [118] and providing safety in scenarios where voice may be the only form of communication [145]. Furthermore, home-based SARs could provide an accessible user interface for cognitive assessment as a side effect of regular HRI. Recent studies have suggested that collecting speech patterns via mobile apps and conversational AI to detect AD and MCI are feasible, often using memory-probing questions that mimic outpatient consultations [79, 271]. However, there is a significant gap in research evaluating these technologies in real-world settings with older adults and people affected by dementia, particularly as in-residence tools for cognitive support.

SARs and conversational AI have demonstrated potential for digital health interventions, extending their application beyond dementia care to support various neurodegenerative and mental health conditions. For instance, SARs have been used to manage pulmonary disease, improving adherence to medication and rehabilitation exercises [27]. SARs have also been employed as training tools in post-stroke rehabilitation [75], integrated into a cardiac rehabilitation clinical programme [40], and explored for providing in-home therapeutic support among older adults with depression [206]. Furthermore, early evidence suggested the feasibility of using SARs to assist individuals with Parkinson's disease through gamified cognitive-motor training [208]. Additionally, large language model (LLM)-powered conversational agents have recently been developed for personalised nutrition coaching to help manage cardiometabolic conditions such as diabetes and hypertension, leveraging the accessibility and scalability of digital platforms, including mobile apps [273].

There are still several challenges in deploying these systems autonomously and longitudinally in real-world settings, including homes. A significant challenge lies in enabling robots to continuously learn and adapt to individuals over time with personalised actions and behaviours. This is particularly important for people with cognitive impairment, whose abilities, preferences, and overall health needs evolve as their condition progresses. Additional challenges include adoption barriers and privacy concerns that limit long-term user engagement, as well as ensuring the clinical utility of interventions and data collected with such technologies.

C. Enhancing Well-being and Home Safety

When integrated into home environments, conversational AI can offer proactive verbal support by responding to safety risks, such as falls or agitation, querying users to verify health symptoms, and initiating alerts in emergencies [148, 200]. Interactions with conversational AI are often instructive and can encourage behaviours to promote well-being. Such interactions can also provide valuable insights into subjective experiences, cognitive status, and overall well-being – insights that are often missed by passive monitoring technologies. However, the deployment of home monitoring technologies raises important ethical and privacy concerns. If not designed carefully and with stakeholder involvement, these can be perceived as intrusive. Understanding the perceptions of stakeholders, including those with lived experience with dementia, and addressing their concerns from the earliest design stages is crucial to ensure long-term engagement, technology acceptance, and usability.

D. Timely and Scalable Assessment of Cognitive Health

The monitoring of cognitive health is essential for early screening and timely intervention, both clinically and in daily care. The 2024 report of the Lancet Commission on dementia prevention, treatment, and care [153] emphasises the importance of timely diagnosis in supporting the well-being of PLWD and their families, facilitating access to services, and ensuring that individuals can benefit from treatments when they are most likely to be effective. Additionally, timely assessment can help reduce unnecessary hospitalisations and improve overall dementia care.

Dementia diagnoses are typically made years after symptom onset, missing a crucial therapeutic window for intervention, which is becoming increasingly important with the recent emergence of anti-amyloid drugs [173]. Recent trials of amyloid- β -targeting antibodies have shown modest efficacy [233, 254], offering hope for disease-modifying treatments in early-stage AD, which should follow an adequate diagnosis. Figure 1.1 illustrates the clinical pathway for dementia detection, assessment, diagnosis and care, highlighting the need for scalable solutions for timely population screening in community care and home settings.

The use of *digital voice biomarkers* for early screening of Alzheimer's disease and related dementias (ADRD) has gained attention in the research community, driven by the need for

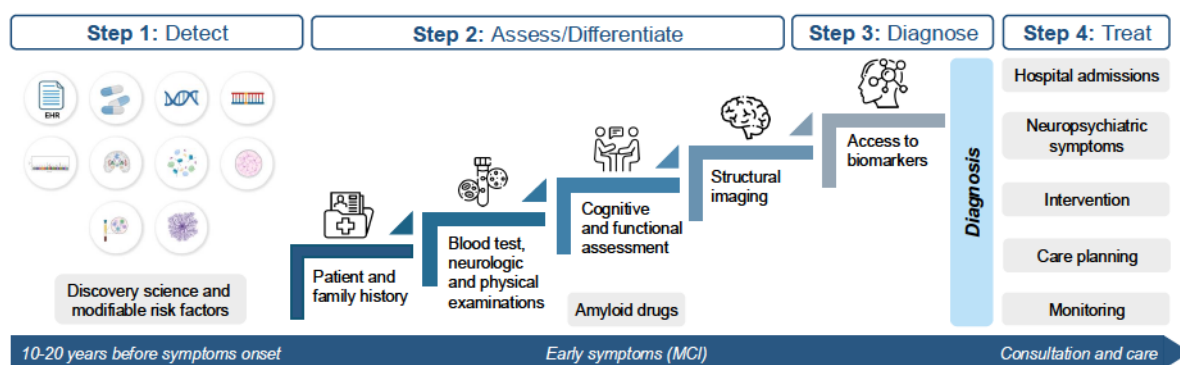


Figure 1.1: Clinical pathway for ADRD detection, diagnosis and care. Step 1 involves early detection through discovery science and identification of modifiable risk factors. Step 2 involves patient history, physical and neurologic examinations, blood analyses, cognitive, functional, and behavioural assessments, and structural imaging. After confirming AD pathology, potential therapeutic interventions, such as amyloid-targeting drugs, are considered. Step 3 highlights the clinical diagnosis using biomarkers. Step 4 focuses on providing appropriate care, including treatment planning, prevention of hospital admissions, monitoring of neuropsychiatric symptoms, cognitive health and overall well-being.

scalable, non-invasive, ecologically valid, and cost-effective solutions for early detection of cognitive decline. Alterations in speech and language can be early predictors of ADRD even before signs of neurodegeneration manifest clinically [4, 77]. This opens new opportunities for the use of conversational AI and SARs to monitor and assess cognitive health from longitudinal verbal interactions. This approach could increase accessibility to earlier detection and intervention. However, the clinical applicability of machine learning (ML) methods for ADRD detection remains unresolved due to limited generalisability, lack of interpretability, and the challenge of obtaining spoken language data for training and validating predictive models.

1.2 Objectives

The goal of this dissertation is to *design and evaluate conversational robots that verbally engage with older populations and PLWD to support well-being, enhance home safety, and monitor cognitive health at home.*

Achieving this goal requires significant interdisciplinary efforts that are guided by specific research questions and objectives, detailed in each chapter. Overall, the objectives of this work were to:

1. Deploy conversational technology in the homes of PLWD to explore long-term user engagement and the potential for behaviour monitoring by analysing activity sequences in smart homes.
2. Evaluate the effectiveness and clinical applicability of ML models for cognitive health assessment from spoken language, using a benchmark DementiaBank dataset, and validate model generalisability using data collected in-residence.
3. Involve stakeholders with lived experience of ageing and dementia in the UK, India and the USA to evaluate technology acceptance, user engagement, and feasibility of home support.
4. Validate the cultural feasibility of robots for dementia care in India.
5. Design an LLM-powered socially assistive robot (SAR) and evaluate its efficacy for providing in-residence cognitive support to older adults.

1.3 Contributions

The work presented in this dissertation follows an iterative process of stakeholder involvement (PLWD, carers, clinicians), technical development, and user evaluation in real-world settings. It offers novel insights and long-term data acquisition for the integration of conversational robots at home to enhance well-being and safety in ageing and dementia care. This work further evaluates predictive modelling of cognitive health from spoken language, with a focus on clinical applicability. This approach could be applied to voice data collected with conversational technology at home to monitor and assess cognitive health. By actively incorporating the voices of stakeholders, this research is shaped by their perspectives from HRI experience and distinct experiences with ageing and dementia. Finally, this work paves the way for a more inclusive and culturally aware robot design and deployment in LMICs, particularly in India. The contributions of this work are summarised below:

Contribution 1: Long-term data acquisition using conversational technology, demonstrating the feasibility of mapping behavioural trends by analysing activity sequences in smart homes.

In a longitudinal deployment study with PLWD at home (N=14), this work evaluated user engagement with commercial conversational technology over time (>1 year in some cases, with an average of over seven months across all households²). By correlating in-home activity detected by Internet of things (IoT) and monitoring devices with the use of conversational technology, this study demonstrated the feasibility of identifying behavioural trends in smart homes. It highlighted the potential of conversational AI in digital health monitoring, specifically for offering proactive verbal support in response to safety risks at home. To the best of knowledge, this was the first study that correlated in-home monitoring data with interactions with conversational technology, targeting ageing and dementia care (Chapter 3).

Contribution 2: Demonstrated the effectiveness of predictive modelling for automated screening of cognitive impairment from spoken language and validated the approach in a real-world context.

This study evaluated predictive models for automated assessment of cognitive health from spoken language, using a benchmark DementiaBank dataset (N=291). This work demonstrated the generalisability and applicability of the ML approach on real-world data collected in-residence (N=22). By identifying the linguistic features most strongly predictive of AD/DRD and performing risk stratification, the interpretability and clinical utility of model predictions were enhanced. The interpretable ML approach showed promise for real-world integration with

²The longest data collection period exceeded one year, with an average deployment duration of over seven months across the 14 households.

in-home conversational technology to monitor and assess cognitive health, enabling earlier detection of higher-risk individuals and timely intervention (Chapter 4).

Contribution 3: Stakeholder involvement to guide the design of conversational AI for well-being and home safety in ageing and dementia care.

By engaging stakeholders (N=20) with lived experience of dementia in participatory workshops, this work underscored the perceived benefits of using conversational AI to support well-being, mitigate safety risks at home, and enhance cognitive engagement. This study presented design guidelines for long-term engagement and appropriate integration of conversational AI into daily living (Chapter 5, Study 1).

Contribution 4: First cultural feasibility study of robots for dementia care in India.

This work designed and validated the first robot for dementia care in India, involving older adults (N=26) and PLWD (N=14) and demonstrating clinical potential for remote cognitive engagement. In a participatory study with stakeholders (N=29), socioeconomic factors and feasibility challenges of deploying conversational robots in India were explored. The study provided novel insights into cultural acceptance, human-robot engagement, and design implications for adapting SARs to support dementia care in the cultural context of India (Chapter 5, Study 2).

Contribution 5: Novel approach to in-residence cognitive support in elder care using LLM-powered SARs.

This study designed an LLM-powered SAR and implemented interactive tasks that involved verbal dialogue based on clinically validated cognitive tools. A user study was conducted with older adults (N=22) in retirement homes, involving three weekly robot-administered sessions to evaluate cognitive task performance, robot acceptance and verbal engagement. This research contributed novel insights into the efficacy of LLM-powered SARs to promote and assess cognitive health in elder care, an approach that had not been tested before (Chapter 5, Study 3).

1.4 Publications

The work presented in this dissertation resulted in the following publications. Some of the work presented has been submitted and is under peer review.

A. Journal publications

1. **Lima, M. R.**, Su, T., Jouati, M., Wairagkar, M., Malhotra, P., Soreq, E., Barnaghi, P. and Vaidyanathan, R. "Discovering behavioural patterns using conversational technology for in-home health and well-being monitoring", In: *IEEE Internet of Things Journal* (2023).
DOI: [10.1109/JIOT.2023.3290833](https://doi.org/10.1109/JIOT.2023.3290833)
2. **Lima, M. R.**, Wairagkar, M., Gupta, M., Rodriguez y Baena, F., Barnaghi, P., Sharp, D. and Vaidyanathan, R. "Conversational affective social robots for ageing and dementia support". In: *IEEE Transactions on Cognitive and Developmental Systems* (2021).
Selected as Research Frontier and featured in IEEE CIS Newsletter
DOI: [10.1109/TCDS.2021.3115228](https://doi.org/10.1109/TCDS.2021.3115228)
3. **Lima, M. R.**, Wairagkar, M., Natarajan, N., Vaitheswaran S. and Vaidyanathan, R. "Robotic telemedicine for mental health: A multimodal approach to improve human-robot engagement", In: *Frontiers in Robotics and AI* (2021).
DOI: [10.3389/frobt.2021.618866](https://doi.org/10.3389/frobt.2021.618866)
4. Wairagkar, M., **Lima, M. R.** et al., "Emotive response to a hybrid-face robot and translation to consumer social robots". In: *IEEE Internet of Things Journal* (2021).
DOI: [10.1109/JIOT.2021.3097592](https://doi.org/10.1109/JIOT.2021.3097592)
5. Natarajan, N., Vaitheswaran, S., **Lima, M. R.**, Wairagkar, M. and Vaidyanathan, R. In: "Acceptability of social robots and adaptation of hybrid-face robot for dementia care in India: A qualitative study", In: *American Journal of Geriatric Psychiatry* (2021).
DOI: [10.1016/j.jagp.2021.05.003](https://doi.org/10.1016/j.jagp.2021.05.003)

B. Conference publications

1. **Lima, M. R.**, O'Connell, A., Zhou, F., Nagahara, A., Hulyalkar, A., Deshpande, A., Thomason, J., Vaidyanathan, R. and Matarić, M. "Promoting cognitive health in elder care with large language model-powered socially assistive robots". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2025.
DOI: [10.1145/3706598.3713582](https://doi.org/10.1145/3706598.3713582)

2. **Lima, M. R.** "Home integration of conversational robots to enhance ageing and dementia care". In: *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024.
DOI: [10.1145/3610978.3638378](https://doi.org/10.1145/3610978.3638378)
3. **Lima, M. R.**, Horrocks, S., Daniels, S., Lamptey, M., Harrison, M., and Vaidyanathan, R. "The role of conversational AI in ageing and dementia care at home: A participatory study". In: *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2023.
DOI: [10.1109/RO-MAN57019.2023.10309459](https://doi.org/10.1109/RO-MAN57019.2023.10309459)
4. **Lima, M. R.**, Soreq, E., Daniels, S., Barnaghi, P., Vaitheswaran, S. and Vaidyanathan, R. "Conversational AI for monitoring behaviour and well-being of people living with dementia", In: *Alzheimer's & Dementia*, 2023.
DOI: [10.1002/alz.082077](https://doi.org/10.1002/alz.082077)

C. Submitted publications

1. **Lima, M. R.**, Capstick, A., Geranmayeh, F., Nilforooshan, R., Matarić, M., Vaidyanathan, R. and Barnaghi, P. "Evaluating spoken language as a biomarker for automated screening of cognitive impairment".
Submitted to: *Nature Communications Medicine*.
arXiv preprint: [arXiv:2501.18731](https://arxiv.org/abs/2501.18731)
2. **Lima, M. R.**, Srinivasan, N., Daniels, S., Vaitheswaran, S. and Vaidyanathan, "Cultural feasibility of conversational robots for dementia care in India".
Submitted to: *Innovation in Aging*.

1.5 Dissertation Outline

This section outlines the structure of the dissertation and provides a summary of each chapter. Each chapter addresses specific objectives and research questions and includes a review of the relevant literature.

- Chapter 2** reviews the state of the art in SARs and conversational technology to support the well-being of older adults and PLWD, and ML approaches for automated assessment of cognitive health. Additionally, the chapter identifies the major research challenges in the field, which this work aims to address.
- Chapter 3** describes a longitudinal deployment of conversational technology in homes of PLWD to explore user engagement and analyse behaviour from activity data in smart homes.
- Chapter 4** introduces an interpretable ML approach for automated assessment of cognitive health from spoken language, with a focus on clinical applicability.
- Chapter 5** presents three user studies conducted in the UK, India and the USA, involving stakeholders (older adults, PLWD, carers, clinicians and dementia care professionals) to evaluate user engagement, acceptance and the feasibility of using conversational AI and SARs in residential settings: *Study 1* explores the role of conversational AI in ageing and dementia care at home; *Study 2* assesses the cultural feasibility in India; and *Study 3* introduces an LLM-powered SAR for in-residence cognitive support.
- Chapter 6** summarises the novel findings and main contributions, discusses limitations and future research directions, and concludes the dissertation.

Chapter 2

Related Work

This chapter reviews the state of the art, design trends, and current applications of SARs and conversational technology to support older adults and PLWD in daily living. Additionally, this chapter describes ML techniques to analyse speech and language for automated cognitive health assessment. Finally, this chapter discusses open research challenges to overcome existing gaps in real-world deployments, long-term engagement between SARs and target populations, adaptation to individual needs and preferences for personalised care, and clinical translation.

Part of the comprehensive review presented in this chapter was published in *IEEE Transactions on Cognitive and Developmental Systems*, 2021 [149]. However, the chapter as a whole situates the dissertation within the current state of the art, reflecting the most recent advancements in the field.

2.1 Socially Assistive Robots in Older Age and Dementia Care

The potential of SARs to support the well-being and independence of older adults and PLWD is well noted in the literature [1, 87, 170]. SAR platforms have been introduced, from mechanically complex [26, 242], mobile [20, 226], embodied humanoid [189, 195], to pet-like robots [199, 260], simpler virtual assistants [212], and commercial smart speakers (e.g., Amazon Alexa, Google Home). Literature on robots for dementia care has argued that SARs can improve attention [242] and neuropsychiatric symptoms [163], reduce agitation [194] and depressive symptoms [195], and establish rapport [2]. However, the development of appropriate robots for dementia care is still in an early stage, with few examples beyond research studies. SARs can be categorised into four roles to support target populations:

1. *Companionship*. Ability to engage or entertain users through conversations about topics of interest or providing personalised recommendations, ultimately reducing feelings of social isolation.
2. *Health monitoring*. Ability to monitor health and well-being, give reminders (e.g., medication), and inform carers or clinicians of potential safety risks at home.
3. *Cognitive stimulation*. Ability to engage in cognitive tasks or games, monitor performance over time, ultimately detecting signs of cognitive decline.
4. *Clinical therapy*. Ability to conduct or assist in cognitive interventions and therapy, autonomously or by teleoperation.

Figure 2.1 shows relevant examples clustered by mobile (i.e., capable of autonomous or teleoperated navigation at home or in assisted living facilities) and small affective SARs (i.e., often simpler in their design, stationary or portable robots).

Interest in and acceptance of SARs among older adults and PLWD is increasing [123], highlighting the need to address ethical considerations, which encompass both *research ethics* and *societal ethics*. Research ethics focus on ensuring informed consent, transparency in the objectives and outcomes of the study, and data privacy. In particular, privacy concerns related to voice data collection using SARs and potential misuse of personal information remain significant among older adults [107, 149, 229, 234], which represents a barrier to technology adoption [118]. Societal ethics, on the other hand, address the broader implications and concerns of integrating SARs into daily life, including intrusiveness at home, loss of privacy, reduced human contact, emotional deception when user expectations of the robot are not met, and attachment to the robot, which can lead to emotional distress [255]. Furthermore, the importance of user trust in SARs for sustained human-robot engagement is well-recognised [141]. For carers, trust in the technology is equally necessary, particularly in its ability to provide safe and effective support without compromising autonomy and well-being [90].

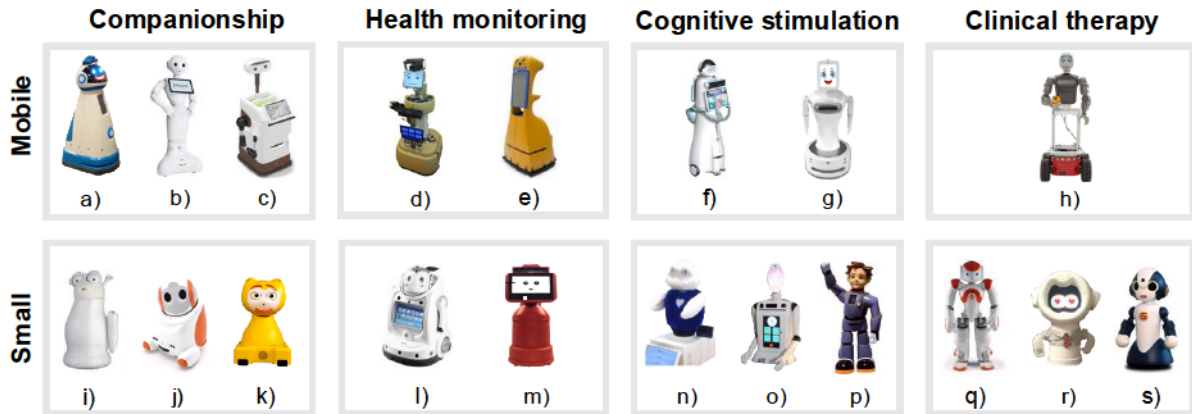


Figure 2.1: SAR platforms used for health and well-being support in elder care. a) GrowMu [196], b) Pepper [189], c) Hobbit [20], d) RAMCIP [84], e) CompanionAble [226], f) Kompai [33], g) Silbot [129], h) Bandit [242], i) Reeti [213], j) PaPeRo [116], k) iCat [116] l) IroblQ [28], m) Screening robot [65], n) Mini [220], o) Ryan [64], p) Milo [197], q) NAO [195], r) Eva [56], s) Sota [224].

A. Comprehensive Review Protocol

Existing surveys evaluated SARs to support target populations [1, 23, 87, 243, 259], including home-based robots for cognitive support [256]. However, no previous reviews targeted the depth of the field in relation to conversational capacity, multimodal interaction, and user engagement. This review aims to fill this gap by surveying conversational SARs specifically targeted at supporting ageing and dementia, both in social and clinical contexts.

Studies published between 2000 and 2021 were selected from scientific publication databases such as IEEE Xplore, ACM, PubMed, Google Scholar, and search keywords related to: “social robot”, “conversational robot”, “elderly”, “dementia”, “verbal communication”, and “dialogue system”. The inclusion criteria consist of HRI studies that: tested or evaluated SARs with target populations; used an embodied SAR with multimodal interactive features, including facial expressions and speech. Studies were excluded if: there were no results from user evaluation; presented work on SARs without verbal communication (e.g., the robotic seal Paro [171]). The most recent study of each robotic platform was selected. For studies published in the same year using the same robot, those with a larger number of participants and longer HRI duration were prioritised. In total, 30 studies that evaluated SARs with target populations were reviewed (see details in Appendix A1).

B. User-Centred Design

Designing robots for end-users. HRI research often employs human-centred design methods to understand individual needs, preferences and technology expectations, which are translated into the conceptualisation of new design features and guidelines [187]. There has been a growing interest in applying user-centred design approaches to design SARs that meet the needs and preferences of end-users for longer-term engagement (see Figure 2.2a).

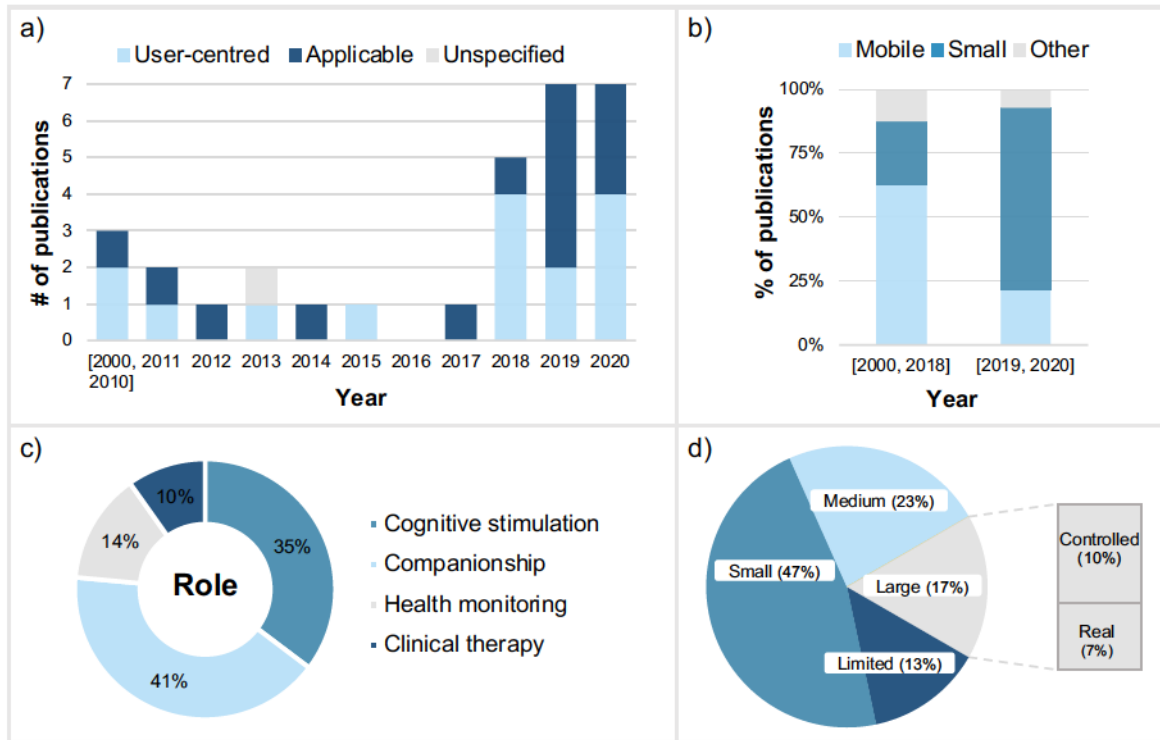


Figure 2.2: Design trends of SARs to support target populations. a) Number of studies per year: robots specifically designed to assist the target population (user-centred); robots designed for other purposes but evaluated with target populations (applicable); unclear robot design purpose (unspecified); b) Percentage of publications per interval using mobile, small, or other type of SARs (e.g., torso); c) Distribution of support roles; d) User evaluation with limited (1-5), small (6-20), medium (21-40), and large (>40) sample sizes. For HRI studies with large sample sizes, the percentage of robots tested in controlled vs. real-world settings is presented.

Designing robots with end-users. Participatory design refers to actively engaging stakeholders in the design of useful technologies while amplifying their voices [235]. Previous studies have used participatory methods to explore the use of SARs to support older adults with age-related health needs and MCI [83, 185], and carers [169, 265] in various contexts, including social engagement [186], mental health support [134], rehabilitation [152], physical assistance in daily tasks [35], home cognitive training [125], and long-term care [155]. These studies gathered feedback in workshops with stakeholders, surveys, focus groups, or user experience with technology to recommend design considerations. For instance, a recent participatory study [135] explored SARs for PLWD and carers, including their conflicting views on robot design preferences, and suggested slow communication as a key design feature.

C. Multimodality

Previous work has argued that physical embodiment and social presence can enhance engagement and enjoyment in HRI [59, 117, 261]. Tapus et al. [241] defined *engagement* as a sustained collaborative connection between humans and robots. Heerink et al. [96] reported social presence as a key robot design feature for older people, while Nishio et al. [179] showed

Table 2.1: Conversational SARs evaluated in clinical applications.

Robot	Clinical application	Participants and setting
Eva [56]	Cognitive stimulation therapy	N=9 PLWD; nursing home
NAO [195]	Memory training	N=21 people with MCI; clinical centre for dementia
Bandit [242]	Cognitive stimulation and engagement	N=10 PLWD; senior living facility
Sota [224]	Memory and dance therapy	N=71 older adults, PLWD; setting unspecified
Ryan [64]	Cognitive behavioural therapy	N=4 older adults with mild to moderate depression; senior living facility
Screening robot [65]	Home-based cognitive and well-being screening	N=30 older adults; setting unspecified

that older people were more verbally engaged with an embodied robot than a virtual one.

There has been a trend in designing smaller, mechanically simpler, often stationary, and more affordable robotic platforms. Whilst around 25% of the studies reviewed between 2000 and 2018 used a simpler and smaller robot, this figure increased to more than 70% in the most recent studies (see Figure 2.2b). Smaller affective robots are easier to deploy in real-world settings, such as homes, and are typically more cost-effective. This design trade-off promotes the deployment of SARs worldwide, including in LMICs. This comes with the need to enhance multimodal interactive features for engaging HRI, such as facial expressions, gestures, and speech. Furthermore, previous research has suggested that smaller affective robots increase acceptance among older adults, as larger mobile platforms may be perceived as intimidating or raise privacy concerns [221]. Importantly, for user acceptance and trust, the robot's appearance must align with its functionality [123].

D. Support Roles

From the four main roles of SARs for health and well-being support (described in Section 2.1) and considering that each robot may combine multiple roles to provide assistance, most studies reviewed aimed at providing companionship (41%) and cognitive stimulation (35%) (see Figure 2.2c). Some of the robotic platforms reviewed were able to monitor health (14%). Furthermore, only six of the 30 reviewed studies were designed to support users through cognitive stimulation, therapy, or clinical screening, as detailed in Table 2.1. There is a clear gap in the clinical translation of SARs for dementia care.

For instance, the humanoid robot Bandit was tested in weekly cognitive therapy sessions with 10 PLWD for six months [242]. The robot improved participants' engagement, attention, and performance in a music game through facial expressions, gestures, and pre-recorded audio. The study employed an adaptive framework to adjust the game's difficulty to individual abilities. This robot was further tested with 33 older adults in simple physical exercises designed for therapeutic intervention [74]. The robot Eva was also evaluated in cognitive

stimulation group therapy sessions over nine weeks [56]. The authors assessed dementia-related behavioural symptoms before and after sessions using questionnaires and gathered qualitative feedback from carers. The study reported a decrease in delusion, agitation, and exaltation and positive short-term effects on mood after robot-guided sessions. Furthermore, the commercial social robot NAO was evaluated as a memory trainer with 21 participants diagnosed with MCI, demonstrating that robot-assisted activities increased attention and decreased depressive symptoms [195].

E. Cultural Adaptation

Previous research suggested that cultural factors influence robot acceptance, highlighting the need to incorporate cultural considerations into robot design [190, 218]. People's perceptions and preferences for robot design vary between cultures [133]. Implicit cultural biases toward robots have been studied to develop culture-aware robots [72, 210]. Previous research has proposed the integration of cultural knowledge in SARs to support individual cultural practices [209]. However, cross-cultural testing of SARs and the development of culturally aware interactions are still rare. Only three of the 30 HRI studies reviewed using SAR platforms included cultural adaptation [129, 196, 219].

Cultural bias in AI systems is a broader issue that affects their effectiveness and acceptance across diverse populations. For instance, previous research showed that large language models (LLMs) trained predominantly on English data reflect biases favouring Western cultural values, particularly when prompted in English [37, 240].

There is a lack of studies exploring SARs for use in LMICs. From the limited studies in robotics for dementia care that have involved stakeholders [35, 83, 135, 169, 265], none have addressed the HRI cultural adaptation needed and the feasibility challenges inhibiting real-world deployments in LMICs. Furthermore, recent advancements in LLMs have expanded multilingual capabilities for robot speech recognition and synthesis. This could enhance user engagement and acceptance of SARs in diverse cultural settings, an area that remains underexplored.

In the context of this dissertation, *cultural adaptation* refers to exploring potential modifications to the robot's design, including its appearance, behaviour, and interaction style, to align with cultural norms and user expectations, targeting sustained engagement and acceptance. *Cultural translation*, on the other hand, explores the integration of SARs into a socio-cultural context, understanding the robot's role and ensuring its utility in daily life. This process considers technology accessibility, caregiving practices, and social structures, including family and healthcare systems.

F. HRI Evaluation

Literature lacks longitudinal robot evaluation studies in real-world settings. In this comprehensive review, sample sizes were categorised as limited (1-5), small (6-20), medium (21-40) and large (>40) based on the number of end-users who tested the technology. The duration of each study was also examined, considering either the total number of HRI sessions or the length of the study in days, weeks, months, or years. Finally, the study setting was explored, distinguishing between controlled environments (e.g., laboratories, simulated home environments) and real-world settings (e.g., homes, senior care facilities).

More than half of the reviewed studies (60%) involved small or limited sample sizes and only 17% of conversational SARs were evaluated with 40 or more participants (see Figure 2.2d). However, none of these involved long-term HRI. Instead, they encompassed a single session [61], one or two sessions per week for up to four months [219, 224], or occasional short interactions with robots placed in senior care facilities for up to two weeks [95, 196]. In the latter, interested users would come closer and greet the robot.

Some of the SARs reviewed have been deployed in real-world settings, specifically in homes [2, 20, 28, 33, 115, 213, 239], clinical or care centres [95, 98, 111, 166, 196]. However, only two of these were conducted with a large sample size and the robot operating autonomously [95, 196]. Many of the studies reviewed were limited to a small number of sessions with the robot (e.g., [195]) or short-term interactions lasting less than a week (e.g., [166]), rather than sustained engagement over time. One major barrier to conducting HRI longitudinal trials lies in obtaining ethical approval for data collection with vulnerable populations such as PLWD. Addressing privacy concerns and information governance for handling sensitive data (e.g., voice and video) is critical to ensuring ethically safe deployments.

2.2 Conversational AI for Health and Well-being Support

There has been an emerging interest in the application of conversational technology in health-care applications, including to support older people at home [128]. The benefits of bringing natural language capabilities into robotics have been reported in the literature [243]. By understanding and responding to natural language, conversational AI can be accessible and personalised companions with potential to monitor health and encourage behaviour [63].

Conversational technology can be categorised into: 1) conversational SARs, which are characterised by a physical presence and multimodal interactive features beyond speech, e.g., facial expressions, embodiment, or gestures; 2) conversational agents and smart speakers, which primarily rely on voice. These systems integrate: automatic speech recognition, which converts input speech into textual vector representations; natural language understanding of user intents; dialogue management to interpret information and keep track of conversation flow, often with a knowledge base to store and retrieve contextual information; and natural language generation to transform the output of dialogue management into textual format, which is converted to spoken language by text-to-speech (TTS) synthesis.

In light of commercial viability, increased worldwide adoption and expanded AI capabilities, conversational technology holds promise to assist older people and PLWD at home. Commercially available smart speakers, such as Amazon Alexa and Google Home, have recently been explored as assistive tools to promote the independence of older populations for routine management [39], remote care [9], or self-management of diabetes [47]. Alexa skills have been proposed to support older adults with ADLs and health monitoring, including medication reminders [9], dressing assistance, depression screening [238], diet tracking, and fall detection [131].

2.2.1 Communication and Adoption Challenges

There has been a growing interest in exploring how conversational AI is used at home by target populations. Qualitative studies have revealed initial insights into user experience with conversational technology, patterns of daily use, and why interest is often lost after the *novelty effect* [118, 227, 248]. Current barriers to long-term adoption have been reported in the literature [104], particularly:

- limitations and errors in speech recognition for effective verbal interaction;
- limited adaptation of interactions to individual needs and cognitive abilities over time;
- privacy concerns related to voice data collection;

The readiness of conversational agents in clinical settings is challenged by additional factors, including the need for persistent Internet connection, user and organisational compliance to exchange personal health information, and the risk of delivering misleading information [230].

Previous research explored communication challenges faced by older adults with MCI [120], highlighting the need for conversational AI to handle user pauses, hesitations, and pre-emptive responses (e.g., when participants respond before the system finishes the sentence), provide instructions for recovering conversation when error messages occur, and assist with confirming responses using both voice and visual prompts.

2.2.2 Opportunities and Risks of LLMs

LLMs are transformer-based models, such as bidirectional encoder representations from transformers (BERT) [60] and OpenAI's generative pre-trained transformer (GPT) [203], trained on large datasets to perform natural language processing (NLP) tasks, including question-answering, translation, summarisation, and content generation [244]. LLMs are largely general purpose in that they are trained with unrestricted text. Fine-tuning LLMs for specific domains, such as mental health, enables them to facilitate natural language interactions similar to those used in mental health assessments and interventions.

Among different applications, one intriguing question yet to be answered is how well LLMs can understand and support mental health and well-being. Furthermore, integrating LLMs with SARs opens new opportunities for human-robot engagement. LLM-powered robots have elicited increased trust in collaborative tasks with humans due to their ability to facilitate more intuitive and natural language interactions [275].

A. Mental Health Assessment

The use of LLMs in mental health care has significant potential to assess symptoms, diagnose conditions, and monitor changes in mental well-being [130]. LLMs have shown the capacity to answer medical questions and assess conditions, sometimes outperforming human clinicians [130]. LLMs have also been explored for screening depressive symptoms and suicide risk, yet researchers observed varying accuracy [80, 270]. Furthermore, LLMs can be trained to ask relevant questions that screen for mental health concerns or predict diagnoses.

Especially promising is the use of LLMs to process symptom descriptions and incorporate patient history and other monitoring data, such as passive sensing data, to support screening and diagnosis [130]. Early evidence suggests that LLM-based chatbots may reduce symptoms of depression, anxiety, and stress [183]. For instance, a chatbot trained on domain-specific health data to engage in mental health assessments has been shown to ask more detailed questions about certain symptoms (e.g., sleep disturbances) compared to human psychiatrists [41]. Research has argued that these chatbots can be trained to establish rapport with users, provide non-judgemental feedback, and sustain therapeutic conversations [137, 158, 164].

B. LLMs in Dementia Research

LLMs hold promise for positive impact in dementia care by enhancing digital screening, promoting cognitive health, enriching social interactions, and supporting carers [249]. Of particular relevance to this dissertation, LLMs hold potential to extract language patterns for automated screening and cognitive decline prediction (see details in Section 2.3). For instance, GPT text embeddings have been used to classify AD from healthy controls [3]. Language models can also summarise patient-doctor conversations [279] and code adverse events in patient narratives [50]. Furthermore, a recent study explored LLMs to generate synthetic clinical data for AD detection, aiming to address the scarcity and imbalance of existing datasets for ADRD research [143].

LLM-powered SARs could further provide emotional support to PLWD and promote cognitive health through human-robot engagement [102]. Additional assistance for PLWD includes cognitive behavioural therapy via phone apps [183], as well as speech-assisted navigation [249].

C. Ethical Concerns and Responsible Use of LLMs

The responsible development, evaluation, and use of LLMs is critical. The knowledge embedded within an LLM is limited to the training and fine-tuning data [85, 130]. Therefore, applying LLMs in healthcare settings presents challenges in ensuring the accuracy, reliability, safety, and adherence to the latest medical guidelines in the information they provide.

As a subset of the ethical considerations discussed in Chapter 2.1, the deployment of LLMs involve both *research* and *societal* ethics. From a research perspective, ensuring the quality and representativeness of training data is essential to minimise the risk of producing inaccurate, biased, or harmful information, including the generation of syntactically sound text that is factually incorrect, known as hallucinations. This is particularly critical in the medical domain, where delivering accurate information is essential. Rigorous evaluation and validation of LLMs are necessary to ensure their accuracy, reliability, and interpretability. This, in turn, can help address the 'black box' nature of these models, facilitating a deeper understanding of their reasoning process to generate outputs that are often presented as recommendations [244].

Furthermore, LLMs may provide different answers to the same prompt in repeated attempts, raising concerns about their reliability [94]. From a societal ethics standpoint, the issue of inconsistent reliability of LLMs can undermine trust and potentially cause harm, including misdiagnoses, inappropriate treatment recommendations, or dismissive replies to health risks [164]. These risks underscore the importance for these models to operate within their knowledge scope and the need for high-quality training data, especially in clinical applications.

2.3 Machine Learning for Cognitive Health Assessment

Timely and accurate assessment of cognitive health is a major unmet need [67]. The development of cost-effective and automated screening tools is crucial for timely intervention, both clinically and in daily care. There has been recent interest in deriving early speech and language features of ADRD as *digital voice biomarkers*, which can be collected in real-world settings outside clinics. Increasing evidence suggests that speech and language can be early predictors of cognitive decline in the early pre-clinical stages of ADRD [4, 77, 82, 122].

Deep learning techniques have been utilised for automatic feature extraction with pre-trained models for audio and text representation [57]. In particular, recent research has explored the potential of LLMs for automated cognitive assessment [3, 11]. These transformer-based models can automatically capture subtle language patterns potentially missed by conventional methods. However, their lack of explainability hinders clinical applicability. Another advantage of deep learning approaches is the ability to extract multilingual embeddings. This is an area of active research [81] and could help address the limited sample sizes in existing speech datasets for ADRD research.

Following feature extraction methods, emerging evidence supports the feasibility and reliability of ML in detecting ADRD and modelling disease progression. For instance, a logistic regression model trained on embedding vectors from neuropsychological tests (NPT) transcripts and demographic data achieved an accuracy of 78.5% and a sensitivity of 81.1% in predicting AD progression within six years [11]. Furthermore, a logistic regression model using acoustic and linguistic features extracted from picture description tasks during NPT achieved an accuracy of 81.9% in binary AD classification [78]. With factor analysis, the authors found four patterns of variance among participants: semantic, syntactic and information impairment, along with acoustic abnormality. Classification models based on conventional acoustic features automatically extracted from speech demonstrated an accuracy of 71.3% [91]. Moreover, fine-tuning transformer-based language models on text transcripts from picture descriptions achieved an accuracy of 89.6%.

Although previous studies have primarily focused on analysing speech and language from voice recordings of NPT [245], their use in real-world settings in pre-clinical populations remains underexplored. Recent studies have suggested the feasibility of collecting speech via mobile applications and voice assistants to detect MCI and ADRD [79, 121, 271]. This opens new opportunities for the use of in-home conversational technologies to monitor cognitive health.

2.4 Research Challenges

SARs and conversational AI offer a promising interface for verbal engagement and support in daily living. However, these technologies have yet to demonstrate a direct impact in supporting older adults and PLWD at home, particularly with long-term user benefits and clinical utility. There is a strong need for systems capable of continuous interaction to store information and adapt to changing health conditions, respond to safety risks (e.g., offering verbal guidance in the event of perceived agitation or confusion), and ultimately enhance independence at home while relieving the burden on carers.

The critical areas in need of further investigation and development are outlined below:

Challenge 1. Longitudinal Deployment and User Engagement

There is a significant gap in longitudinal data collection using conversational technology at home. Privacy concerns and the challenge of maintaining user engagement after the initial novelty phase need to be addressed. It is essential to involve stakeholders, including those with lived experience of dementia, in every stage from design to deployment to identify meaningful support scenarios, address user concerns, and design non-intrusive and safe HRI for long-term utility at home. Additionally, cultural feasibility challenges must be considered for successful deployment across diverse cultural and economic contexts.

Challenge 2. Intelligent Adaptation

Adapting HRI to individual health and cognitive needs is critical for personalised care. Integration with home monitoring devices can enhance adaptive capabilities based on nonverbal indicators of physical or mental state, yet this remains a complex challenge. While speech and language patterns hold promise for inferring mood and cognitive state from HRI, reliable ground truth data for comparison remain scarce. Further participatory studies are essential to tailor natural language interactions to the evolving needs and expectations of target populations.

Challenge 3. Clinical Translation

Research is needed on effective in-home data collection using conversational technology to capture meaningful and actionable data with clinical utility. Despite the potential of digital voice biomarkers, the clinical applicability of ML methods for automated cognitive assessment remains unresolved due to limited generalisability, lack of interpretability, and the challenge of obtaining spoken language data for training and validating predictive models. Regular speech collected from HRI at home presents a promising direction to promote and monitor cognitive health over time.

Behaviour Monitoring with Conversational Technology at Home

This chapter describes a longitudinal data collection and deployment of commercial conversational technology (Amazon Alexa) in the homes of PLWD (N=14). In-home activity and interaction data were collected over >1 year in some cases, with an average of over seven months across all households. The study evaluated user engagement with technology beyond the novelty phase. Additionally, it demonstrated the feasibility of mapping behaviour in smart homes by correlating activity data from remote monitoring technologies with interactions with a conversational agent. This research suggests the promise of conversational AI in digital health monitoring, specifically for providing automated verbal support and alerts in response to safety risks at home.

The work presented in this chapter was published in *IEEE Internet of Things Journal*, 2023 [148], presented as an abstract at the *Alzheimer's Association International Conference*, 2023 [147], and in oral form at the *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2024 [144].

3.1 Overview

The ability to map in-home behaviour and detect deviations or changes from previously observed patterns can be leveraged to provide proactive support through personalised verbal interactions. Conversational agents could directly query users about their subjective perceptions of health and well-being when behavioural changes are detected and inform stakeholders to enable timely interventions. Such interactions could also provide valuable insights into subjective experiences, cognitive status, and overall well-being – insights that are often missed by passive monitoring technologies.

This study explored the integration of conversational technology in smart homes and the potential of behaviour monitoring to support well-being and home safety in dementia care. In-home activities detected by remote monitoring technologies (e.g., motion, use of kitchen appliances, vitals) were correlated with interactions with Alexa. The analysis inspected: long-term user engagement with Alexa, particularly after the novelty phase; activity sequences to identify behavioural trends; and whether participants continued to engage with Alexa after clinical events, which were logged by a monitoring team in contact with participants.

The protocol for Alexa deployment in households of PLWD and data collection using monitoring technologies were conducted as part of the Minder study, an in-home monitoring platform developed by the UK Dementia Research Institute Care Research and Technology Centre (DRI-CR&T) to passively collect behaviour data, detect warning signs, and alert the integrated NHS team to facilitate early interventions. The contributions of this work, which analysed in-home behaviour using conversational technology, are described below.

Contributions:

1. Long-term data collection using conversational technology in the homes of PLWD.
2. Demonstrated technical feasibility of identifying behavioural trends in smart homes by analysing activity sequences captured from remote monitoring technologies and interactions with a conversational agent.
3. The proposed approach highlighted the potential for conversational AI to provide proactive verbal support and automated alerts in response to safety risks at home.

Ethics Approval: The study was approved by the NHS Health Research Authority London-Surrey Borders Research Ethics Committee (Minder Health Management Study); TIHM 1.5 REC 19/LO/0102.

Code Availability: The code used in this study is hosted on Code Ocean, a computational reproducibility platform (<https://codeocean.com/capsule/7912475/tree/v1>).

3.2 Background

3.2.1 Home Activity Monitoring

Opportunities in the use of smart home technology for older adults and PLWD are well noted in the literature: from diagnostic assessment to tailored care, health monitoring, cognitive support, and completion of ADLs [17, 170, 172, 200]. Advances in IoT technologies have spurred significant progress in activity recognition, habit assessment, and anomaly detection within smart homes [5, 54, 58, 228, 276], including monitoring systems that aim to support independent living of older adults and people affected by dementia [14, 15, 36, 154, 231]. If not designed carefully and with end-user involvement, these can be perceived as complex and intrusive and may raise ethical concerns regarding privacy [6, 48, 149, 172, 180].

Models capable of recognising behaviour are of particular relevance to this study. Home activity data can be used to analyse activity sequences with temporal and spatial information, such as a user's morning routine [136, 168]. This analysis can detect changes in behaviour over time, which may indicate shifts in lifestyle, functional or cognitive abilities [228, 278].

Recent work has proposed habit representation methods using activity data collected from smart environments, focusing on the sequence and duration of activities [136]. Along similar lines, a real-time monitoring framework has been proposed to recognize habits and detect anomalies to support seniors [168], yet no results were obtained from data collection with PLWD. Furthermore, sequence comparison and clustering methods have been applied to activity vectors to identify daily routines, indicating the potential to assist with ADLs [228].

Further studies have analysed abnormal behaviour of PLWD, identifying differences in daily routine patterns [14]. However, this analysis was limited to a dataset of three households. A common gap identified in these studies is the lack of longitudinal data collected from target populations in real-world contexts. Researchers have also explored correlations between changes in daily routine and alterations in cognitive and physical health [55]. The authors evaluated the approach using continuous data collected from smart home sensors and 18 senior residents. While there has been work in each of these areas individually, to the best of knowledge, no research to date has correlated in-home activity captured by IoT technologies with voice interactions with conversational AI to track behaviour.

3.2.2 Similarity Methods

The investigation of ADLs to understand human behaviour includes sequence comparison and clustering approaches to identify patterns from activity sequences captured by IoT technologies [126, 228]. Inspired by sequence mining algorithms used in bioinformatics to analyse gene sequences [160, 161], researchers have applied similar techniques to study human activity. Different similarity measures have been studied in the context of sequence analysis, which

can be categorised as follows: distances between probability distributions, counts of common attributes, and optimal matching between sequences considering the necessary operations to transform one sequence into another [223, 236].

The choice of a suitable similarity measure for comparing activity sequences largely depends on the sequential features being considered (e.g., temporal information, duration, and order) [207, 228]. Popular distance metrics are used to calculate the similarity between pairs of categorical sequences, such as the *Hamming distance* [92], which calculates position-wise comparisons of pairs of sequences of equal length, and the *Levenshtein distance* [140], given by the smallest number of editing operations needed to turn one sequence into another. However, these metrics are not suitable for dealing with temporal sequences, where common and consecutive elements should be considered. An alternative approach has been proposed to capture the sequentiality of events [86].

3.2.3 Topic Modelling

Topic modelling is an unsupervised learning technique used to identify hidden patterns in a text corpus [100]. This technique can be applied to analyse interactions with conversational AI and inspect user interests and preferences over time.

Conventional models, typically based on Latent Dirichlet Allocation (LDA), employ a bag-of-words model, in which each unique word is modelled independently of the others [44]. However, these models involve simplistic assumptions and pre-processing steps that often dismiss semantic relationships between words, resulting in less coherent and interpretable topics [46, 177]. With recent developments in NLP, pre-trained language models have been proposed to capture semantic and contextual information from text more effectively (e.g., BERT [113], GPT [204]). Similarly, Top2Vec [12] and BERTopic [89] have been proposed to infer topics while keeping the original text structure [66, 142].

3.3 Methods

The work presented in this chapter was conducted as part of the UK DRI-CR&T's Minder study¹. Minder is a digital care platform designed to deliver personalised care while promoting the independence and safety of PLWD at home. The platform leverages in-home monitoring technologies to passively collect behavioural data, monitor health and well-being, and identify potential risks that could lead to hospitalisation, such as falls, neuropsychiatric problems or infection. Minder aims to support timely interventions and facilitates continuous interaction between PLWD, carers, and a dedicated NHS team to enhance clinical care.

A. Terms and Definitions

Behavioural event. In this study, a behavioural event is defined as a sensor trigger captured in a smart home with an associated timestamp. Each behavioural event indicates an activity (e.g., motion in the house, taking vitals, interacting with a voice agent).

Activity sequence. This study explores in-home daily activity using conversational technology. An activity sequence is defined as a sequence of ordered behavioural events that occur in the 10-minute period (defined based on domain knowledge) preceding or following Alexa use, with an associated start and end timestamp (elaborated in Section 3.3.4).

Behavioural pattern. A behavioural pattern is defined as a set of activity sequences with a high degree of similarity that occur for at least three weeks, determined by domain knowledge (e.g., a user takes vitals in the morning before interacting with Alexa for a month). Details on the approach used to quantify the degree of similarity between pairs of activity sequences are presented in Section 3.3.4.

Behavioural change. A behavioural change is identified when a new set of activity sequences emerges, evidenced by a lower degree of similarity to previously observed patterns compared to adjacent activity sequences (e.g., a user stops taking vitals before interacting with Alexa, which had previously been identified as a trend).

Novelty phase. This analysis focuses on discovering behavioural trends in households with PLWD beyond the novelty phase. In line with previous home-based robotics studies, the novelty period is defined as the first three months of use [237].

3.3.1 Research Questions

This study aimed to address the following research questions:

- **RQ1:** Do end-users engage with conversational technology beyond the novelty phase?

¹The UK DRI-CR&T's [Minder](#) digital care platform was developed to promote the independence of PLWD at home.

Table 3.1: Participant cohort and data acquisition. Total days refers to the duration of data collection, from the first to the last date of Alexa interactions, but participants did not use the device daily. The total number of Alexa triggers encompasses both the questionnaire trigger and random interactions throughout the data collection period. Unique events comprise all the behavioural events for each participant, including Alexa interactions.

PID	Sex	Diagnosis	Total days	Start date	End date	Total triggers	Unique events
P1	Male	Vascular dementia	169	2021-05-07	2021-10-23	202	100436
P2	Male	Dementia in Parkinson's	388	2021-05-13	2022-06-05	1824	239611
P3	Male	MCI	381	2021-05-16	2022-06-01	96	214879
P4	Male	AD	387	2021-05-14	2022-06-05	448	210271
P5	Male	AD	244	2021-09-28	2022-05-30	368	105326
P6	Female	AD	260	2021-09-08	2022-05-26	600	95762
P7	Female	AD	263	2021-09-08	2022-05-29	555	89475
P8	Male	AD	23	2021-09-08	2021-10-01	19	9716
P9	Female	AD	102	2021-09-08	2021-12-19	102	33569
P10	Female	AD	107	2021-09-20	2022-01-05	69	75
P11	Male	AD	73	2021-09-25	2021-12-07	72	15687
P12	Male	Lewy Body Dementia	220	2021-10-28	2022-06-05	2125	131032
P13	Male	AD	244	2021-10-05	2022-06-06	610	152162
P14	Male	AD	242	2021-10-05	2022-06-04	656	125112

- **RQ2:** Can the analysis of in-home activity and interactions with conversational technology identify trends and changes in behaviour?
- **RQ3:** Do participants continue using Alexa following clinical events (e.g., hospitalisations)?

3.3.2 Participants and Setting

A year-long longitudinal study was conducted in homes of PLWD to explore the integration of conversational AI in smart homes. Participants (N=14, 75-94 years, four females, P1 to P14) had a diagnosis of dementia or MCI and lived in their homes with a carer. Table 3.1 presents the cohort of participants and data collection information. The total data collection timeframe varied between participants due to two recruitment stages followed by Alexa deployment and individual usage.

3.3.3 Longitudinal Deployment

In-home activity and interaction data were collected using IoT technologies and commercial conversational technology (Amazon Alexa), respectively.

A. Technology Overview

Each household was equipped with the same monitoring technologies to collect in-home activity, namely: passive infrared (PIR) sensors installed in the bedroom, the lounge/living

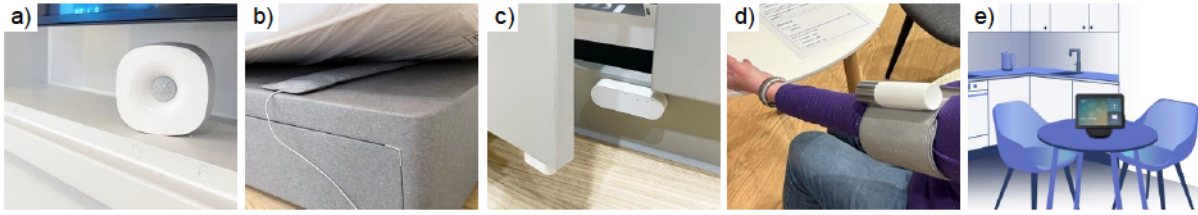


Figure 3.1: Technology used for in-home data collection. Home monitoring and conversational technology used to collect activity and interaction data, respectively: a) PIR, b) sleep mat, c) door sensor, d) blood pressure monitor, e) smart speaker.

room, the kitchen, the bathroom, and the hallway to capture motion; door sensors placed on the front door and the back door to detect when a door was opened or closed; a sleeping mat to collect information about a person getting in or out of bed; physiological devices to capture vital signs, including a pulse oximeter, scale, thermometer, and blood pressure cuff. In addition, Alexa was deployed in each home and daily interactions were analysed over time (see Figure 3.1).

B. Data Overview

A household analysis was conducted without using voice to ensure privacy and anonymity. Therefore, the interaction data comprised text utterances of each Alexa trigger (generated through embedded speech-to-text) and the corresponding timestamp. Two types of Alexa data were considered:

- *Questionnaire*: Participants were encouraged to complete a daily well-being questionnaire that included six well-being questions about mood, agitation, anxiety, sleep quality, tiredness, and activity plans.² This data type was used to inspect the frequency and time of questionnaire triggers over time. Timestamps for the start and end of the questionnaire were extracted and answers to each question were excluded.
- *Random interactions*: Participants were further encouraged to interact with Alexa freely (e.g., ask for the weather, news, or entertainment). This includes all Alexa triggers excluding the questionnaire start, end, and answers. This was used to explore common topics of interest.

Following a previous approach for in-home activity data collection [69], a total of 13 behavioural events were analysed related to location at home, getting into and out of bed, opening or closing doors, vitals³, and user-initiated interactions with Alexa (see Table 3.2). The activity data were extracted offline from DCARTE⁴, a framework that allows anonymised data access by DRI-CR&T researchers. The activity and interaction datasets were aggregated, grouped by participant, and sorted by timestamp, resulting in a total of over 1.5 million unique

²A bespoke Alexa skill prototype was developed for this study.

³Note that all physiological measurements were grouped into one vitals event.

⁴UK DRI-CR&T DCARTE (<https://zenodo.org/record/4592930>)

Table 3.2: Behavioural events captured by IoT technologies. Each behavioural event represents an activity captured by remote health monitoring devices or the Alexa smart speaker. Note that all physiological readings were grouped into a single vitals event and three Alexa types of interaction were considered.

Behavioural event	Monitoring devices	Activity detected
Lounge	PIR sensors	Motion
Kitchen	PIR sensors	Motion
Bedroom	PIR sensors	Motion
Bathroom	PIR sensors	Motion
Hallway	PIR sensors	Motion
Bed in	Sleep mat	Getting into bed
Bed out	Sleep mat	Getting out of bed
Front door	Door sensor	Door opening/closing
Back door	Door sensor	Door opening/closing
Vitals	Pulse oximeter, scale, thermometer, blood pressure cuff	Taking vitals
Start questionnaire	Amazon Alexa	Interacting with Alexa
End questionnaire	Amazon Alexa	Interacting with Alexa
Random interaction	Amazon Alexa	Interacting with Alexa

observations collected over 3103 person-days (see details in Table 3.1). Each observation represented one event with a timestamp.

As part of the recruitment and deployment approach followed in this study, a monitoring team maintained communication with participants to clarify technology capabilities, mitigate privacy concerns, and record health events impacting PLWD routine (e.g., falls, infections, hospitalisations). These notes were used to validate the quantitative findings. The proposed pipeline of analysis is illustrated in Figure 3.2.

3.3.4 Similarity of Activity Sequences

This study investigated behavioural patterns by analysing in-home activity data and voice interactions using conversational AI. The pairwise similarity of activity sequences occurring within the 10-minute period preceding or following Alexa triggers was calculated. Behavioural patterns were then explored by grouping activity sequences based on higher or lower similarity scores. An activity sequence, $\vec{s}$, is defined as a vector of temporally ordered sensor triggers (i.e., behavioural events):

$$\vec{s} = [s_1, \dots, s_n], s_i \in E,$$

where E is the finite set of 13 behavioural events considered in this study (see Table 3.2), and s_i is the event in position i of the sequence.

Each activity sequence includes temporal window and duration parameters, in addition to a defined target event, i.e., the type of interaction: *Start* of the well-being questionnaire; *End* of the questionnaire; and *Random* refers to other topics of interest.

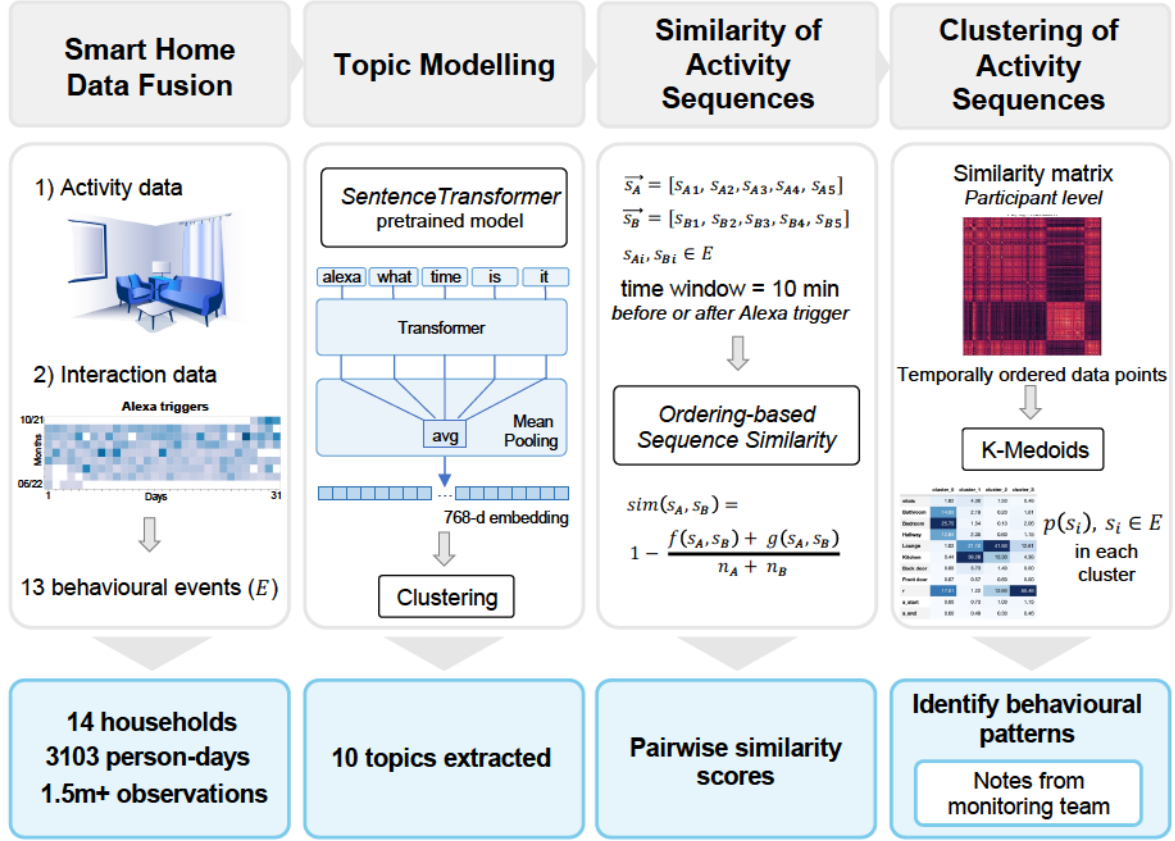


Figure 3.2: Illustration of the behaviour analysis pipeline. The main methods used in the study include: 1) the longitudinal fusion of environmental and voice data, comprising a set of 13 distinct behavioural events, as outlined in Table 3.2; 2) the topic modelling approach employing sentence embeddings and clustering techniques; 3) the analysis of activity sequences on a household level using a pairwise similarity measure; and 4) clustering to discern behavioural patterns. Additionally, notes from the monitoring team were used to explain and validate the quantitative findings.

Through exploratory data analysis, an optimal window of five consecutive events was selected and sequences were filtered to a maximum duration of 10 min near the target trigger. Note the duration of each event varies. Therefore, activity sequences may comprise repetitive events as long as the total vector size is five (i.e., $n = 5$) and the maximum duration is 10 min. Below are examples of equal-length sequences that preceded the trigger of the daily questionnaire ($s_n = \text{Start}$) for a given participant:

$$\begin{aligned} &Lounge > Kitchen > Lounge > Lounge > Start \\ &Vitals > Vitals > Kitchen > Lounge > Start \end{aligned}$$

To quantify the level of similarity between activity sequences composed of chronologically ordered behavioural events, $\vec{s}$, pairwise sequence similarity was calculated. Given the focus on capturing the temporal order of events, *ordering-based sequence similarity* [86] was applied. This categorical sequence mining technique considers the number of common elements and their order within the sequence.

Let $\vec{s}_A = [s_{A1}, \dots, s_{An}]$ and $\vec{s}_B = [s_{B1}, \dots, s_{Bn}]$ be two equal-length activity sequences. The

similarity score between s_A and s_B (the vector symbol was omitted for simplicity) was calculated as follows:

$$\text{sim}(s_A, s_B) = 1 - \frac{f(s_A, s_B) + g(s_A, s_B)}{n_A + n_B}, \quad (3.1)$$

where $f(s_A, s_B)$ quantifies the similarity in the position of elements in the sequence (i.e., the order), $g(s_A, s_B)$ counts the number of non-common elements, and n_A and n_B denote the vector size of s_A and s_B , respectively.

For a behavioural event $e \in E$ and activity sequences s_A and s_B , let L_A^e be the number of times e appears in s_A , and $s_A^e(k)$ the k^{th} position of e in s_A . C_{AB} denotes the set of common events in s_A and s_B . U_{AB} denotes the set of events that appear in s_A but not in s_B . Then, $f(s_A, s_B)$ and $g(s_A, s_B)$ are calculated as follows:

$$f(s_A, s_B) = \frac{\sum_{e \in C_{AB}} (\sum_{k=1}^{K_{AB}^e} |s_A^e(k) - s_B^e(k)|)}{\max(n_A, n_B)}, \quad (3.2)$$

and

$$g(s_A, s_B) = \sum_{e \in U_{AB}} L_A^e + \sum_{e \in U_{BA}} L_B^e, \quad (3.3)$$

where $K_{AB}^e = \min(L_A^e, L_B^e)$.

Using the two activity sequences shown above as an example with simplified notation $s_A = \{L, K, L, L, S\}$, $s_B = \{V, V, K, L, S\}$: since L , K and S appear in both sequences, $C_{AB} = \{L, K, S\}$. Looking at the position of common events in each sequence, $s_A^L = \{0, 2, 3\}$, $s_B^L = \{3\}$, $s_A^K = \{1\}$, $s_B^K = \{2\}$, $s_A^S = s_B^S = \{4\}$, therefore $f(s_A, s_B) = (|0-3| + |1-2| + |4-4|)/5 = 0.8$. Calculating the non-common events, V appears twice in s_B , hence $U_{BA} = \{V\}$ and $g(s_A, s_B) = 2$. Following Equation 3.1, $\text{sim}(s_A, s_B) = 1 - (0.8 + 2)/10 = 0.72$.

The target event, s_5 , was predefined to analyse activity sequences within the 10-minute window preceding Alexa use (e.g., for activity sequences preceding the questionnaire trigger, $s_5 = \text{Start}$). Similarity matrices were then analysed on a household level, using the computed pairwise similarity scores. Higher values within these matrices indicate a higher degree of similarity between pairs of activity sequences.

Subsequently, similarity matrices were used to cluster activity sequences by grouping those with higher or lower similarity scores. Given that the data points to cluster (i.e., the activity sequences) are not in a vector space, K-Medoids clustering was applied, a method based on the partition around medoids algorithm [112]. The silhouette method was used to determine the number of clusters. Altogether, the approach involved computing pairwise similarity and performing clustering of activity sequences to identify behavioural patterns and examine changes in behaviour using conversational technology at home.

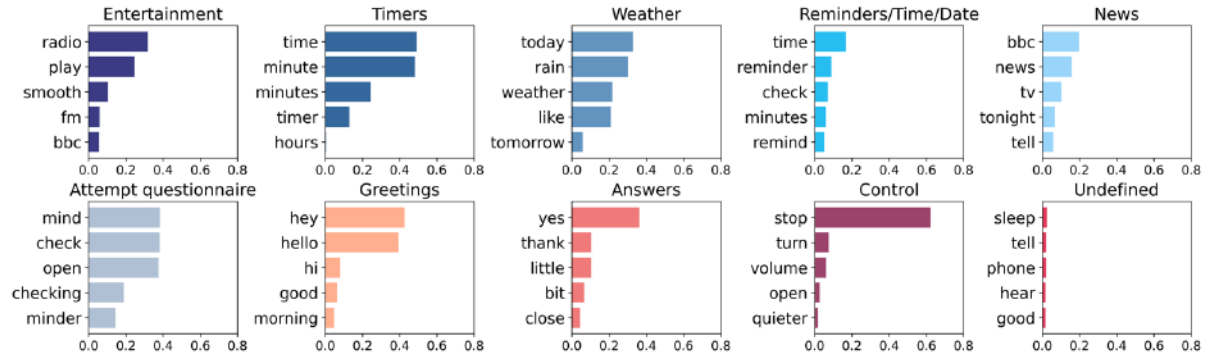


Figure 3.3: Topic modelling. Top words for each topic are shown, ranked by their TF-IDF scores reflecting word importance relative to all words within their respective topic. The *undefined* topic contained a diverse range of utterances, resulting in low TF-IDF scores.

3.3.5 Topic Modelling of Interaction Data

Topic modelling was applied to analyse Alexa interactions of type *random* (i.e., Alexa usage unrelated to the well-being questionnaire, as described in Section 3.3.3). A pre-trained language model was used for this analysis.⁵ K-Means clustering⁶ was applied in two iterations to the obtained utterance vectors, using the silhouette method to determine the optimal number of clusters.⁷ In the first round, K-Means clustered the vectors into 16 groups. Upon manual inspection, a set of generic utterances was labelled as the *undefined* topic. Similar clusters were then combined, resulting in eight distinct clusters, including the *undefined* topic. In the second round, the *undefined* topic was further clustered into 15 topics using K-Means. These topics were then merged with similar clusters from the first round, integrating two newly emerged clusters and identifying a new *undefined* topic.

Figure 3.3 shows the 10 identified topics from Alexa interactions and the most frequent words associated with each topic using term frequency-inverse document frequency (TF-IDF)⁸. These include 1) confirming Alexa actions or speech recognition, 2) control commands, 3) requesting entertainment such as music, radio, or games, 4) setting timers, 5) asking for weather updates, 6) attempting to start the daily questionnaire; 7) inquiring about time, dates, or setting reminders, 8) requesting news, 9) greeting Alexa, and 10) remaining Alexa interactions of type *random*.

⁵A pre-trained sentence transformer model [211] was used to embed each user utterance into a 768-dimensional vector ([all-mpnet-base-v2](#) from Hugging Face).

⁶Preliminary research on topic modelling revealed that non-negative matrix factorization and LDA produced unsatisfactory results.

⁷The performance of different clustering methods was evaluated using the silhouette score as a measure of coherence. K-Means marginally outperformed Hierarchical Clustering, Gaussian Mixture Models, and Spectral Clustering (see details in Appendix Table A2.1).

⁸TF-IDF is a statistical measure used to evaluate the importance of a word within a specific document relative to its occurrence across the entire corpus.

3.4 Results

This study analysed over 1.5 million events captured from IoT technologies, remote health monitoring devices, and conversational technology across a unique cohort of 14 households with PLWD. First, user engagement with Alexa was investigated. Four participants who interacted the most with Alexa were selected to report case studies (P2, P6, P12, and P14; see the total number of Alexa triggers in Table 3.1). The monthly average number of Alexa interactions (i.e., the total number of triggers) was examined both during and beyond the novelty phase on a household level. Furthermore, activity sequences were analysed to identify trends and changes in behaviour on a household level. Alexa usage was also inspected in the week following clinical events.

3.4.1 Long-Term Engagement at Home

This study explored the use of conversational technology by people affected by dementia and their carers at home over seven months on average⁹ (see Table 3.1). Particular interest was given to the *novelty effect* on user engagement (see RQ1 in Section 3.3.1), which has been argued as a key barrier to long-term adoption by older populations [118, 248]. The common topics triggered by participants when interacting with Alexa were also inspected.

A. Engagement Beyond the Novelty Phase

Figure 3.4 shows an overview of Alexa interactions over time for P12 and P14. P12 interacted with Alexa consistently over time, particularly during the novelty phase, as evidenced by a rise in *random* interactions. However, there was a gradual decline in Alexa usage during the post-novelty phase (i.e., after the first three months). P12's daily interactions with Alexa peaked in December 2021 (N triggers=42) during the novelty period. P14 also used Alexa consistently but exhibited intervals of non-interaction. P14's daily interactions peaked in November 2021 (N triggers=27), on the second day of use.

Engagement with the well-being questionnaire varied across participants. For instance, P14 consistently completed the questionnaire monthly, with a noticeable decline in engagement after seven months, in May 2022. In contrast, P12 gradually reduced the frequency of questionnaire triggers, eventually stopping after March 2022.

The weekly average number of Alexa interactions (of both questionnaire and *random* types) was compared during and beyond the novelty phase. This analysis included the 11 participants with a total Alexa usage of at least four months. A significant decrease in engagement was observed after the novelty phase across the cohort ($W=5$, $p<.025$, common-language effect size (CLES)=0.66¹⁰). A significant decrease in engagement with the daily well-being

⁹Note that technology usage varied across participants, averaging over seven months (222 days).

¹⁰A Wilcoxon signed-rank test was used and false discovery rate (FDR) was applied with corrected p-values compared against the significance level $\alpha = 0.025$. The CLES is also reported.

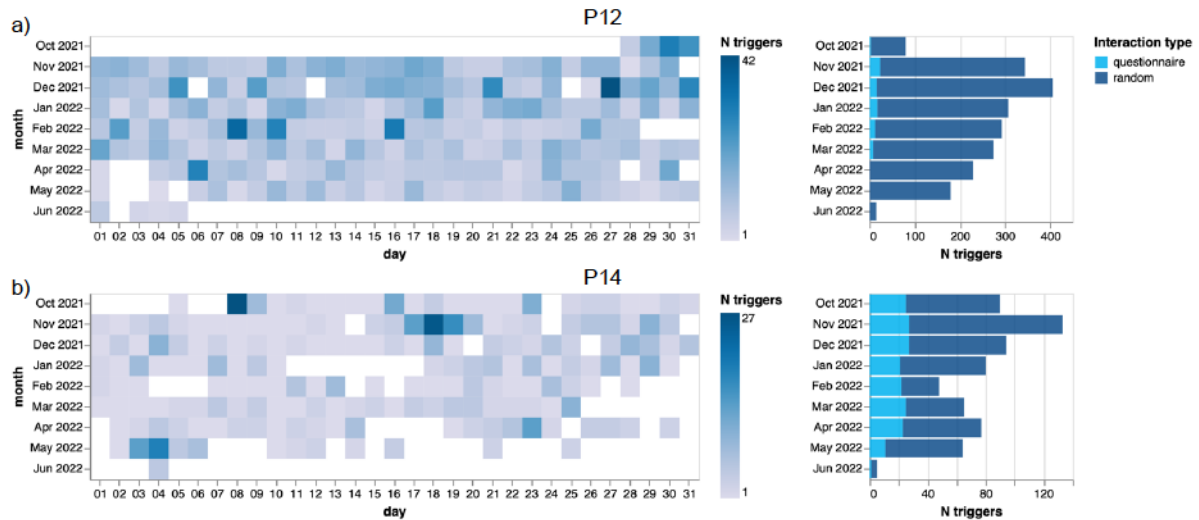


Figure 3.4: Longitudinal Alexa interaction data. The left plots show the daily counts of Alexa events per day. The right plots show monthly usage of interactions (both questionnaire and random) over the duration of Alexa use for: a) P12 and b) P14. N triggers quantifies the total number of interaction triggers.

questionnaire was also observed in the post-novelty period ($W=0$, $p<.025$, $CLES=0.72$). Moreover, of the participants who continued using Alexa beyond the novelty phase, two stopped completing the well-being questionnaire after the first month of use.

Figure 3.5 shows an overall decline in Alexa usage in the post-novelty period for participants with total Alexa usage exceeding eight months. In all these households, the weekly number of interactions peaked during the novelty phase. For example, P2's weekly Alexa interactions reached a peak in the first week of usage (13 May 2021, N triggers=302), as highlighted in the grey box in Figure 3.5. However, a marked decrease in usage was observed after the first month. Interestingly, interactions ceased for a 4-week period in February 2022. Health notes from the monitoring team revealed that P2 was hospitalised during this period, which explains the absence of Alexa usage. Furthermore, an increase in weekly interactions was observed in April 2022, coinciding with P2's return home on 31 March 2022 after hospitalisation (as logged by the monitoring team).

B. Frequent Topics of Interest

The common topics triggered by participants when interacting with Alexa (interactions of type *random*) were examined using a topic modelling approach with sentence embeddings [113] and K-Means clustering on the obtained utterance vectors. Figure 3.6a shows the top 10 most triggered topics across the cohort. Participants frequently used Alexa for control commands (e.g., change volume, start or stop actions), entertainment (e.g., play music, radio, or games), and trigger the well-being questionnaire, which participants were encouraged to complete daily. They also often used Alexa to set reminders or ask about current time, date, or day of the week, get weather information, set timers, and ask for the news. Some participants used

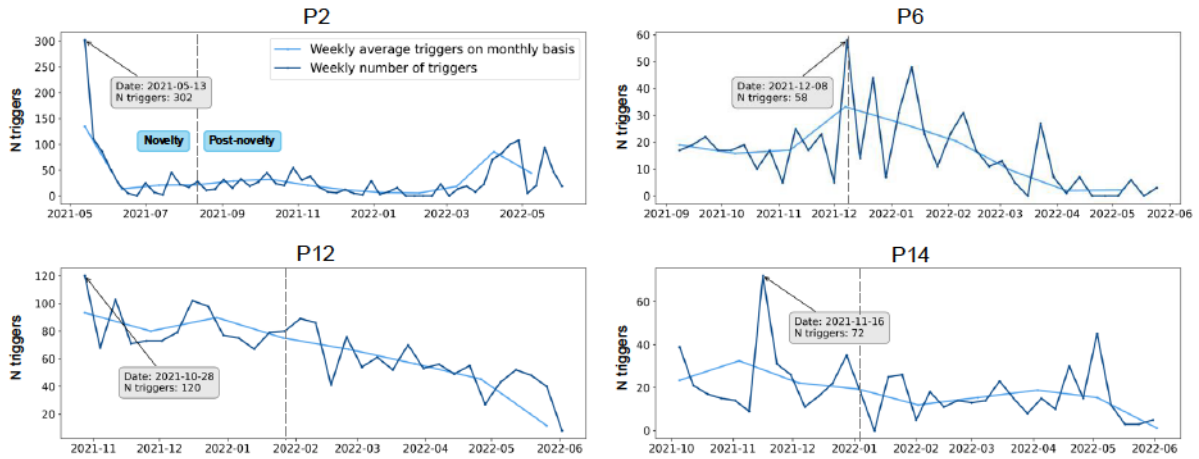


Figure 3.5: Engagement with Alexa during and after the novelty period. Weekly average number of Alexa interactions per month and total weekly interactions for P2, P6, P12, and P14. *N triggers* represents the total number of triggers. The vertical dashed line marks the transition from the novelty phase (first three months of usage) to the post-novelty period. Grey boxes indicate the date with the highest number of interactions and total triggers. Start dates and usage periods varied among participants (as detailed in Table 3.2), so usage data is standardized based on the first Alexa interaction for each participant.

the device to simply greet Alexa.

Figure 3.6b and Figure 3.6c illustrate the proportion of topics triggered for selected participants. After the novelty phase, P2 primarily used Alexa for morning weather updates (85.3% of weather requests occurred in the morning during this period, as detailed in Appendix Table A2.2) and setting timers in the morning (53.7%), especially after January 2022. Topics related to news, entertainment, and reminders were also frequently triggered over time. P6 showed an increased interest in using Alexa for entertainment in the morning, both during the novelty phase (55.6%) and afterwards (51.9%). This participant also frequently asked about the current date, time, and reminders in the morning during the first three months (66.7%), a trend that continued in the post-novelty phase (52.6%).

Overall, participants continued to find certain capabilities of Alexa engaging and used them regularly. While the observed decrease in Alexa usage, often attributed to unmet expectations or perceived lack of utility [49], aligns with previous research involving older demographics [118, 248], these findings suggest that future improvements in the design of home-based conversational agents are necessary. Specifically, addressing long-term user engagement and facilitating more personalised interactions could foster better integration in households with PLWD. Based on these findings, challenges and design opportunities for future conversational agents are discussed in Section 3.5.

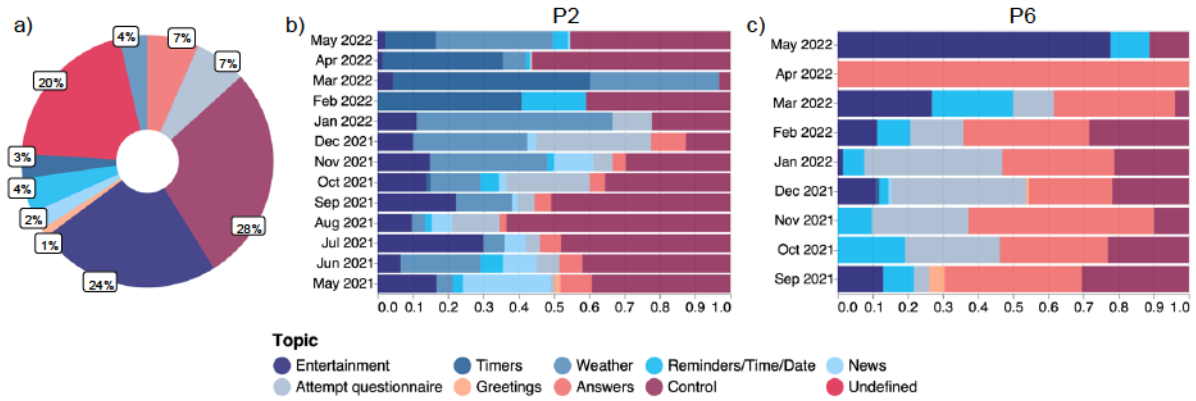


Figure 3.6: Topic distribution. a) Frequency of topics triggered across the cohort; b) Proportion of topics triggered monthly by P2 and c) P6, excluding the *undefined* topic.

3.4.2 Mapping Behaviour from Activity Sequences

This study aimed to investigate whether conversational technology and environmental sensors can map behaviour to support dementia care (see RQ2 in Section 3.3.1). In-home activities detected by IoT technologies were correlated with Alexa use (as outlined in Table 3.2).

In-home activity sequences in the 10-minute period preceding or following Alexa triggers were analysed. The probability of behavioural events within these sequences was first calculated, considering the two types of Alexa interactions, i.e., *questionnaire* and *random*. Figure 3.7 shows the probability of each behavioural event, captured by IoT technologies, for four representative users (P2, P6, P12, P14), within the 10-minute temporal window considered.

Participants were contacted to confirm the locations of their Alexa devices within their homes. By examining the probability of behavioural events in the activity sequences, the primary locations of Alexa were confirmed. P2 and P14 had the device placed in the lounge, as evidenced by higher probabilities of lounge events before and after Alexa triggers (see darker areas in Figure 3.7). Similarly, P6 and P12 had their device in the kitchen, supported by higher probabilities of kitchen events. The higher probability values for the corresponding behavioural events were consistent across the three types of activity sequences examined: (A) activities preceding the questionnaire trigger, (B) activities following the questionnaire completion, and (C) activities prior to random Alexa interactions.

In addition to identifying the primary Alexa locations, the frequent occurrence of specific events, such as kitchen activities for P2 or hallway for P12, indicates that these activities often occurred within the 10-minute period before or after Alexa triggers. For instance, P2's most common activity sequences before starting the well-being questionnaire ($s_5 = \text{Start}$, column A in Figure 3.7) was: *Kitchen > Lounge > Kitchen > Lounge > Start*. This sequence suggests that P2 typically completed the questionnaire in the lounge after moving from the kitchen.

Furthermore, the pairwise similarity for each pair of activity sequences was computed, as de-

	P2			P6			P12			P14		
	A(%)	B(%)	C(%)	A(%)	B(%)	C(%)	A(%)	B(%)	C(%)	A(%)	B(%)	C(%)
Bed in	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.00	0.00
Bed out	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00
Vitals	16.91	10.10	2.69	7.19	3.04	3.05	0.00	0.00	0.29	7.67	16.62	5.02
Bathroom	1.01	1.92	6.82	0.31	0.34	1.13	2.73	2.50	3.70	0.71	0.15	3.03
Bedroom	0.72	1.28	14.54	0.62	2.03	1.02	7.73	5.42	12.62	1.14	1.50	2.77
Hallway	0.87	3.37	6.00	19.06	21.62	17.65	11.82	27.92	23.73	0.00	0.00	0.00
Lounge	39.88	46.47	25.52	0.00	0.00	0.00	4.55	8.75	8.38	63.49	62.13	53.46
Kitchen	28.90	29.65	18.89	50.94	58.78	47.74	47.73	40.00	38.31	17.61	11.08	11.51
Back door	1.01	1.28	1.61	5.31	5.41	6.11	0.45	0.00	0.71	2.27	1.05	1.99
Front door	0.43	0.96	0.45	0.31	3.38	0.68	1.36	1.25	2.82	0.00	0.00	0.26
Random	8.24	3.04	22.01	15.62	4.73	12.10	22.27	12.50	6.74	5.97	6.44	12.63
Start	0.43	1.76	0.86	0.00	0.68	8.82	0.00	1.25	1.62	0.14	0.90	5.36
End	1.59	0.16	0.48	0.62	0.00	1.70	1.36	0.42	0.74	0.99	0.15	3.98

Figure 3.7: Probability of behavioural events before or after Alexa use. Probability (%) of each behavioural event occurring within the 10-minute period preceding or following Alexa use for P2, P6, P12, and P14. Three Alexa events were analysed, represented as columns: (A) activity sequences preceding the start of the well-being questionnaire $s_5 = \text{Start}$; (B) activity sequences following the completion of the well-being questionnaire $s_1 = \text{End}$; (C) activity sequences preceding any random Alexa interaction $s_5 = \text{Random}$. For each participant and activity sequence type the probability of an event $s_i \in E$ (the finite set of 13 events) was calculated by summing occurrences of event s_i in all sequences and dividing by the total number of possible occurrences. Note that the total number of occurrences excludes the first or last event of each sequence, s_1 or s_5 , since these were predefined and equal for all sequences analysed.

tailed in Section 3.3.4. This allowed us to form a distinct similarity matrix for each household. Each data point in the matrix corresponds to an activity sequence of five behavioural events that occurred in the 10-minute period preceding or following Alexa use (see Figure 3.8). Pairwise similarity scores were subsequently used to cluster the activity sequences, using the K-Medoids method, to identify groups of similar activity sequences and inspect differences between clusters. Given the chronological ordering of data points in each similarity matrix, it became feasible to identify common household behaviours during specific time periods, i.e., groups of activity sequences with a high degree of similarity for a given time period.

Figure 3.8a shows an example of P2's household behaviour over time, from May 2021 until June 2022. Border A shows a distinct shift in the similarity of activities within that period compared to prior months and subsequent weeks. Four distinct clusters were identified in P2's activity. One of these clusters (denoted as cl_1 for clarity) became more prominent during a 6-week period, labelled as X1 in Figure 3.8a. During X1, 84.7% of activity sequences belonged to the dominant cluster, cl_1 . Moreover, outside of period X1, only 11.9% of activity sequences belonged to cl_1 . Further details are presented in Appendix Table A2.3 and A2.4.

This indicates that during X1, there was a distinct pattern in P2's activity sequences. Notably, the probability of Bedroom events in activity sequences within cl_1 increased to 25.8%

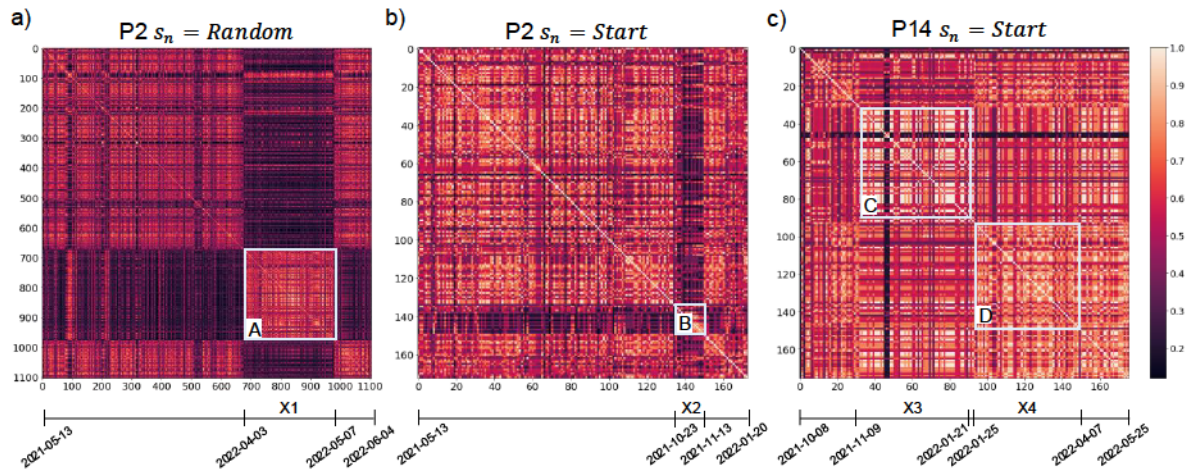


Figure 3.8: Similarity matrices of in-home activity sequences. a) similarity of activity sequences preceding Alexa interactions of type random ($s_5 = \text{Random}$) from May 2021 until June 2022 for P2; b) similarity of activity sequences that preceded the activation of the well-being questionnaire ($s_5 = \text{Start}$) from May 2021 until January 2022 for P2; c) similarity of activity sequences that preceded the activation of the well-being questionnaire ($s_5 = \text{Start}$) from October 2021 until May 2022 for P14. The axis numbers represent chronologically ordered activity sequences. The dates of the selected periods, along with the dates of the first and last analysed sequences, are shown below each matrix. Lighter tones represent a high degree of similarity and darker tones represent a low degree of similarity. The colour bar applies to all matrices.

compared to a maximum of 0.1% in other clusters (see Table 3.3). Additionally, the probability of Lounge events declined from a maximum of 41.5% in other clusters to 1% in cl_1 . Furthermore, during X1, 61.1% of household activity preceding Alexa interactions of type *random* occurred in the morning.

These findings suggest a change in the Alexa device location from the lounge to the bedroom and a trend toward triggering Alexa in the morning. This observation aligns with the higher probability of Bedroom events, as observed in Figure 3.7 (14.5%, column C). The monitoring team verified with P2 that the Alexa device had been moved to the bedroom during period X1 as a consequence of the participant's health condition, who, after a hospitalisation, primarily stayed in the bedroom for one month.

This approach and interpretation were subsequently applied to other similarity matrices for further analysis. Figure 3.8b shows the similarity matrix for P2's activity sequences preceding the questionnaire trigger, spanning from May 2021 to January 2022. The selected border, labelled B, was further analysed. Note that even though P2 discontinued the daily questionnaire in late January 2022, their Alexa usage for *random* interactions remained consistent until the end of data collection, as observed in the similarity matrix previously analysed (Figure 3.8a).

One of the three clusters identified using K-Medoids (denoted as cl_3) was predominant during a 4-week period, labelled X2. During X2, 77.8% of data points belonged to cluster cl_3 ,

Table 3.3: Case study 1 – probability of behavioural events for each cluster identified. The cluster size indicates the number of activity sequences within the cluster and the corresponding percentage of the total time period represented in the similarity matrix in Figure 3.8a. Note that bold numbers highlight key results reported in the text.

Behavioural event	cl_1 (%)	cl_2 (%)	cl_3 (%)	cl_4 (%)
Bed_in	0	0.2	0	0
Bed_out	0	0	0.1	0
Vitals	1.8	4.3	1.5	0.5
Bathroom	14.0	2.2	0.2	1.6
Bedroom	25.8	1.5	0.1	2.1
Hallway	12.8	2.4	0.6	1.2
Lounge	1.0	21.1	41.5	12.6
Kitchen	0.4	36.3	13.3	4.4
Back door	0	3.7	1.4	0
Front door	0.1	0.6	0.6	0
Random (Alexa)	17.5	1.2	12.6	50.5
Start (Alexa)	0	0.7	1.0	1.2
End (Alexa)	0	0.5	0.3	0.5
Cluster size	364 (33.0%)	313 (28.4%)	271 (24.6%)	156 (14.1%)

Table 3.4: Case study 2 – probability of behavioural events for each cluster identified. These values refer to the similarity matrix analysed in Figure 3.8b.

Behavioural event	cl_1 (%)	cl_2 (%)	cl_3 (%)
Bed_in	0	0	0
Bed_out	0	0	0
Vitals	5.5	5.3	47.9
Bathroom	1.4	0.7	1.1
Bedroom	0.5	0.4	1.6
Hallway	0.5	0.7	1.6
Lounge	61.8	36.3	19.7
Kitchen	13.2	49.3	16.5
Back door	1.4	1.1	0.5
Front door	0.5	0	1.1
Random (Alexa)	12.7	4.6	8.5
Start (Alexa)	0.5	0.4	0.5
End (Alexa)	2.3	1.4	1.1
Cluster size	55 (32.8%)	71 (41.0%)	47 (27.2%)

compared to just 21.3% outside of X2 (see Appendix Table A2.3 and A2.4). Furthermore, the analysis revealed that 94.7% of household activity preceding the questionnaire trigger during X2 occurred in the morning. The increased probability of Vitals events in activity sequences within cl_3 (47.9%, compared to a maximum of 5.5% in other clusters, presented in Table 3.4) suggests a trend of taking vitals before completing the daily questionnaire in the morning.

Similarly, Figure 3.8c illustrates P14's household behaviour over time. The selected borders, C and D, were further inspected. Four clusters were identified using K-Medoids, with two clusters (cl_1 and cl_4) being prominent across two distinct 11-week periods, labelled X3 and

Table 3.5: Case study 3 – probability of behavioural events for each cluster identified. These values refer to the similarity matrix analysed in Figure 3.8c.

Behavioural event	cl_1 (%)	cl_2 (%)	cl_3 (%)	cl_4 (%)
Bed_in	0	0	0	0
Bed_out	0	0	0	0
Vitals	2.1	46.7	9.1	0.4
Bathroom	1.1	0	0	0.7
Bedroom	2.1	0	4.6	0
Hallway	0	0	0	0
Lounge	50.0	43.5	27.3	89.1
Kitchen	38.7	1.1	4.6	3.9
Back door	2.5	1.1	6.8	1.8
Front door	0	0	0	0
Random (Alexa)	3.5	3.3	43.2	3.5
Start (Alexa)	0	0	2.3	0
End (Alexa)	0	4.4	2.3	0.7
Cluster size	71 (40.3%)	23 (13.1%)	11 (6.3%)	71 (40.3%)

X4. During X3, 64.8% of data points were associated with cluster cl_4 , while only 29.5% of activity sequences were associated with cl_4 outside of X3. During X4, 75% of data points were associated with cluster cl_1 , compared to only 24.2% outside of X4 (see Appendix Table A2.3 and A2.4). Furthermore, all activity sequences within both X3 and X4 were concentrated in the morning hours.

Period X3 suggests that P14 consistently completed the questionnaire while in the lounge (the probability of lounge events within cl_4 is 89.1%). On the other hand, period X4 indicates the predominance of activity in both the lounge (50% within cl_1) and kitchen (38.7% within cl_1), as seen in Table 3.5. Overall, these two periods suggest a trend in behaviour related to the use of conversational technology to complete the daily questionnaire. Specifically, P14 routinely triggered the questionnaire in the morning, either while in the lounge or following kitchen activities.

Furthermore, a noticeable darker area within period X3 on the matrix indicates a change in behaviour over nine days in early December 2021. During this period, P14's activity sequences consistently included the Vitals event, e.g., *Lounge > Lounge > Vitals > Vitals > Start*. This suggests that P14 routinely took vitals in the 10 minutes prior to completing the Alexa questionnaire, a pattern identified as highly dissimilar to surrounding activity sequences.

By fusing activity data captured by IoT technologies with interactions with conversational technology, the analysis demonstrated technical feasibility to map trends and changes in behaviour from in-home activity. This ability to map behaviour provides a basis for personalising future interactions. By tracking daily contexts at home, conversational AI could proactively initiate conversations about relevant topics at appropriate times. Furthermore, conversational AI could inform carers and clinicians of changes in behaviour detected, ultimately enhancing the well-being and safety of PLWD at home.

3.4.3 Use of Conversational Technology After Clinical Events

The study also investigated whether participants continued to use Alexa during the week following clinical events (see RQ3 in Section 3.3.1). Dates of individual health events (e.g., falls, infections, hospitalisations) were logged by the monitoring team in regular contact with participants. Alexa triggers (including both *questionnaire* and *random* interactions) were considered specifically within the seven days following each health event.

A total of 38 health events were evaluated across the cohort, which comprised events that took place on a day at home (e.g., falls), events of longer duration than a day during which PLWD stayed at home (e.g., Covid-19 infections), and events of longer duration than a day out of home (e.g., hospitalisations). Furthermore, the total number of events considered in this study corresponded to 10 participants (no health events were logged by the monitoring team for the remaining participants).

This analysis showed that participants continued interacting with Alexa even after exhibiting a clinical event ($W=561$, $p<0.017^{11}$). The same observation was verified when inspecting single dates and periods of health events during which PLWD stayed home ($W=435$, $p<0.017$). Interestingly, participants continued using Alexa in the 7-day period after returning from hospitalisations ($W=10$, $p<0.017$).

Despite the limited number of observations analysed across the cohort, these findings suggest the potential for conversational agents to provide personalised assessments of health and well-being after the occurrence of health events, which is further discussed in the next section.

¹¹A one-sample Wilcoxon signed rank test was conducted and FDR was applied, with corrected p-values compared against the significance level of $\alpha = 0.017$.

3.5 Discussion

This chapter presents a year-long longitudinal deployment of conversational technology in homes of PLWD (N=14), analysing user engagement and behaviour in smart environments to support dementia care. This section summarises the main findings of the study and their implications for future research towards real-world translation of conversational technology for utility in digital health monitoring.

A. Summary of Findings

A1. User engagement decreased after the novelty phase.

The analysis considered 13 behavioural events (outlined in Table 3.2) related to in-home activities and voice interactions with Alexa, both to complete a daily well-being questionnaire and trigger other topics of interest. User engagement was explored over time, particularly beyond the novelty phase (i.e., after the first three months). A significant decrease in engagement was observed after the novelty phase across the cohort, although certain topics prevailed in individual daily routines during the post-novelty phase (see RQ1 in Section 3.3.1). This decline can be attributed to a lack of perceived utility and interactions that adapt to individual needs, preferences, and cognitive state over time.

Carers frequently reported frustration with Amazon Alexa's invocation name, as PLWD often had difficulty remembering it. Speech recognition errors were also noted, with Alexa occasionally failing to understand user commands or handle longer pauses in speech when participants needed more time to formulate answers – a challenge reported in previous research on older adults using voice technology [118]. These technical barriers likely compounded user disengagement, particularly after the initial novelty wore off.

Another limitation identified that likely affected long-term engagement was the repetitive nature of the daily questionnaire participants completed during the study, which did not vary in content. The development of adaptive questionnaires that proactively engage with users for health and well-being self-assessments and follow up on previous responses represents a promising area of future research. User engagement with conversational technology will be further evaluated and discussed in this dissertation through user studies conducted in the UK, India, and the USA (see Chapter 5).

A2. In-home activity analysis showed the capacity to identify behavioural trends.

This study investigated the daily contexts in which Alexa was triggered at home by analysing activity sequences in the 10-minute period preceding or following its use (see details in Section 3.3.4). The behaviour analysis correlated in-home activities detected by IoT technologies and remote health monitoring devices with user-initiated Alexa interactions. This study demonstrated the technical feasibility to identify trends and changes in behaviour (see RQ2

in Section 3.3.1). Case studies for selected participants revealed distinct behavioural patterns, including relocating the Alexa device to the patient's bedroom as a response to health and caregiving needs after hospitalisation, a consistent morning routine of taking vitals before completing the well-being questionnaire, and a repetitive sequence of morning kitchen activities followed by triggering the questionnaire over several weeks.

A3. Alexa use continued after clinical events, suggesting opportunities for proactive support.

Participants continued using Alexa during the week following clinical events, including after returning home from hospitalisations (see RQ3 in Section 3.3.1). While the limited number of health events analysed across the cohort prevents broader conclusions with depth of clinical outcome, this preliminary finding suggests opportunities for automated verbal interactions targeting proactive and personalised health support after the occurrence of health events.

B. The Promise of Conversational AI in Digital Health Monitoring

This research argues that conversational AI integrated into the living environment holds potential to support ageing and dementia care. If well incorporated at home, particularly through integration with IoT technologies, conversational AI could map behavioural trends and provide automated verbal support and alerts in response to safety risks at home.

Conversational technology could directly query users for self-assessments of health and well-being, verify safety risks detected by other monitoring technologies, and encourage behaviours that promote well-being, such as suggesting activities or hydration at appropriate times. This provides a basis for the verification of changes in health and behaviour captured by IoT technologies in smart homes. Conversational AI could also offer verbal support in the event of perceived agitation or confusion and promptly notify relevant stakeholders, which can facilitate early intervention. Furthermore, verbal interactions could be personalised based on users' topics of interest, such as suggesting entertainment, reminding them of a favourite TV show, or sharing relevant news.

This research envisions conversational technology playing a crucial role in proactive and personalised dementia care with the potential to: 1) verify health or safety risks captured by IoT monitoring technologies in smart homes; 2) proactively query users for self-assessments of health and well-being following clinical events; and 3) trigger automated alerts to inform carers and healthcare professionals, facilitating timely interventions.

C. Limitations and Future Work

The small sample size ($N=14$) limits broader conclusions and generalisability of the findings, but the depth of longitudinal data collected is unique given time (>1 year), real-world setting (households of PLWD) and participant engagement.

The use of in-home monitoring technologies presents several challenges. In this study, participants had various monitoring technologies and an additional interactive device in their homes. Some participants expressed privacy concerns about having their voice interactions recorded and stored, which led to reduced adoption rates. These concerns must be addressed with stakeholder involvement early in the design, recruitment, and deployment process to ensure the technology infrastructure is non-invasive and privacy aware. This approach will be further explored in Chapter 5 through participatory studies with stakeholders aimed at better understanding their needs, experiences, perceived benefits and concerns.

Adapting interactions is essential for scalability in smart homes. This should incorporate previous user responses on subjective perceptions of health and well-being, individual cognitive abilities, and detected changes in behaviour at home. The proposed behaviour analysis could be integrated into proactive, personalised interactions, including the use of automated questionnaires, to offer insights into health and well-being, which are not captured by passive monitoring systems alone.

Future work could address the challenge of long-term engagement with conversational technology. Although participants in this study generally used Alexa for extended periods, usage declined over time, particularly after the novelty phase. This decline is attributed to a perceived lack of utility and insufficient adaptation of interactions. This research argues that to effectively support target populations at home, conversational technology should: 1) be easy to use; 2) be adequately integrated with the environment to proactively respond to contextual cues; 3) adapt to individual needs, preferences, and cognitive abilities over time to provide personalised support; 4) promote the autonomy of the carer, so they can undertake their caring tasks while receiving forms of relief; and 5) facilitate meaningful human connections, such as between PLWD, carers, and clinicians.

This study analysed behaviour at the household level, which could be extended to a more individualised approach. However, when multiple individuals, including visitors, interact with Alexa, the behaviour analysis may become skewed. Speaker diarisation techniques could help address this issue. Furthermore, future work could involve larger and more diverse populations to uncover real-world use cases that can be effectively implemented in smart homes, benefiting both users and clinical practices. This work identified the following use cases for future investigation: 1) tracking cognitive health from spoken language; 2) predicting comorbidities, such as the risk of agitation, from in-home activity and behaviour.

3.6 Summary

This chapter presented findings from a year-long deployment of commercial conversational technology (Amazon Alexa) in the homes of people affected by dementia (N=14). User engagement was evaluated over time, particularly beyond the novelty phase. In-home activity sequences were analysed using data captured from remote monitoring devices (e.g., motion, vitals, use of kitchen appliances) and conversational technology. To the best of knowledge, no study to date has integrated conversational AI with continuous home monitoring data, specifically targeting well-being support and safety in ageing and dementia care.

There was a significant decrease in engagement after the novelty phase across the cohort ($p < .025$), highlighting the need to adapt voice interactions to individual needs and preferences for personalised and proactive support in dementia care. The analysis demonstrated the technical feasibility of identifying behavioural trends by correlating activity data from remote monitoring technologies with interactions with conversational AI in smart homes. Additionally, there was engagement with conversational technology after clinical events, such as hospitalisations, suggesting future opportunities for adaptive questionnaires to directly query users for self-assessments of health and well-being.

The study highlighted the potential to integrate conversational technology at home for behaviour monitoring to support well-being and home safety. Key to realising this potential is the development of adaptive interactions that can: 1) provide proactive verbal support in response to safety risks detected by IoT and home monitoring technologies; 2) directly query PLWD to verify health symptoms; 3) trigger alerts to inform relevant stakeholders about risks at home, facilitating timely interventions; and 4) encourage positive behaviour through verbal prompts and suggestions tailored to individual needs.

Voice interactions with conversational technology could also offer valuable insights into subjective perceptions of health and well-being, user concerns, and cognitive status, which are overlooked by passive monitoring technologies alone.

The following chapter investigates the use of spoken language – which could be collected regularly using conversational AI at home – for automated assessment of cognitive health.

Predictive Modelling of Cognitive Health from Spoken Language

This chapter evaluates predictive models for automated assessment of cognitive health from spoken language, with a unique focus on clinical applicability. Using a benchmark DementiaBank speech dataset (N=291), interpretable ML techniques were evaluated for ADRD detection in binary classification and cognitive state prediction via Mini-Mental State Examination (MMSE) scores from spoken language during a picture description task. To validate model generalisability, a pilot study was conducted with speech data collected in-residence from older adults (N=22). By identifying the linguistic features most strongly predictive of cognitive decline and performing risk stratification, the interpretability and clinical utility of model predictions were enhanced. The proposed predictive modelling approach shows promise for real-world integration with in-home conversational technology to monitor and assess cognitive health, enabling earlier detection of higher-risk individuals and timely intervention.

The work presented in this chapter was submitted to Nature Communications Medicine (under review).

4.1 Overview

This study explored predictive models for automated assessment of cognitive health from speech and language with a focus on clinical applicability. The analysis included ADRD detection in binary classification and prediction of cognitive performance via MMSE scores to assess the severity of cognitive decline. The predictive modelling approach used spontaneous speech recordings from two DementiaBank datasets for training and testing (N=291) [157]. These were obtained from a picture description task during NPT.

To assess model generalisability in real-world settings, a separate pilot study is presented with speech data collected from older adults living in retirement homes (N=22). Acoustic and linguistic features were extracted using both conventional and deep learning approaches, including LLMs. For feature selection, interpretability was prioritised to inform clinicians of changes in language patterns that indicate cognitive decline, and accessibility for efficient scalable real-world application.

The final model incorporated 100 NLP-based lexical features. The linguistic features most predictive of ADRD risk were identified and thresholds for risk stratification were evaluated, aiming to optimise healthcare resource allocation by identifying higher-risk individuals. The contributions of this work are described below.

Contributions:

1. Assessment of the clinical applicability of cognitive health predictive modelling from spoken language and validation of model generalisability in a real-world context.
2. Insights that inform clinicians of linguistic features associated with higher ADRD risk.
3. Demonstrated potential for integration of the ML approach with in-home conversational technology for accessible, long-term monitoring of cognitive health.

Ethics Approval: The pilot study data collection was approved by the University of Southern California Institutional Review Board (IRB) UP-24-00154 (see Appendix A6).

Code Availability: The code used in this study is publicly available (<https://github.com/mariarlima>).

4.2 Background

4.2.1 Cognitive Health Screening

There is a pressing need for accurate, accessible, and cost-effective diagnostic methods for the early detection of cognitive decline. Dementia is typically diagnosed years after symptom onset, missing a crucial therapeutic window of intervention that is becoming increasingly important with the recent emergence of anti-amyloid drugs [173]. Traditional ADRD diagnostic methods rely on identifying fluid biomarkers such as *Tau* and β -amyloid related proteins, or neuroimaging techniques such as positron emission tomography (PET) and magnetic resonance imaging (MRI) [225]. While informative, these techniques are invasive, expensive, and inaccessible for scalable population screening [138]. Furthermore, brain imaging is useful only after signs of neurodegeneration manifest, further limiting the therapeutic window of opportunity. Administering NPT through an in-person interview remains the primary method to evaluate cognitive functions, including attention, memory, language, and visuospatial abilities. However, NPT are limited by clinician availability, are often qualitative in nature, and are susceptible to errors and high inter-rater variability. Furthermore, results can be affected by non-cognitive factors (such as mood disorders and fatigue) and expertise is required when interpreting the results to avoid false-positive diagnoses [67].

4.2.2 Voice Biomarkers

The use of *digital voice biomarkers* presents a scalable, non-invasive, ecologically valid and cost-effective solution for assessing and monitoring cognitive health. Changes in acoustic and linguistic characteristics have been linked to cognitive decline, including slower speech rate, more disfluencies (e.g., frequent pauses, hesitations, repetitions), reduced noun use, and increased use of pronouns, verbs, and adjectives [78, 105, 151, 247]. Neuroimaging studies also indicated that semantic fluency and naming performance correlate with neurodegeneration in the temporal and parietal lobes [258], areas commonly affected in AD.

The Alzheimer's Dementia Recognition through Spontaneous Speech only (ADReSSo) challenge has been proposed for systematic comparison of ML approaches for AD detection and severity prediction using spontaneous speech data from the *Cookie Theft* picture description task from the Boston Diagnostic Aphasia Examination (BDAE) [88]. The challenge baseline [157] achieved an accuracy of 64.8% and 77.5% on the test set using a Support Vector Machine (SVM) classifier on acoustic and linguistic features, respectively. Late fusion improved the accuracy to 78.9%. Chen et al. [45] achieved an accuracy of 81.7% using a logistic regression model and modality fusion. Qiao et al. [201] reported 83.1% accuracy using BERT, ensemble techniques, and additional disfluency features. More recently, Agbavor et al. [3] applied GPT embeddings and reported an accuracy of 80.3% using an SVM model, and showed that the embeddings outperformed conventional acoustic features.

4.3 Methods

This section describes the speech datasets used for model training and evaluation, the ML pipeline, methods for linguistic feature extraction, and the risk stratification approach to further assist clinical decision-making. The following ADRD severity groups were considered based on MMSE scores according to the UK National Institute for Health and Care Excellence dementia guidelines [250]: cognitively normal (CN) (26, 30], MCI (20, 26], moderate [10, 20], severe [0, 10) (following interval notation).

4.3.1 Research Questions

This research aimed to answer the following research questions:

- **RQ1:** How effective are ML models in detecting ADRD and predicting individual MMSE scores on a benchmark DementiaBank dataset?
- **RQ2:** Does the ML approach generalise to real-world data collected in-residence?
- **RQ3:** How does risk stratification change model performance?
- **RQ4:** Which spoken language features are most strongly predictive of cognitive decline?

4.3.2 Speech Datasets

This research evaluated the effectiveness of various ML models using speech recordings from DementiaBank (N=291) and validated model generalisability on an independently collected speech dataset (N=22). Table 4.1 describes the demographic characteristics of each study cohort used for training, testing and real-world pilot testing.

A. DementiaBank Data

The ADReSSo dataset from DementiaBank [157] was used to train and evaluate the ML models. This dataset was selected because it includes spontaneous speech recordings produced by CN participants and people with an ADRD diagnosis, who were asked to describe the *Cookie Theft* picture, a standard task of the BDAE [88] (see Appendix A4). The recordings were acoustically pre-processed with noise reduction and volume normalisation. This dataset contained 237 speech samples (5 hours) with a 70:30 train-test ratio balanced for demographics (see Table 4.2).

To verify the generalisability of the best-performing models for ADRD detection and MMSE prediction, external validation was performed with two datasets beyond the ADReSSo held-out set. The Lu corpus from DementiaBank was used as an *external test set* [127]. This dataset comprised 54 speech samples (1 hour) from the same picture description task with binary labels for CN participants and those with a ADRD diagnosis (see Table 4.2). Note

Table 4.1: Demographic characteristics of the study cohorts. Mean (standard deviation (SD)) is reported for age and MMSE.

	Training	Test	Additional Test	Pilot Study
Total	166	71	54	22
Age (years)	68 (6.8)	67.3 (6.9)	79.3 (9.7)	76.2 (8.0)
Sex (% male)	34%	38%	41%	41%
MMSE	22.9 (7.0)	23.9 (6.6)	–	24.9 (3.9)
Cognitive group	79 CN, 87 AD	35 CN, 36 AD	27 CN, 27 ADRD	8 CN, 14 MCI
Language	All English	All English	All English	14 English, 8 Spanish

Table 4.2: DementiaBank datasets. Characteristics of participants in the ADReSSo training and test datasets, as well as the Lu corpus. Note that the latter does not provide individual MMSE scores.

	ADReSSo train (N=166)		ADReSSo test (N=71)		Lu test (N=54)	
Cognitive Group	AD	CN	AD	CN	ADRD	CN
Participants	87	79	36	35	27	27
Sex (% male)	33%	34%	39%	37%	48%	33%
Age	69.7 (6.8)	66 (6.3)	68.5 (7.1)	66.1 (6.5)	79.0 (9)	79.7 (10.6)
MMSE	17.4 (5.3)	29 (1.2)	18.9 (5.8)	28.9 (1.3)	–	–

that the additional DementiaBank test set does not provide MMSE scores, so this dataset was not used for severity prediction modelling.

B. Pilot Study Data

Separately, an additional speech dataset was collected from 22 older adults (46 min) living in retirement homes, who completed the same verbal picture description task. The dataset included both English and Spanish speakers. Although participants did not have a dementia diagnosis, they were grouped into two cognitive groups: CN and those with mild to moderate cognitive impairment, using standard MMSE cutoff of 26 as suggested in previous work [222]. Participants reported their races and ethnicities: 14 African-American, 7 white, 8 Hispanic Latino; one participant did not disclose. This newly collected data in a real-world setting is referred to as *pilot study*. The pilot study was approved by the University of Southern California IRB UP-24-00154.

4.3.3 Acoustic and Linguistic Features

Acoustic and linguistic features were extracted using both NLP-based methods for interpretable features and pre-trained deep learning models. Additionally, six multimodal feature combinations were explored using early fusion methods [21]. The feature sets extracted were used as input to various ML models, as illustrated in Figure 4.1. The ML pipeline for feature extraction and predictive modelling is illustrated in Figure 4.2.

A. Acoustic Features

The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features were ex-

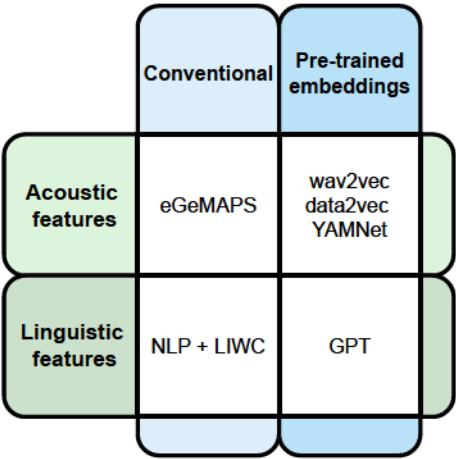


Figure 4.1: Feature extraction from speech and language. Acoustic and linguistic features were extracted using both conventional and pre-trained deep learning models.

tracted directly from open-source Speech and Music Interpretation by Large-space Extraction (OpenSMILE) [71] with proven usefulness for paralinguistic acoustic feature extraction [70]. It consists of 88 features per speech sample, including frequency, spectral, and energy-related parameters that capture various aspects of voice quality, prosody, and speech dynamics. Additionally, three deep neural embeddings were used to extract feature representations from audio data: wav2vec [19] using a vector size of 768, data2vec [18] using a vector size of 768, and YAMNet¹, which predicts audio events from 521 classes, using a vector size of 1024.

B. Linguistic Features

Transcripts were obtained from each audio file (i.e., one per participant) using OpenAI's Whisper [202] for automatic speech recognition (ASR). Given the high Spearman's rank correlation (mean $r=0.98$, $SD=0.03$, $p<.05$) between linguistic features manually transcribed from participant-only and combined speaker data across the 237 DementiaBank audio files, the remainder of the analysis proceeded without speaker diarisation. From the transcripts obtained with ASR, linguistic features were extracted using two methods:

- Token embeddings, using transformer-based pre-trained language models to create a 1536-dimensional vector representation for each participant's transcript;²
- NLP to extract lexical features based on domain knowledge.

The latter was used to train ML models with clinical applicability by providing interpretable insights into the linguistic patterns that contribute to model predictions. This transparency facilitates more informed decision-making by clinicians and increases confidence in the results.

¹The YAMNet model.
²We used OpenAI's GPT embeddings (text-embedding-3-small model).

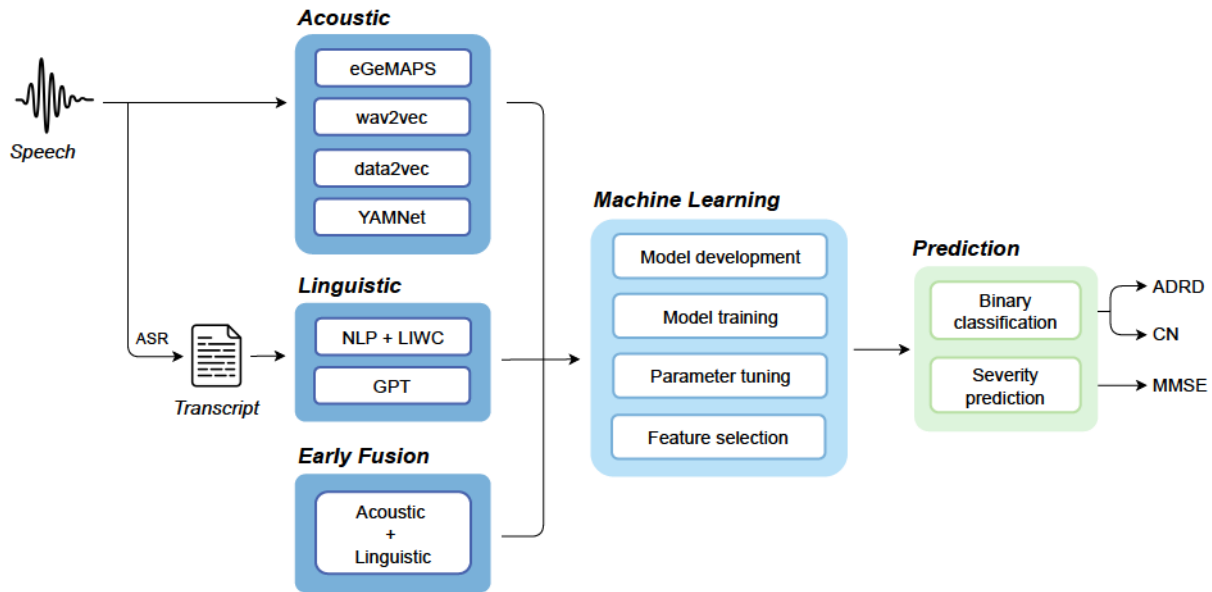


Figure 4.2: Proposed ML pipeline for cognitive health assessment. Analysis used for screening of cognitive health and MMSE prediction from spoken language.

Five lexical diversity features were computed, including type-token ratio corrected for text length, Brunet's index, Honore's index, propositional idea density, and consecutive duplicate words (see details in Table 4.3). These were combined with semantic psycholinguistics features extracted using Linguistic Inquiry and Word Count (LIWC)³, a method that counts words in psychologically meaningful categories [193]. Previous studies using LIWC demonstrated its ability to distinguish MCI and ADRD [16, 281], and loneliness among older adults [264].

After pre-processing and selection of LIWC subcategories, a 100-dimensional vector was extracted from each participant's transcript (see Table 4.3). To maintain consistency in linguistic feature extraction, GPT-4o translation was applied to Spanish transcripts before extracting English-based LIWC features, ensuring an end-to-end pipeline from data collection through pre-processing to analysis.

4.3.4 Predictive Modelling

A. Machine Learning Models

The analysis involved an ADRD detection binary classification task and an MMSE severity score regression task. For the first, the following models were tested for detecting ADRD from spontaneous speech data: Logistic Regression (LR), SVM, Random Forest (RF), Multilayer Perceptron (MLP), and Extreme Gradient Boosting Decision Tree (XGBoost). For the second task, the following models were tested in predicting MMSE severity scores: Ridge Regression

³The LIWC-22 English-only dictionary was used, as it includes a more comprehensive and diverse vocabulary compared to older multilingual versions.

Table 4.3: NLP-based linguistic features extracted for predictive modelling. Description of the lexical diversity and semantic psycholinguistics features extracted, along with references to studies that have explored these features for MCI and ADRD detection.

Features	Description
<i>Lexical diversity (N=5)</i>	
Type-token ratio	The ratio of unique words (types) to total words (tokens) in a transcript, adjusted for text length. Lower values indicate less diverse vocabulary usage and less lexical richness [32, 34].
Brunet's index	A measure of lexical richness based on the variation in word types (part-of-speech) relative to the total word count. Lower values indicate higher lexical richness [32, 34, 181, 281].
Honore's index	A measure of lexical richness that establishes the number of words used only once by the participant as a proportion of the total number of words used. This statistic is based on the notion that the larger the number of words used by a speaker that occur only once, known as hapax legomena, the richer the lexicon. Higher Honore's index values reflect richer vocabulary used by the participant [32, 34, 78, 181, 281].
Propositional idea density	The number of distinct propositions (facts or notions) divided by the total word count, measuring semantic complexity. In this metric, propositions correspond to verbs, adjectives, adverbs, prepositions, and conjunctions. Nouns are not considered propositions, as the main verb and all its arguments count as one proposition. Lower values suggest simpler language, while higher values reflect a greater number of ideas in a concise form [68, 214].
Consecutive duplicate words	The proportion of duplicated words or terms with reference to the total number of words as a measure of repetitiveness [281].
<i>Semantic psycholinguistics (N=95)</i>	
LIWC	Different categories are extracted using the LIWC-22 Dictionary, including: word count, summary language variables (e.g., analytical thinking, clout, authenticity, and emotional tone), general descriptor categories (words per sentence, percent of target words captured by the dictionary, and percent of words in the text that are longer than six letters), standard linguistic dimensions (e.g., percentage of words in the text that are pronouns, articles, adverbs, verbs), word categories tapping psychological constructs (e.g., affect, cognition, biological processes, drives) personal concern categories (e.g., home, leisure activities), informal language markers (assents, fillers) [16, 281].

(RR), Support Vector Regression (SVR), Random Forest Regressor (RFR), MLP Regressor, and XGBoost Regressor.

Hyperparameters were tuned using 10-fold cross-validation (CV), as detailed in Table 4.4. Furthermore, the best model was evaluated on the ADReSSo held-out test set that was not used during model development, as well as the real-world pilot dataset. Bootstrapping was used to estimate performance variability on the test set with 10 bootstrap repeats, with each run using a bootstrap sample of the training set to ensure reproducibility.

B. Evaluation Metrics

Classification performance was measured using sensitivity, specificity, receiver operating characteristic area under the curve (ROC-AUC) and accuracy, as described below. Each metric provides important information about the model's performance, and their combined use helps provide a comprehensive picture of the model's predictive ability. Understanding these met-

Table 4.4: ML models and parameters evaluated. Models used in ADRD binary classification and MMSE prediction tasks. Hyperparameters were tuned using 10-fold CV.

Model	Parameters
<i>Classification task</i>	
LR	L1 (Lasso) or L2 (Ridge) regularisation: $[10^{-5}, 10^2]$, solver: 'liblinear' (more efficient for small datasets) or 'saga' (supports both penalties).
SVM	Regularisation: $[10^{-4}, 10^3]$, gamma: 'scale', 'auto', or $[10^{-6}, 1]$, kernel: 'linear' or 'rbf'.
RF	Gini entropy, estimators: $[50, 500]$, max depth: $[3, 20]$.
MLP	Learning rate: $[0.001, 0.01]$, activation: 'logistic', batch sizes: $[16, 32, 64, 128, 166]$ (166 is the total number of recordings available for training), hidden layer size: 400, L2 regularisation: $[10^{-4}, 10^{-3}]$, stochastic gradient descent with adaptive learning rate.
XGBoost	Learning rate: $[0.01, 0.5]$, estimators: $[50, 500]$, max depth: $[1, 10]$, subsample ratio: $[0.01, 0.99]$, L1 regularisation: $[0, 0.001]$.
<i>Regression task</i>	
RR	L2 regularisation: $[10^{-3}, 10]$.
SVR	Regularisation: $[10^{-2}, 10^2]$, gamma: 'scale' or 'auto', kernel: 'linear' or 'rbf'.
RFR	Estimators: $[50, 200]$, max depth: $[5, 10]$, min samples split: $[2, 5]$, min samples per leaf: $[1, 2]$.
MLP Regressor	Learning rate: $[10^{-3}, 10^{-1}]$, batch sizes: $[16, 32]$, hidden layer size: optimised, L2 regularisation: $[10^{-3}, 10^{-2}]$.
XGBoost Regressor	Learning rate: $[0.01, 0.3]$, estimators: $[50, 200]$, max depth: $[2, 6]$, subsample ratio: 0.5, column sample ratio: 0.5, L1 regularisation: $[0, 1]$, gamma: $[0, 0.4]$.

rics can help healthcare providers to assess the reliability and usefulness of these models in clinical practice.

ROC-AUC: summarises the model's performance across all classification thresholds by plotting the true positive rate against the false positive rate. ROC-AUC was selected as the primary evaluation metric as it is based on the predicted probability scores, providing a comprehensive assessment of the model's ability to distinguish true ADRD cases while minimising false positives across all classification thresholds.

Sensitivity (and equivalently, recall): the proportion of true positive predictions among all actual positive cases. It measures the model's ability to correctly identify individuals who are at risk of ADRD. A high sensitivity indicates that the model is effective at identifying true positive cases.

$$\text{Sensitivity} = \frac{TP}{TP + FN},$$

where TP, TN, FP, and FN refer to True Positives, True Negatives, False Positives, and False Negatives, respectively.

Specificity: measures the model's ability to correctly identify individuals who are not at risk of ADRD. A high specificity indicates that the model is effective at identifying true negatives, i.e., those who are not at risk.

$$\text{Specificity} = \frac{TN}{FP + TN}$$

Accuracy: The proportion of correct predictions among all cases, both true positives and true negatives.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Regression performance (MMSE prediction) was measured with mean absolute error (MAE) and root mean square error (RMSE), as described below.

MAE: measures the average magnitude of the errors in the predictions. It is the average absolute difference between the predicted and actual values. A lower MAE indicates better predictive performance.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|,$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of samples.

RMSE: the square root of the average of the squared differences between the predicted and actual values. RMSE gives higher weight to larger errors, making it more sensitive to outliers than MAE. A lower RMSE indicates a better fit to the data.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

C. Model Calibration and Reliability

The chosen RF-NLP model was verified to be well-calibrated (see Figure 4.3a). The performance of the model using the best hyperparameters is reported by averaging the selected evaluation metrics across all validation folds.

There was some variability in the calibration across the bootstrap runs, but the general trend remained close to the diagonal, indicating that the model is fairly robust. The reliability analysis showed that the risk scores can be considered reliable when considering both positive and negative ADRD predictions together. This can be seen by observing the small gaps in the top plot of Figure 4.3b and the distance between the average accuracy and average confidence in the lower plot.

D. Model and Feature Selection

Each classifier was evaluated on the selected feature sets (acoustic, linguistic, and fusion, as described in Section 4.3.3) and the model with the highest ROC-AUC on the validation set was selected for further comparison. From this process, two models trained on linguistic features were selected. From these two best-performing models, the one using lexical-based NLP features was selected for final analysis due to its explainability and interpretability of features for clinical utility.

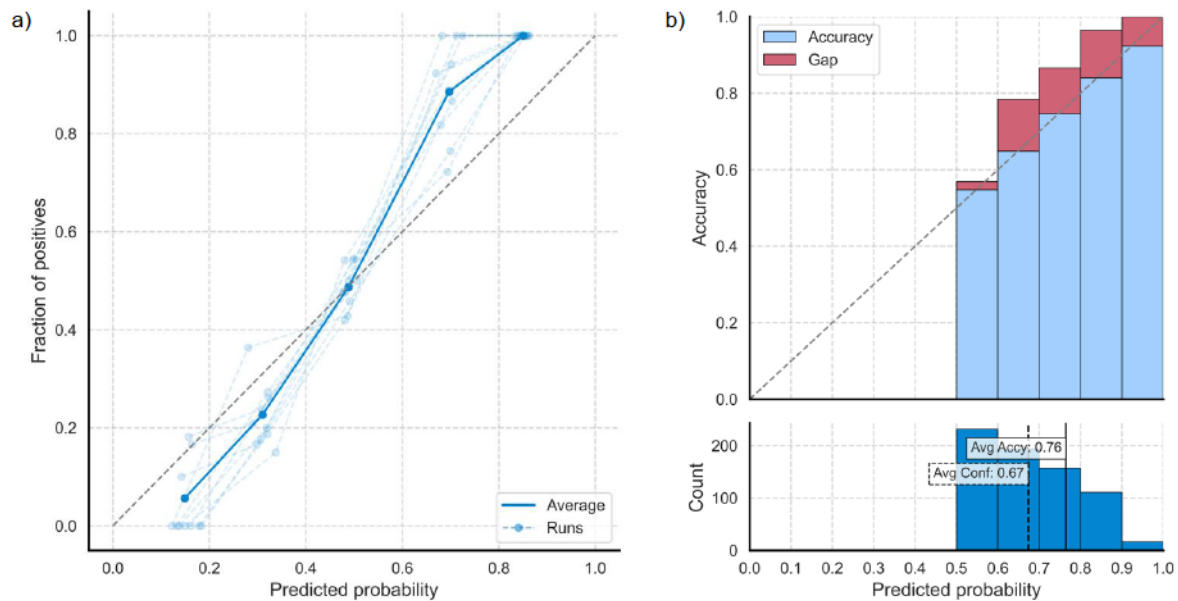


Figure 4.3: Model calibration and reliability plots. a) The calibration plot shows the mean ADRD predicted risk against the proportion of positive cases for the best model, evaluated on the test set. The 10 bootstrap runs are represented in lighter colour; b) The reliability plot. The top shows the model confidence (for positive and negative ADRD cases) against accuracy on the test set. The gap represents the difference between average accuracy and confidence per bin, with an ideal gap of 0. The bottom shows a histogram of model confidence levels on the test set.

4.3.5 Risk Scores

To support clinical decision-making, risk scores from the best-performing model were stratified into three ADRD risk groups: Green (low risk), Amber (medium risk), and Red (high risk). The thresholds were determined via 10-fold stratified cross-validation on the validation set. By varying the thresholds, the sensitivity and specificity for the different risk groups can be balanced. The thresholds were varied with a resolution of 10%, and evaluated performance metrics for the Green (positive prediction) and Red (negative prediction) groups on the validation set. Following a selective classification approach [76], the Amber group was excluded, assuming higher model uncertainty.

This approach aims to enhance clinical utility by prioritising more confident predictions in the Green and Red groups, which can streamline triage processes and better inform clinical decisions by identifying individuals at higher risk of ADRD. Given the small size of the dataset, when similar results were obtained for different thresholds on the validation set, smaller thresholds were selected to prevent overfitting.

4.4 Results

4.4.1 Model Performance in ADRD Detection

The analysis presents the effectiveness of LR, SVM, RF, MLP, and XGBoost in detecting ADRD from the extracted acoustic and linguistic features. Models trained with linguistic features achieved higher performance than those using acoustic features for ADRD classification, as described in Table 4.5. The best-performing model was MLP using GPT embeddings (referred to as MLP-GPT). However, the RF model of 50 decision trees with depths of 16 (chosen through hyperparameter optimisation) using lexical-based features from NLP (referred to as RF-NLP) achieved a comparative mean ROC-AUC on the validation set with only 4% lower performance than MLP-GPT, whilst offering more interpretability and efficiency.

Furthermore, both models showed comparable performance in identifying true positive and true negative cases on the benchmark DementiaBank test set. While the RF-NLP model exhibited a higher number of false negatives compared to the MLP-GPT model (107 vs. 65 cases, as seen in Figure 4.4), its enhanced interpretability remains a key advantage. This is partly explained by the RF-NLP model using a more compact (size 100) and interpretable linguistic feature set compared to the GPT embeddings (see Section 4.3.3). Additionally, this model does not require the use of pre-trained transformer-based models, which lack explainability. These more explainable features could better inform clinicians of linguistic changes indicative of cognitive deterioration. Therefore, considering the study focus on clinical utility of digital cognitive health screening, the proceeding results are presented for the RF-NLP model.

On the test set, the model achieved a ROC-AUC, the primary evaluation metric, of 85.7% (95% confidence interval (CI)=83.8–87.6). This model performance is comparable to previous studies using the same dataset [3, 157, 201] and outperforms previous results based on informative linguistic features [45]. Furthermore, while these previous studies reported performance from a single run on the test set, the experiments in this analysis were performed using 10 bootstrap repeats to ensure superior reproducibility. Previous studies focused on accuracy as the main performance metric, whereas this study prioritised ROC-AUC since it is based on predicted probability scores (instead of discrete class labels) and measures the ability to classify true ADRD cases while minimising false positives.

A. Model Generalisation

To further assess model generalisation, the RF-NLP model was evaluated on an additional DementiaBank dataset [127] never seen during model training and used only for a final test. Results suggested good model generalisation on unseen data using the NLP features, with a comparably high ROC-AUC of 84.6% (95% CI=82.8–86.4). Table 4.6 presents the performance results on the validation and test data sets using NLP-based features.

Table 4.5: Results of the feature sets tested for ADRD detection in binary classification. Evaluation metrics include sensitivity, specificity, ROC-AUC and accuracy, reported as mean (SD)% for the 10-fold CV. The bold row shows the chosen model for interpretable predictive modelling.

Feature	Model	Sensitivity	Specificity	ROC-AUC	Accuracy
<i>Linguistic</i>					
NLP	RF	78.8 (16.7)	72.1 (13.4)	83.5 (8.9)	75.3 (9.4)
GPT	MLP	79.3 (13.5)	76.2 (18.1)	87.5 (7.7)	77.7 (12.3)
<i>Acoustic</i>					
eGeMAPS	LR	69.3 (11.7)	70.5 (20.3)	73.1 (12.3)	69.8 (8.3)
wav2vec	LR	78.1 (9.8)	67 (16.3)	78.6 (11.1)	72.9 (10.2)
data2vec	MLP	75.3 (14.5)	73 (16.5)	81.9 (13.0)	74.4 (12.3)
YAMNet	SVM	72.8 (14.3)	69.6 (15.9)	77.5 (12.4)	71.1 (11.2)
<i>Fusion</i>					
eGeMAPS + NLP	LR	72.4 (10.6)	76.2 (18.1)	77.4 (10.4)	74.1 (9.1)
eGeMAPS + GPT	LR	76.0 (12.0)	86.2 (13.1)	87.1 (8.3)	80.7 (10.7)
data2vec + NLP	XGBoost	83.8 (7.9)	78.6 (11.1)	87.4 (10.3)	81.3 (6.9)

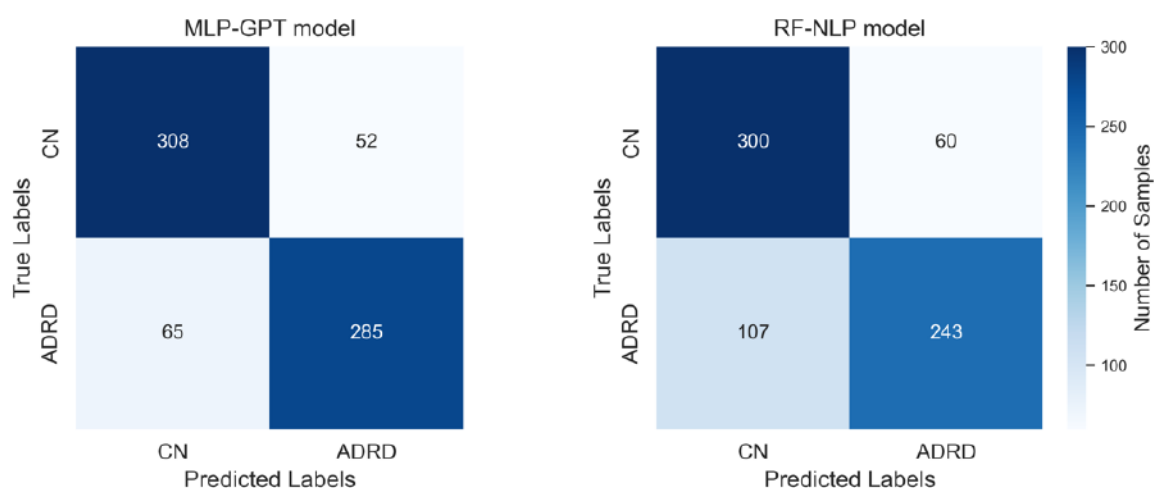


Figure 4.4: Confusion matrices for the best-performing models in ADRD detection. The total number of true positives (ADRD) and true negatives (CN) are shown. The RF-NLP model, while slightly less accurate, offers greater interpretability of linguistic features, making it more suitable for clinical settings.

Table 4.6: Model performance in binary classification using interpretable linguistic features. Classification results of the best-performing model RF-NLP using interpretable linguistic features. Evaluation metrics include sensitivity, specificity, ROC-AUC and accuracy, reported as mean (SD)% for the 10-fold CV, and as mean (95% CI)% for the test sets (benchmark DementiaBank, external test set from DementiaBank, and the pilot dataset) with 10 bootstrap repeats.

	Sensitivity	Specificity	ROC-AUC	Accuracy
Validation	78.8 (16.7)	72.1 (13.4)	83.5 (8.9)	75.3 (9.4)
Test	69.4 (66.4 – 72.5)	83.3 (78.0 – 88.7)	85.7 (83.8 – 87.6)	76.5 (74.4 – 78.6)
External Test	80 (77.2 – 82.8)	74.1 (69.6 – 78.6)	84.6 (82.8 – 86.4)	77 (74.6 – 79.5)
Pilot Study	70.0 (58.0 – 82.0)	52.5 (39.3 – 65.7)	65.4 (54.9 – 70.1)	63.6 (54.7 – 72.6)

Table 4.7: Model performance on demographic group splits. Mean (95% CI)% accuracy of the RF-NLP model for the female/male and age groups on the test set with 10 bootstrap repeats. The likelihood of a positive ADRD prediction for each demographic is also presented, along with the proportion of positive and negative labels on the test set.

Split	Total	Accuracy	$P(\hat{y}=1 \mid \text{Demographic})$	Pos : Neg
Female	44	76.4 (74.6 – 78.1)	42.3 (38.6 – 46.0)	1 : 1.1
Male	27	76.7 (72.2 – 81.2)	43.3 (37.9 – 48.8)	1 : 0.9
50 – 59 years	14	83.6 (78.2 – 89.0)	42.1 (34.7 – 49.5)	1 : 1.3
60 – 69 years	27	75.9 (72.6 – 79.3)	37.4 (31.5 – 43.3)	1 : 1.2
70 – 80 years	30	73.7 (71.0 – 76.3)	47.7 (44.7 – 50.7)	1 : 0.8

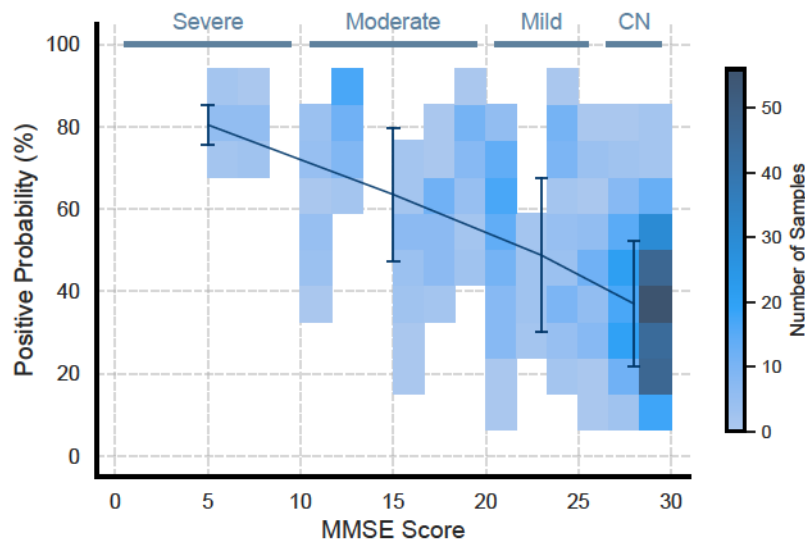


Figure 4.5: Predicted positive cases per cognitive group. The RF-NLP model predicted probabilities for ADRD detection. The total number of samples per cognitive group (CN, mild, moderate, severe, based on MMSE) is shown considering the values from 10 bootstrap repeats. Lower MMSE values reflect worse cognition.

B. Demographic Parity

The model performance on different demographic groups was evaluated based on sex (female, male), and age (50-59, 60-69, 70-80 years). To do this, the predictions obtained by the proposed model were split (with thresholds of $>50\%$ =Positive and $<50\%$ =Negative) on the test set by demographic group and the mean accuracy and SD for each group were calculated. These results, along with the number of participants in each demographic group and binary label proportions, are shown in Table 4.7. The likelihood of a positive prediction across demographics does not show high variations, even with an imbalance in the number of participants of the younger group.

Furthermore, Table 4.7 shows that participants in the oldest age group (70-80 years) have a higher likelihood of a positive ADRD prediction than the younger groups. Overall, this analysis

Table 4.8: Model performance on MMSE group splits. Mean (95% CI)% accuracy of the RF-NLP model for the MMSE groups on the test set with 10 bootstrap repeats. The likelihood of a positive ADRD prediction for each group is also presented. Note that the test set included only two participants with MMSE scores in the severe group, which explains the metrics obtained.

Cognitive Group	Total	Accuracy	$P(\hat{y} = 1 \mid \text{MMSE})$
CN	36	81.9 (76.8 – 87.1)	20.8 (15.0 – 26.6)
Mild	11	48.2 (43.8 – 52.6)	39.1 (34.7 – 43.5)
Moderate	21	78.6 (73.7 – 83.5)	78.6 (73.7 – 83.5)
Severe	2	100	100

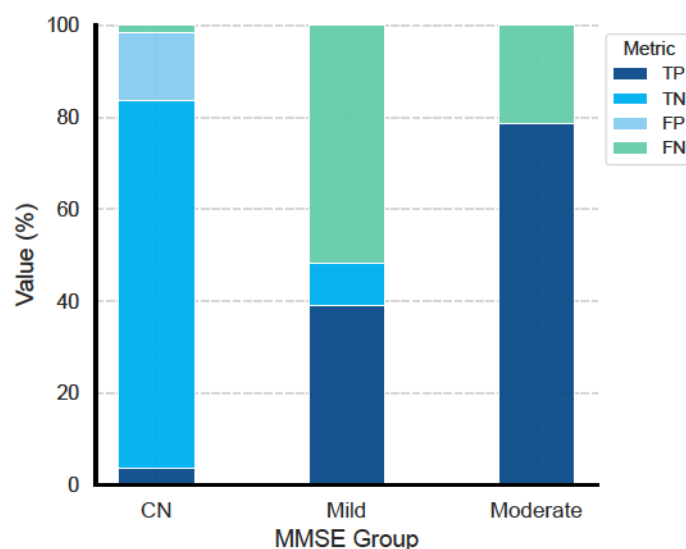


Figure 4.6: Distribution of model predictions for MMSE groups. TP, TN, FP, and FN refer to True Positives, True Negatives, False Positives, and False Negatives, respectively.

suggested that the RF-NLP model is a fair classifier, demonstrating high demographic parity.⁴

4.4.2 Correlation with Cognitive Scores

Figure 4.5 demonstrates that the worse the cognitive impairment, the higher the ADRD predicted positive probability of the RF-NLP model. Although the classifier was trained with binary labels (i.e., not based on severity), the predicted probabilities obtained were correlated with cognitive impairment as measured by MMSE. This property of the modelling approach was particularly interesting as it demonstrates that the model is well-calibrated with both the risk of dementia and its severity without it being explicitly trained on the latter. Furthermore, the model demonstrated lower confidence and higher variability in its predictions for the MCI group. This is anticipated because individuals with MCI exhibit more subtle changes in language [109], increasing model uncertainty in distinguishing from CN individuals.

Model performance was further evaluated on different cognitive groups based on participants' MMSE scores (see Table 4.8). Results showed that the higher the cognitive impairment (lower MMSE score), the higher the likelihood of a positive ADRD prediction of the RF

⁴Measuring fairness (<https://pair.withgoogle.com/explorables/measuring-fairness/>).

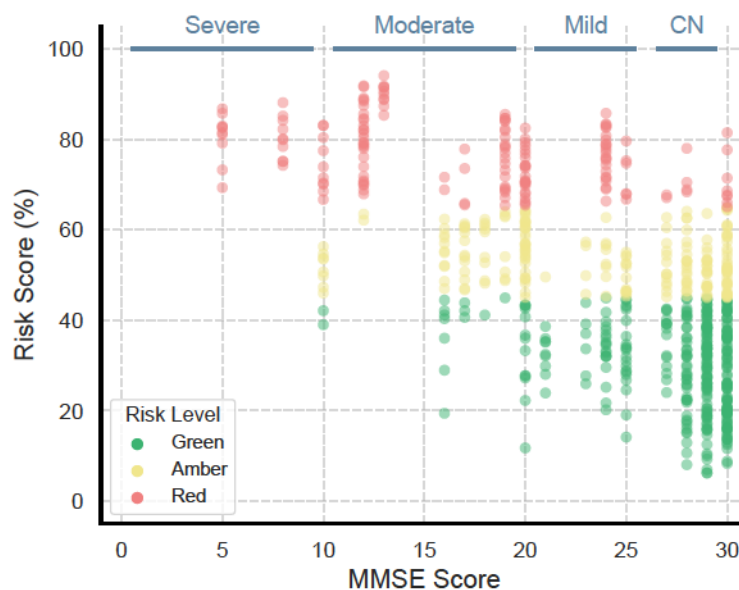


Figure 4.7: Risk level distribution by MMSE scores. Distribution of the Green, Amber and Red risk groups across each MMSE score on the test set for the RF model using explainable linguistic features. The prediction results are reported considering 10 bootstrap repeats.

model trained on NLP-based linguistic features. This finding supports Figure 4.5, which shows that the model is more confident (i.e., makes fewer mistakes) for the moderate and severe cognitive impairment groups compared to the mild and CN. Additionally, Figure 4.6 shows that the number of true positives increases for participants with moderate cognitive impairment compared to mild, while the number of true negatives is the highest for the CN group. This supports the previous finding that the model is well-calibrated with both the risk of dementia and its severity, despite not being explicitly trained on the latter.

4.4.3 Risk Analysis

To improve model flexibility and clinical applicability, risk thresholds were calculated on the validation set that represent minimal (Green), medium (Amber), and high (Red) risk of AD RD (see details in Section 4.3.5). Thresholds were selected as [0%, 45%], (45%, 65%], and (65%, 100%] (following interval notation) for Green, Amber, and Red risk groups, respectively.

Figure 4.7 shows the distribution of AD RD risk levels for each MMSE score on the test set. Furthermore, grouping the predictions of Red and Green risk levels following a selective classification approach [76] enhanced model performance to a ROC-AUC of 88.7% (95% CI=86.2–91.2) sensitivity of 67.6% (95% CI=62.1–73.2) specificity of 96.7% (95% CI=93.3–100) and accuracy of 83.6% (95% CI=80.6–86.5) on the test set. Of particular note, when excluding the Amber risk group, the model better captured true negative cases (i.e., CN). Such risk analysis approach can be beneficial for triaging in a clinical context. By categorising patients into different risk groups based on their use of language, clinicians can prioritise those who require immediate attention and further diagnostic workup. This method allows for a

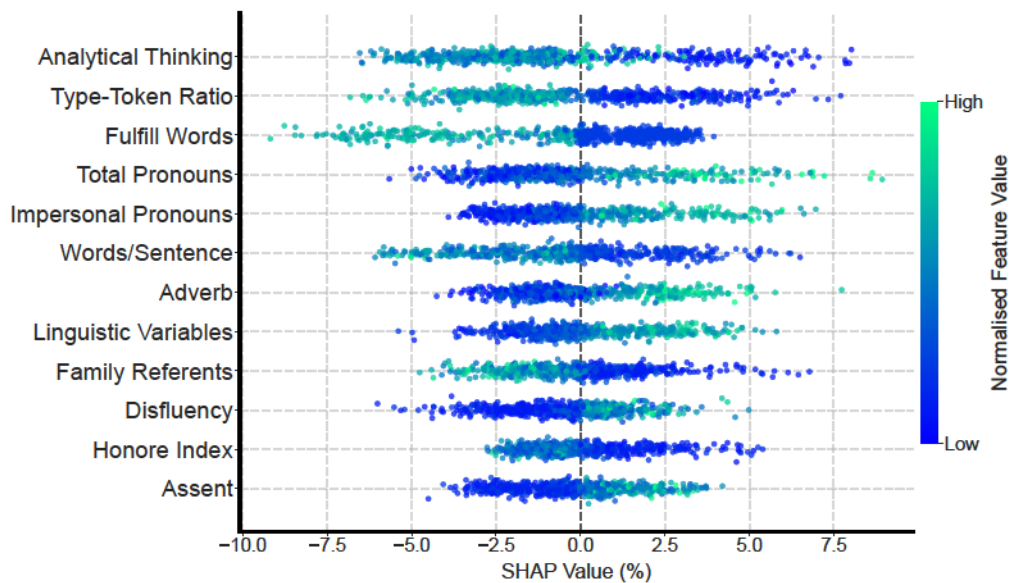


Figure 4.8: SHAP results. The feature importance for the top 12 most important features on the test set and their corresponding feature values from the RF-NLP model. Lower SHAP values suggest reduced risk of ADRD. The colour represents the normalised feature value, and the position in the x-axis represents the contribution that value made to the prediction.

more efficient allocation of medical resources, ensuring that higher-risk individuals receive timely intervention.

4.4.4 Linguistic Feature Importance

The most important features influencing predictions were evaluated using SHapley Additive exPlanations (SHAP) [156]. This method calculates the contributions of individual features to risk scores, providing explainable predictions. The SHAP results on the test set can be seen in Figure 4.8.

This analysis indicated that lower ADRD risk (i.e., true negative cases) was associated with higher levels of *analytical thinking*, higher lexical diversity (as measured by type-token ration), more frequent use of *fulfill words* (i.e., words expressing satisfaction or completion, indicating higher semantic complexity), greater average *words per sentence* and more frequent references to *family-related* words. Conversely, SHAP analysis indicated that more frequent use of *pronouns*, particularly *impersonal pronouns* (e.g., 'that', 'it', 'this'), as well as *adverbs* (e.g., 'there', 'so', 'just'), higher disfluency, and notably increased use of *assent* words (e.g., 'yeah', 'ok', 'yes'), contributed to higher ADRD risk (i.e., true positive cases). Additionally, lower Honore's index values – indicative of reduced vocabulary richness and increased repetitiveness – were linked with higher ADRD risk predictions.

The increased reference to family-related words by CN participants suggests greater cognitive inference ability and more detail provided in picture descriptions.⁵ Furthermore, the

⁵Note the *Cookie Theft* picture illustrates family activities in a kitchen setting (see Appendix A4).

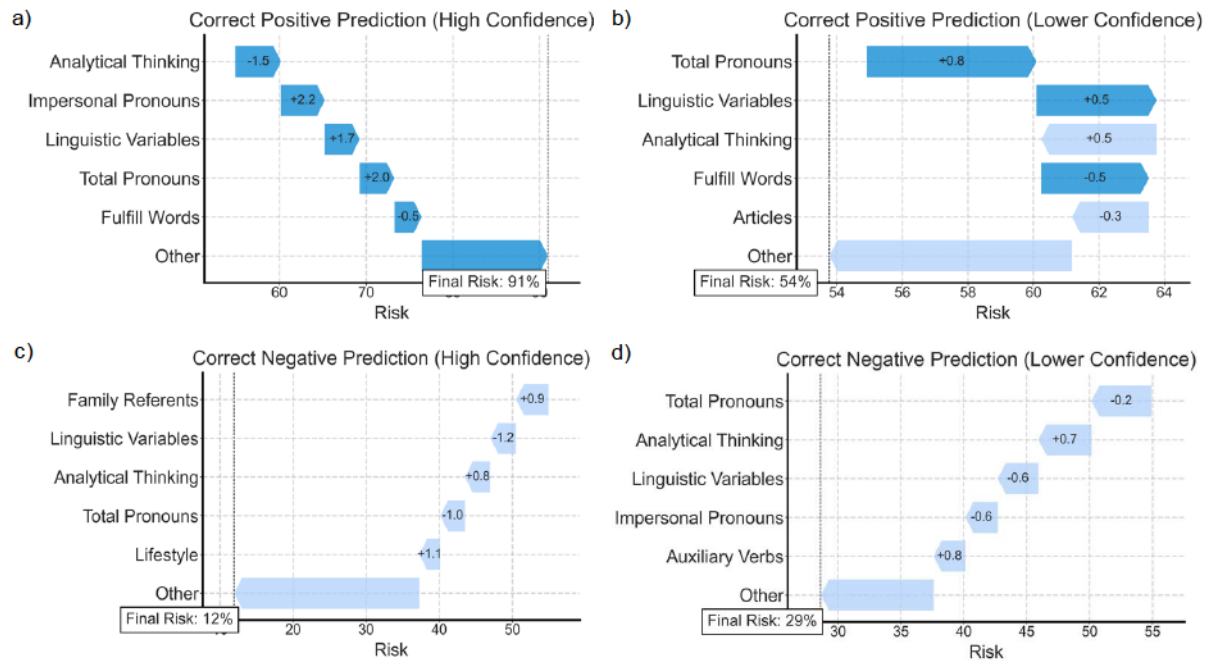


Figure 4.9: Feature importance results for individual predictions. SHAP values of a single prediction shows how each feature contributed to individual predictions based on SHAP values: a) a correct positive prediction (i.e., risk of ADRD) with a high risk score of 91% (participant in the moderate group, MMSE=13); b) a correct positive prediction with a lower risk score of 54% (participant in the mild group, MMSE=25); c) a correct negative prediction (i.e., CN) with a risk score of 12% (participant in the CN group, MMSE=29); d) a correct negative prediction with lower confidence and a risk score of 29% (participant in the CN group, MMSE=27); Here, the values on the arrows represent the normalised feature value in standard deviations from the mean.

increased frequency of pronoun and adverb usage among participants with greater cognitive impairment may suggest difficulty in retrieving specific terms, relying on a more restricted and less diverse vocabulary to describe the scene. This could also be indicative of prolonged cognitive processing times, increased hesitations, word-finding difficulty and reduced linguistic complexity.

Single predictions can also be evaluated to understand contributions to a specific risk score (see Figure 4.9). This visualisation can enable clinicians to explore the factors behind each model prediction. For example, Figure 4.8a shows an example of a correct positive prediction with 91% risk driven by lower *analytical thinking*, more frequent *impersonal pronouns* and overall *linguistic variables*, and decreased use of *fulfill words*. Furthermore, Figure 4.9c shows a correct negative prediction (i.e., participant belongs to the CN group) with a predicted risk of 12%. Significant contributors to this prediction include frequent references to *family* and *lifestyle*, higher levels of *analytical thinking*, and less frequent use of *pronouns* and other linguistic variables.

Table 4.9: MMSE prediction results for the 10-fold CV. MAE and RMSE results of the five regression models evaluated, reported as mean (SD) for the 10-fold CV.

Model	MAE	RMSE
RFR	4.8 (0.5)	5.9 (0.7)
XGBoost	4.8 (0.5)	5.9 (0.8)
SVR	5.3 (0.9)	6.4 (1.2)
RR	6.0 (1.2)	7.5 (1.3)
MLP	6.6 (1.7)	8.3 (1.8)

Table 4.10: Severity prediction results using the best-performing RFR-NLP model. Evaluation metrics include MAE and RMSE, reported as mean (SD) for the 10-fold CV, and as mean (95% CI) for the test set with 10 bootstrap repeats.

	MAE	RMSE
Validation	4.8 (0.5)	5.9 (0.7)
Test	3.7 (3.7 – 3.8)	4.7 (4.6 – 4.8)
Pilot study	3.3 (3.1 – 3.5)	4.2 (3.9 – 4.4)

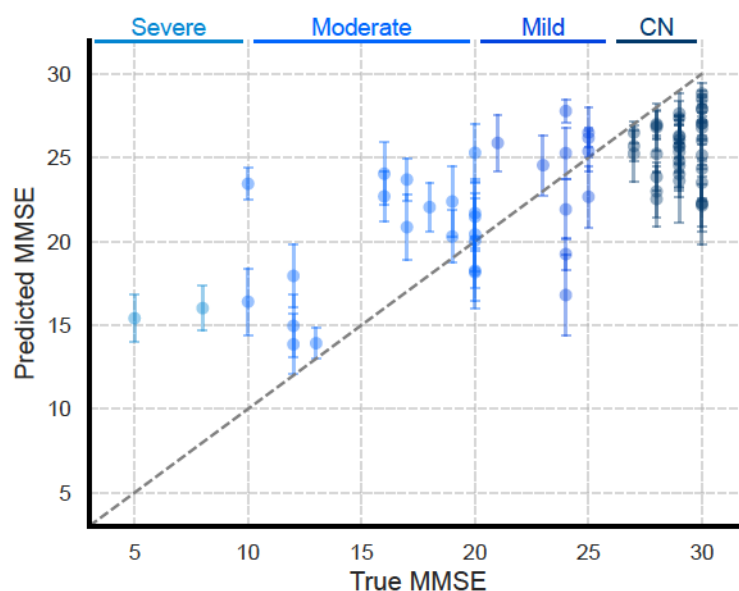


Figure 4.10: Model performance in individual MMSE score prediction. Predicted MMSE versus true MMSE for each participant on the test set. The error bars represent the SD of the values from the 10 bootstrap repeats for each participant.

4.4.5 Model Performance in Severity Prediction

The analysis presents the effectiveness of RR, SVR, RFR, MLP Regressor, and XGBoost Regressor in predicting MMSE scores. Results showed the best-performing model using NLP features was RFR (referred to as RFR-NLP) with a minimum of two samples per leaf chosen through hyperparameter optimisation (see Table 4.9). This model achieved a MAE of 3.7 (95% CI=3.7–3.8) on the test set (see Table 4.10).

Figure 4.10 shows the absolute error per individual prediction on the benchmark DementiaBank test set, categorised by participants across different cognitive groups. Additionally,

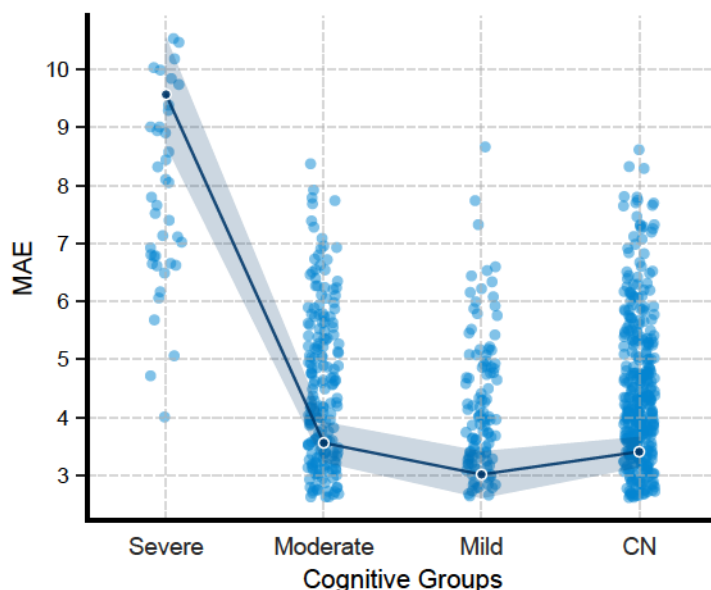


Figure 4.11: Model performance in MMSE prediction across cognitive groups. Average MAE for predictions on the test set with 10 bootstrap repeats. The error intervals represent the SD of the values from the bootstrap repeats.

Table 4.11: Participant word transcription proportion by cognitive group. Results are reported as mean (SD)% for transcripts analysed on the test set.

	CN	Mild	Moderate	Severe
Word proportion (%)	91.1 (15.2)	84.7 (13.8)	73.5 (17.6)	40.6 (29.4)

Figure 4.11 illustrates the average MAE per cognitive group on the same test set, showing all predictions with 10 bootstrap repeats.

The model showed better predictive power for higher MMSE scores, which could be due to the uneven distribution of the available data, with severe cognitive impairment representing only 5.7% of the test set. The higher MAE observed for the severe cases could also be attributed to noise on the linguistic features used by the ML model for predictions. Notably, there was a moderate positive correlation (Spearman's rank $r = 0.53$, $p < .05$) between the proportion of participant-only transcribed speech and MMSE scores, indicating that those with worse cognition required more intervention from the administrator during the task. Particularly, participants in the moderate and severe cognitive groups, with lower mean participant transcription word proportion of 73.5% and 40.6%, respectively, received more frequent intervention from the administrator (e.g., "What else is going on?", "Can you tell me what else is going on in that picture?") (see Table 4.11). This increased administrator input introduced noise into the linguistic features used by the ML model for predictions, which could explain the higher MAE obtained in MMSE prediction for the severe group.

4.4.6 Real-World Pilot Study

The analysis was extended to an independently collected speech data, applying the model without re-training, to assess the generalisability and applicability of the ML approach in a real-world context. Multilingual speech samples were collected from 22 older adults living in retirement homes who completed the same picture description task in English or Spanish (see details in Table 4.1). The RF-NLP model achieved a ROC-AUC of 65.4% (95% CI=54.7–72.6) on this new dataset, as reported in Table 4.6. Although this classification performance is lower than that observed in the two benchmark DementiaBank test sets used, particularly on the ability to detect true negatives (i.e., Specificity), it shows promise for the predictive modelling approach as a complementary tool to inform clinicians about higher risk of cognitive decline from speech data collected in real-world settings.

The lower model performance observed on this pilot dataset may be attributed to the predominance of participants in the mild and CN cognitive groups (see details in Table 4.1). These groups generally exhibited lower predicted probabilities for distinguishing cognitive impairment (see Figure 4.5), so their higher representation in the pilot dataset likely contributed to this effect.

Using the same risk thresholds (see Section 4.3.5) and grouping predictions of Red and Green levels enhanced model performance to ROC-AUC of 67.3% (95% CI=61.4–73.1) and Accuracy of 60% (95% CI=51.5–68.5), with higher Specificity of 73.5% (95% CI=55.4–91.6) at the expense of lower Sensitivity of 53% (95% CI=37.4–68.7).

Additionally, when predicting MMSE scores (severity prediction), the RFR-NLP model achieved a MAE of 3.3 (95% CI=3.1–3.5), improving results from those obtained on the test set. Despite the small sample size, this pilot study underscored the potential of using linguistic features from spoken language transcripts obtained from spontaneous speech (e.g., picture descriptions) as indicators of cognitive state and automated screening of cognitive impairment, even when collected in real-world settings beyond clinics, such as retirement homes.

4.5 Discussion

This chapter presented an explainable ML pipeline for automated assessment of cognitive health and ADRD severity prediction via MMSE scores from spoken language, with a focus on clinical applicability. This section summarises the main findings of the chapter and their implications for future research towards automated cognitive assessment and monitoring in real-world settings beyond clinics.

A. Summary of Findings

This study used a benchmark DementiaBank dataset (N=291) obtained during verbal picture descriptions in NPT. To validate model generalisability, a separate real-world pilot study with multilingual speech data were collected in-residence from 22 older adults. ML models were trained on various acoustic and linguistic features, extracted using conventional methods based on domain knowledge and transformer-based pre-trained language models. This study focused on clinical applicability, therefore the following factors were prioritised in the analysis: 1) feature interpretability to inform clinicians of changes in spoken language patterns that indicate cognitive decline; 2) accessibility for efficient in-residence data collection. The final model incorporated 100 NLP-based features – including lexical diversity and semantic psycholinguistic features – extracted from individual transcripts obtained through ASR.

A1. ML models trained on interpretable linguistic features effectively identified ADRD and predicted individual MMSE scores on a benchmark DementiaBank dataset, demonstrating generalisability in a real-world setting.

The best-performing RF-NLP model achieved a ROC-AUC of 85.7 (95% CI=83.8–87.6) on the test set (N=71) in ADRD detection. ROC-AUC was selected as the primary evaluation metric to assess the model's ability to classify true ADRD cases while minimising false positives. On the unseen pilot dataset (N=22), the model achieved a lower ROC-AUC of 65.4 (95% CI=54.9–70.1), though the sensitivity was maintained (70% vs. 69.4% in the DementiaBank test set). It is important to note that the main clinical utility of using voice and spoken language as a biomarker lies in the ability for early screening, making sensitivity an important metric for correctly identifying individuals at risk. However, other steps should be taken to minimise the burden and effect of false positives. For example, as demonstrated with the DementiaBank data, using more structured assessments (e.g. describing a picture) could be used to improve the specificity of the model as a second-tier screening.

The pilot data included 22 individuals with a mean MMSE of 24.9 (see Table 4.1), with most participants in the MCI and CN groups. The higher proportion of these participants likely contributed to model uncertainty, as individuals in these groups often exhibit less pronounced changes in language [109], making it more difficult for the model to distinguish between cognitive groups.

Furthermore, a RFR achieved a MAE of 3.7 (95% CI=3.7–3.8) on MMSE prediction, outperforming previous results using the same DementiaBank dataset [3, 157]. The model then achieved a comparable MAE of 3.3 (95% CI=3.1–3.5) on the pilot dataset, demonstrating its generalisability and real-world applicability. This result is particularly interesting given that the model is trained only on the DementiaBank dataset, and is exclusively used for predicting cognitive state through MMSE scores on the pilot dataset. This suggests that the model has captured transferable knowledge and could operate as an out-of-the-box solution without requiring retraining.

A2. Risk stratification enhanced the identification of higher-risk individuals and clinical applicability of the ML approach.

Identification of risk groups aimed for the effective management of high-risk alerts (Red) by identifying people with increased risk of cognitive deterioration while minimising false alerts. Risk stratification resulted in improved model performance considering the low-risk (Green) and high-risk (Red) groups, evidenced by an increase in ROC-AUC, with a notable 13% increase in specificity on the test set. Similarly, on the pilot dataset, the analysis led to a higher ROC-AUC and a 21% increase in specificity at the expense of reduced sensitivity. Although the small size of the training dataset limits broader conclusions on clinical effectiveness, this risk stratification analysis offers a comprehensive approach to selective classification, which had not been previously explored in the context of ADRD screening from spoken language. This approach is proposed for future studies with larger cohorts as a way to alert clinicians to people with increased risk, enhance resource allocation, ultimately enabling more personalised and timely interventions.

A3. Feature importance analysis revealed the language patterns most strongly predictive of cognitive decline.

Feature importance analysis enhanced the interpretation and clinical applicability of the proposed ML pipeline by identifying the linguistic features most strongly predictive of ADRD risk. The findings revealed that increased reliance on pronouns, particularly impersonal pronouns (e.g., 'that', 'it', 'this'), greater disfluency, particularly with assent words (e.g., 'yeah', 'ok', 'yes'), and lower lexical diversity with repetitive language all contributed to higher ADRD risk, in line with previous literature [32, 175, 181, 267]. The frequent use of pronouns and assent words likely indicates empty, vague, or non-specific speech, a known characteristic of cognitive decline [93].

The results showed that language associated with reduced analytical thinking, decreased use of words reflecting a psychological state of completion (e.g., 'enough', 'full', 'complete'), and higher use of adverbs (e.g., 'there', 'so', 'just') contributed to positive predictions. These findings suggest that participants with ADRD exhibit a decline in words reflecting cognitive processes related to structured and logical thinking. Words related to psychological com-

pletion, which typically indicate higher semantic complexity, were less common, potentially reflecting difficulties in articulating complete thoughts. Additionally, the results showed that longer sentences and more frequent references to family were associated with lower ADRD risk (i.e., true negative cases). These findings suggest that CN participants tended to provide more detailed and contextually rich descriptions of the *Cookie Theft* picture with greater inferences regarding relationships. Further investigation by language and cognition experts is needed to generalise these findings.

B. Spoken Language as a Biomarker for Automated Screening of Cognitive Health

The monitoring of cognitive health is essential for early screening and timely intervention, both clinically and in daily care. The 2024 report of the Lancet Commission on dementia prevention, treatment, and care [153] emphasises the importance of timely diagnosis in supporting the well-being of PLWD and their families, facilitating access to services, and ensuring that individuals can benefit from treatments when they are most likely to be effective.

There is hope for disease-modifying treatments for early-stage AD, with recent trials of amyloid- β -targeting antibodies showing modest efficacy, creating a clear therapeutic window of opportunity for intervention, before which adequate diagnosis should be made [233, 254]. Additionally, timely assessment can help reduce unnecessary hospitalisations and improve overall dementia care. The report also highlights the growing role of mobile and wearable devices in detecting neurodegeneration, given their widespread use and ability to continuously monitor physical and cognitive changes.

The use of digital voice biomarkers for early ADRD screening has gained attention in the research community, driven by the need for scalable, non-invasive, and cost-effective solutions. Alterations in lexical and semantic content, syntactic complexity, and fluency can be predictors of cognitive changes linked to ADRD, even before clinical signs of neurodegeneration emerge [4, 77]. The proposed interpretable predictive modelling approach could be integrated with home-based conversational AI. With consistent use, these technologies hold potential to become accessible and personalised tools that could ultimately track the trajectory of cognitive status over time through spoken language, alerting clinicians to individuals who may need more comprehensive diagnostic evaluation. Integrating additional in-home behavioural data, comorbidities, and individual health events such as hospitalisations or infections [38, 148], could further improve prediction performance and enhance clinical applicability.

C. Limitations and Future Work

This study includes limitations that point toward future research directions. The use of ASR systems, such as OpenAI's Whisper [202], introduces transcription errors, particularly for participants with severe cognitive impairment and higher speech disfluency. While ASR can affect the extraction of linguistic features, this was an intentional design choice to assess

the feasibility of automated screening of cognitive health in real-world settings where human annotation is impractical. The quality of the DementiaBank audio data used for training can also impact the accuracy of linguistic and acoustic features extracted for model development.

Cognitive screening tools developed primarily in white, English-speaking populations may not generalise well to more diverse populations due to differences in education and cultural backgrounds [174]. Although the pilot study contained 36% of speech recordings in Spanish (see Table 4.1), to improve the generalisability of the findings, future studies should include larger, more culturally diverse populations and explore predictive linguistic features in languages beyond English. Additionally, speaker diarisation was not performed. Despite testing existing systems to integrate with the end-to-end pipeline for automated analysis, they proved unreliable in separating participant and administrator speech. However, when inspecting manual transcriptions from DementiaBank, a high correlation was found between linguistic features obtained from participant-only transcriptions and those that also included short segments of administrator speech, which justified the decision. Furthermore, the rate of participant-only transcribed speech was correlated with the MMSE scores, with participants in the moderate and severe cognitive groups receiving more intervention from the administrator, which introduced noise into the linguistic features used by the ML model for predictions. As automatic speaker diarisation techniques improve, incorporating this pre-processing step could enhance future analysis.

This study evaluated ML models for automated screening of cognitive impairment using acoustic and linguistic features, as well as feature combinations using early fusion methods. The most clinically interpretable feature sets and their corresponding best-performing models were selected for further analysis. Future work could conduct an ablation study to assess how each feature contributes to model performance by systematically removing specific acoustic or linguistic features and observing the impact on performance. This fine-grained approach could further exclude less important features from predictions, reducing overfitting and ultimately enhancing model performance and interpretability.

Future studies could integrate additional predictive features, such as age, gender, education, and family history of dementia, which could improve models for ADRD screening and severity prediction. Moving beyond binary classification by including MCI or other neurodegenerative disorders as prediction classes would also expand the scope and utility of this analysis. Moreover, the acoustic and linguistic features were extracted from spontaneous speech during a picture description, a standardised cognitive task. While standardised tasks provide a replicable framework and facilitate benchmarking comparisons, future work could combine features extracted from unstructured speech interactions, such as natural conversations with conversational AI. Longitudinal data collection using these technologies could advance scalable and accessible screening tools for earlier detection of cognitive decline.

4.6 Summary

This chapter evaluated the use of spoken language as a biomarker for automated assessment of cognitive health. An interpretable ML approach was introduced to evaluate the effectiveness in detecting *ADRD* in binary classification and predicting the severity of cognitive decline via *MMSE* scores.

Speech recordings from a benchmark DementiaBank dataset (N=291) were used for model training and testing with 100 NLP-based interpretable features extracted from transcripts. This research demonstrated the generalisability and applicability of the ML approach on a real-world dataset collected in-residence (N=22). In cognitive impairment detection, the best-performing model achieved a ROC-AUC of 85.7% (95% CI=83.8–87.6) on the benchmark DementiaBank test set and 65.4% (95% CI=54.9–70.1) on the real-world data. In *MMSE* prediction, a MAE of 3.7 (95% CI=3.7–3.8) was achieved, with a comparable performance of 3.3 (95% CI=3.1–3.5) on the pilot data.

By evaluating thresholds for risk stratification – an approach that had not been explored previously in the context of automated *ADRD* screening – this analysis enhanced the clinical utility of model predictions, aiming to optimise healthcare resource allocation by identifying higher-risk individuals. Additionally, this research provided insights that inform clinicians of linguistic features associated with higher *ADRD* risk, including: reduced analytical thinking, increased reliance on pronouns and adverbs, greater disfluency, lower lexical diversity and reduced use of words reflecting a psychological state of completion.

The proposed ML approach shows promise for real-world integration with in-home conversational technology for accessible and scalable monitoring of cognitive health. Such an approach could facilitate earlier detection and timely intervention. However, to realise this potential, further research is needed to address the feasibility of in-home use, long-term user engagement and technology acceptance. The following chapter investigates these factors through comprehensive user studies and stakeholder engagement.

Robot Evaluation and User Engagement in Real-World Settings

This chapter presents three user studies conducted in the UK, India and the USA, involving a total of $N=111$ stakeholders – older adults, PLWD, carers, clinicians and dementia care professionals – to evaluate user engagement, acceptance, and the feasibility of using conversational AI and SARs in residential settings. The chapter aims to explore the perceived benefits and concerns of using conversational AI to support well-being at home, offering insights for long-term user engagement and appropriate home integration. Study 1 explores the role of conversational AI at home through participatory workshops. Study 2 describes a cultural feasibility study in India. Finally, Study 3 presents the design and evaluation of an LLM-powered SAR for in-residence cognitive support through human-robot verbal dialogue.

The work presented in this chapter resulted in the following contributions:

- **Study 1** was published and presented in oral form at the *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2023 [145], and the *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2024 [144].
- **Study 2** was published in *Frontiers in Robotics and AI*, 2021 [150], the *American Journal of Geriatric Psychiatry*, 2021 [176], and was submitted to *Innovation in Aging* (under review).
- **Study 3** was accepted for publication and oral presentation at the *ACM CHI Conference on Human Factors in Computing Systems*, 2025 [146].

5.1 Study 1: The Role of Conversational AI in Dementia Care at Home

This study used a community-based participatory approach to involve people with lived experience of dementia and frailty. Three participatory workshops were conducted, including group activities and prototype testing of voice-based HRI. The activities encouraged stakeholders to think critically about technology features and use case scenarios of support that are suitable, acceptable, and desirable for long-term utility in dementia care at home. The contributions of this study are described below.

Contributions:

1. Expert-informed insights on the role of conversational AI in supporting healthy ageing and dementia care at home.
2. Design guidelines on the feasibility for long-term adoption and appropriate integration of conversational AI in the daily lives of target populations.

Ethics Approval: The study was conducted under the UK DRI-CR&T's initiatives for patient and public involvement and engagement (PPIE).

5.1.1 Background

Human-centred design methods are key to developing useful and acceptable conversational AI applications for long-term use by target populations [187]. These methods involve participatory design and community-based research, including PPIE [251, 263], to gain insights into stakeholders' attitudes and perceptions towards adopting the technology. This can help identify the perceived benefits, barriers and concerns for the specific support context, which can further be translated into the conceptualisation of new design features and guidelines [185].

Despite increasing interest in conversational AI in healthcare applications, there is a lack of studies that involve stakeholders to explore real use cases of support, preferences, and desired technology features to guide the development of useful and accessible conversational technology for dementia care at home. Further community-focused research is needed to understand the role of conversational AI in supporting people affected by dementia, particularly in promoting independence and well-being at home.

5.1.2 Methods

This study followed human-centred design methods to involve a range of stakeholders – PLWD, carers, family members, and older adults with mild to moderate frailty – in the scoping of this technology. Through participatory design activities, participants were invited

to share their preferences, expectations, and concerns for home-based conversational agents to ensure the technology is designed to meet end-user needs. Discussions focused on the usefulness of use case scenarios within the context of dementia care, the role of technology and its desired interactive features, the benefits for end-users, as well as barriers that would inhibit the adoption of this technology.

5.1.2.1 Research Questions

This research was guided by the following questions:

- **RQ1:** How can conversational technology support people affected by dementia at home?
- **RQ2:** How do stakeholders conceptualise conversational technology now and in the future?
- **RQ3:** What are the perceived benefits and barriers to the adoption of conversational technology at home?

5.1.2.2 Participants

Three PPIE workshops were conducted with 20 stakeholders recruited through community groups and local day centres. In the first workshop, participants had direct experience caring for individuals affected by dementia ($N=9$, mean age=63, $SD=9.6$, 9 females, P1 to P9). The second involved older adults with age-related health needs ($N=3$, mean age=65, $SD=6.9$, 1 female, P10 to P12). Participants in the third workshop were older adults living with dementia and/or mild to moderate frailty ($N=8$, mean age=74, $SD=7.4$, 3 females, P13 to P20). All workshops took place in London, two at the DRI-CR&T and one at an Age UK day centre for older people in South London.

Participants were generally familiar with existing commercial examples of voice technology, such as Siri, Amazon Alexa, and Google Home, but had little to no experience using voice technology. Furthermore, some participants ($N=8$) were part of the cohort involved in the long-term data collection of in-home activity using remote monitoring devices and Amazon Alexa (see Chapter 3). Upon providing informed consent, learning about the purpose of the workshop, and having an opportunity to ask questions, participants engaged in participatory design activities in groups of three or four.

5.1.2.3 Participatory Workshops

Each PPIE workshop lasted three hours and was carried out following PPIE practice guidelines [251]. The activities and discussions were structured into the following categories: 1) technology role, 2) technology features, 3) imagining in daily life, 4) benefits and opportunities, 5) barriers to use and concerns. Group activities were designed to explore stakeholders' views and preferences on different functionalities and features of the technology, as detailed in

Table 5.1: Scenarios and prompts used in participatory workshops.

Category	Scenarios and prompts
Technology role	Support Monitor Alert Remind Entertain Inform
Technology features	Multimodality Embodiment Mobility Personalization
Imagining in daily life	Location in the house Portability Proactivity Smart home integration

Table 5.1. Figure 5.1 shows participants engaging in the workshops. Each group had a facilitator to moderate discussions and record participants' perspectives. To explore the perceived benefits and barriers to technology adoption, the facilitator led the group through reflective open-ended questions to encourage further discussion.

Six contextual scenarios of interaction were selected to explore the role of the technology at home: offer verbal support in ADLs (e.g., taking vitals); monitor health and wellbeing (e.g., daily questionnaire); trigger alerts to the carer (e.g., upon a detected change in sleep patterns); set reminders (e.g., medication); entertain (e.g., play music or engage users in conversation); inform (e.g., tell the news or events of the day). The scenarios were reviewed by an occupational therapist to ensure that they were appropriate and useful for discussion with the target population. Each scenario was presented on printed cards with dialogue that simulated interactions (Figure 5.1c). These cards included a persona example and a prompt for the interaction, either user-initiated or automatic. Participants read each of the scenarios interactively before proceeding with the remaining activities. This helped contextualise the breadth of applications this technology can have in dementia care.

Participants tried out two demos of conversational agents: Amazon Alexa through an Alexa skill prototype developed for this study that provided verbal assistance in taking a blood pressure reading from a smart monitor (see Figure 5.1d), and a virtual affective robot capable of showing eight facial expressions (based on previous affective robotics work [150])¹. The robot was integrated with a bespoke conversational agent² that asked well-being questions.

¹Note that details on the robot are provided in Study 2 (see Section 5.2).

²Rasa, an open source conversational AI platform, was used for implementation (<https://rasa.com/>).



Figure 5.1: Activities, material and demos used in the participatory workshops. a) and b) stakeholders engaging in participatory activities; c) dialogue cards designed to simulate HRI support scenarios; d) a participant interacting with a conversational agent prototype that provides verbal and visual guidance for taking a blood pressure reading. Photos used with consent.

A grounded theory approach was used for thematic analysis [43] to identify the emerging themes reflecting stakeholders' perspectives. Two researchers independently coded the information gathered during the workshops. Additionally, participants completed a questionnaire to evaluate individual preferences for technology capabilities such as adaptation over time, emotion recognition, integration with home activities, data collection and access.

5.1.3 Results

The diversity of experiences in ageing and dementia care encouraged participants to share valuable insights on how conversational technology may effectively support people affected by dementia in their homes. The participatory activities focused on the scenarios and prompts described in Table 5.1 along with open discussions on the perceived benefits and barriers. The analysis revealed three themes: 1) the perceived role and useful support scenarios for dementia care, 2) preferences for technology features and interaction capabilities promoting home support, and 3) feasibility considerations for long-term adoption.

5.1.3.1 Perceived Utility for Home Support

Participants discussed use case scenarios for conversational technology in the context of supporting people affected by dementia at home (Table 5.1). Three primary roles were iden-

tified: helping PLWD remain independent for longer, particularly by supporting daily function; increasing safety at home, particularly by mitigating risk; and enhancing cognitive engagement. These roles were seen as particularly important given that target populations may be at increased risk of accidents and behavioural changes as daily tasks become increasingly challenging.

A. Supporting Daily Function

Overall, participants conceptualised conversational agents as valuable interactive tools to stimulate action by giving automatic prompts. Supporting target users in carrying out ADLs was perceived as a useful scenario of support. Participants suggested this could be achieved by sending reminders of daily tasks, providing verbal instructional guidance when needed, and suggesting actions, *"I would find it useful if it reminds me of events like 'you have a hospital appointment today', I currently have posters all over my kitchen"* (P16, older adult who lives alone). A particular focus was placed on the usefulness of reminders for routine management and medication tracking. Participants felt that this could support independence and help to alleviate the burden placed on carers.

Furthermore, they stressed reminders and other prompts should be personalised to the user's needs and preferences for increased engagement. The ability to track cognitive decline and dementia progression was suggested as a promising technology feature. However, there were some concerns regarding the accuracy of this approach when the technology is not used consistently.

B. Monitoring and Mitigating Risk

The usefulness of integrating voice technology with other devices or home activities was highlighted, particularly to trigger alerts that might support or manage risk and safety at home. Overall, a monitoring system was suggested, which would include alerts automatically sent to the carer, *"a way to watch the individual whilst the carer is elsewhere [for safeguarding]"* (P8, carer). Several participants discussed the ability to act upon falls at home by calling emergency care or friends and family, for which the portability and hands-free nature of the technology was suggested as a key feature to enable timely and effective responses.

Furthermore, emphasis was placed on the potential for the technology to reduce episodes of acute care (such as in the event of a fall) by promptly addressing the user's verbal concerns, questions or commands. Participants also suggested using the technology to regularly check on individuals recovering from hospital admissions. As one participant noted, *"after I've had a seizure [...] I wouldn't be able to use a phone or click a pendant alarm. Having something you could just ask to call an ambulance would be incredibly helpful and likely save people."* (P12, older adult with epilepsy). Keeping a record of health events in the house and sharing it with healthcare providers were deemed beneficial, *"it would provide me peace of mind"*

(P12), as was using the technology as an alternative careline service. Participants referred to other valuable home integration examples, such as alerting carers when the oven is left on or encouraging users to maintain nutrition and hydration.

C. Enhancing Cognitive Engagement

Discussions indicated the strong potential to encourage individuals to engage in positive activities that may benefit their well-being. Carers with experience using smart speakers valued their entertainment capabilities (e.g., playing music and audiobooks, hearing the news).

Participants recognised the potential for cognitive stimulation through the use of quiz games and engaging in conversations about topics of interest, which was seen as beneficial for sustaining interaction, "*[conversational technology] gives them the sense of a companion, makes them feel that somebody cares*" (P4, carer). Importantly, it was noted the technology should use the context and content of previous interactions to provide support in a personalised way. Furthermore, participants remarked conversational AI should respect end-users and not infantilise them, "*[Giving reminders and providing support in ADLs could be useful] as long as it does not make me feel like an idiot*" (P17, PLWD), which could compromise autonomy and independence, "*It could make me reverse back to being too needy*" (P18, PLWD).

5.1.3.2 Robot Design Preferences

As part of exploring stakeholders' preferences on technology features, the activities prompted people to consider how they might want to interact with the technology in their homes. This process involved assessing the appeal of screen-based interactions, embodiment and facial expressions.

A. Modes of Interaction

A key takeaway was that regardless of its interaction capabilities, the technology should be simple and easy to use. While voice was the preferred interaction mode, participants acknowledged the supportive role of a screen for visual prompts, such as images, weather updates, diary reminders, or video tutorials for assistance in ADLs (e.g., how to take a blood pressure reading). The screen could also increase accessibility, "*It could make the technology more accessible for people with hearing aids*" (P8).

Discussions with carers were held regarding whether screens might add another layer of complexity to the interaction for PLWD. It was acknowledged, however, that future older generations will become more familiar with screens and technology in general. Many (not all) of the participants were open to the idea of having a face that could show facial expressions. Some suggested this might "*add a friendly face to a robot voice*" (P5, carer), deliver a "*more engaging, comforting, and relatable experience*" (P6, carer), and in some cases "*could make people more responsive*" (P7, family member). A participant further suggested the potential

impact on mood, *"You might be miserable, and if it starts smiling at you, it could lift up your mood"* (P16). However, discussions indicated the preference for an affective robot would be influenced by its visual design and cultural context. Uncertainty and fear were expressed towards human-like features, and participants generally preferred a cartoon-like face.

B. Imagining in Daily Life

The insights indicated that the location of the technology in the house plays an important role in user acceptance and that the technology's intended purpose should drive its location, appearance, and interactive features. Participants expressed reservations about the technology being in the bathroom or bedroom due to privacy concerns. However, when discussing the technology being used to support safety at home, participants' perspectives changed, *"If it was part of an emergency support system, you'd want to make sure you have coverage everywhere in the home"* (P20, older adult). They would want to verbally access the technology or have it integrated with the home environment to access help if they slipped in the shower or had a fall in the bedroom.

Discreet integration into the home environment was further suggested. When asked about mobility, a general preference for a static location and/or portability was expressed with the ability to "hear" from multiple places in the house (e.g., both upstairs and downstairs). Having a robot move around the house was often referred to as "scary". However, participants expressed the technology shouldn't have to be wearable to function, as this would create an additional barrier to adoption. Furthermore, even if static, embodiment was viewed as a potential driver of engagement, particularly as a *"way to take the scariness of a voice coming out of nowhere"* (P4, carer).

C. Communication Style

Participants emphasized the need for thoughtful language, tone, and personalisation when designing conversational agents for people affected by dementia. Discussions uncovered mixed perceptions towards empathetic interactions. There was an overall preference for a conversational agent that is task-based and practical rather than one that transmits false empathy. For instance, expressions such as "I'm glad that..." were perceived as disingenuous, emphasising the need to design conversational agents that communicate in a way that is genuine and authentic to avoid unmet expectations. Others valued an encouraging, softly spoken interaction style to maintain user comfort and independence. As one participant pointed out, *"[current voice] technology is pretty brusque"* (P17). Interestingly, carers generally favoured the first perception, while PLWD and older adults typically preferred the second. These differing (and occasionally conflicting) perspectives align with previous studies on robots for dementia care [135, 265].

Additionally, participants pointed out that care should be taken with open-ended questions. While recognizing that these have the potential to spark engagement, they could also generate anxiety among PLWD, *"They may not know what to say, they may think they forgot about something and feel stressed"* (P6). Instead, they advocated for an instructional tone and noted older adults often respond well to clear instructions. For instance, rather than asking, *"Would you like me to remind you?"*, the conversational agent should state, *"It's time to take your medication"*. Discussions also indicated that continuous prompting may leave users feeling coerced into performing a particular task, undermining their sense of autonomy, *"It can be frightening that you have to do it, that it prompts you all the time"* (P16).

D. Home Integration

Participants agreed that conversational technology should adapt to users' changing needs and learn home routines for proactive support (Figure 5.2a). Perceptions varied on the value of detecting emotions through voice. While some viewed it as a useful feature to detect mood and potentially flag changes, others expressed reluctance for such a human-like feature and noted that emotions could easily be assessed incorrectly in the situation of multiple people interacting with the technology. Despite mixed views, the feature was generally viewed favourably.

Participants explored options for selecting data captured by other home sensors and devices to provide personalised assistance, particularly in a smart home context (Figure 5.2b). Vital signs and sleep patterns were the most selected sources of data (selected by 88% and 82% of respondents, respectively). Capturing motion around the house and kitchen habits were also perceived as relevant (selected by 71% and 59% of respondents, respectively).

Participants also described their preferences for sharing data collected from interactions with conversational AI and other devices for home activity monitoring (Figure 5.2-c). Most participants (88%) were comfortable sharing data with healthcare professionals. This was followed by academic researchers (65%), and family members and carers (59% each). Participants felt less comfortable sharing data with friends and 'big tech' companies, such as Amazon or Google.

5.1.3.3 Perceived Barriers to Adoption

Discussions highlighted practical factors that could inhibit the long-term adoption of technology by target populations.

A. Familiarity and Anticipation

Discussions revealed the challenge of introducing voice technology to current older generations, who may see the technology as beneficial but do not feel ready to start using it due

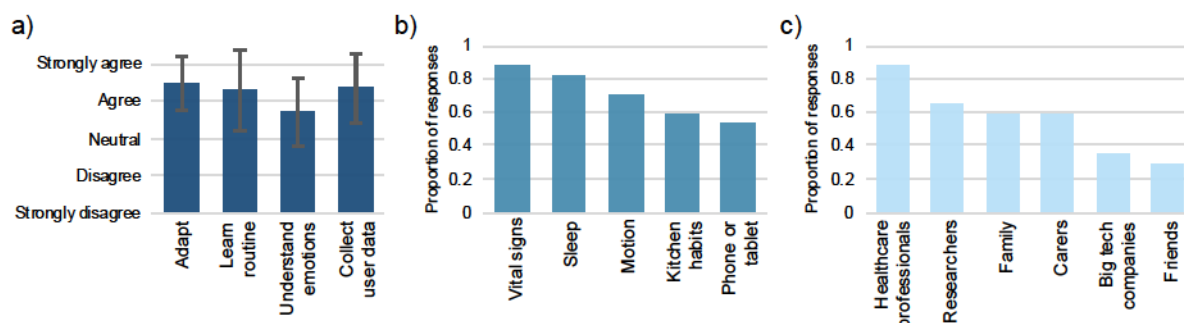


Figure 5.2: Stakeholder preferences for conversational AI features and support functions. a) conversational AI features surveyed, including adaptation to user needs and preferences over time, routine learning, emotion understanding, and data collection for support; b) types of data collected at home; c) data sharing.

to unfamiliarity, *“I would not be ready for that yet. If I become ready it could become an asset”* (P17). Participants remarked that this challenge is likely to change in the future as upcoming generations become more familiar with voice technology. Participants suggested introducing the technology early in the disease and implementing a training period could help to prevent confusion or discomfort.

B. Cost and Accessibility

The cost of the technology was identified as a major barrier to adoption. Further accessibility concerns arose during discussions around the implementation of an IoT ecosystem to integrate conversational AI technology with other home devices. For instance, the setup and technology maintenance could create a reliance on the carer. Additionally, some participants were concerned about connectivity, specifically the overreliance on Wi-Fi, which, if failing to work, could prevent alerts from being triggered in the event of an emergency, posing potential safety risks. As one of the most beneficial roles identified for long-term support was interconnecting with health services, the latter was found to be a priority barrier to overcome.

C. Fear and Mistrust

Trust emerged as a significant barrier to the adoption of conversational technology among target populations. Participants expressed concerns about the technology potentially collecting private or sensitive data, such as financial conversations or even ‘spying’ on people. Such mistrustful attitudes were often shaped by media coverage and stories heard from others and further amplified by a general lack of knowledge of the technology’s capabilities and purpose. Discussions indicated a common concern of not being able to control the technology if something goes wrong, including deactivating it if needed, which could hinder its adoption in the future.

Participants further remarked on the need for transparency regarding the technology’s actions and capabilities to avoid confusion and mistrust, for instance, by verbally reminding the user

of what it can do to assist them. It was suggested that target users are more likely to accept and feel comfortable using AI if they have a clear understanding of what it offers.

D. Autonomy

Participants showed receptiveness towards the idea of the technology being proactive in reaching out to users (e.g., for monitoring purposes through the use of automatic verbal prompts with information or alerts). However, discussions indicated that the frequency and communication style (i.e., language, tone) of such prompts may come across as intrusive. It was suggested that the user or their carer should have the option to choose their preferred prompts and respective timings to ensure that the interaction remains comfortable and non-intrusive, *"I want to be in control, I don't want to be said and led all the time"* (P17). It was also noted that hearing a voice coming unexpectedly could be frightening, particularly when a person is alone at home, and that promoting familiarity with the technology and its capabilities would be crucial to avoid such a scenario.

5.1.4 Discussion

Based on the perspectives and experiences shared by stakeholders, this study presents guidelines for designing conversational agents to effectively support healthy ageing and people affected by dementia in their home environments. These guidelines, along with their implications for long-term engagement and usability, are discussed below.

Guideline 1: *Introduce the technology early by implementing a training period.*

The findings indicated that general unfamiliarity with technology, particularly among the current older generation, inhibits long-term adoption and acceptance of conversational agents. It is important to introduce voice technology as early as possible (in the early stages of dementia) and provide a training period to promote familiarity with its features and capabilities, this way preventing confusion or discomfort.

Guideline 2: *Integrate conversational AI with home monitoring technologies to support independence and safety.*

Functions that support independence and safety at home and help alleviate the burden placed on carers were identified as particularly beneficial and should be explored further. These include reminders for routines or medication tracking and prompts for daily activities. Applications of the technology should leverage the benefits of hands-free and portable communication in scenarios where voice may be the only form of communication to enhance safety. When integrated with other home monitoring technologies, conversational agents hold strong potential to monitor health and well-being and act upon changes in behaviour detected from household activities (such as a fall or lack of kitchen activity in the morning). The conversational agent could directly query users to verify symptoms, provide verbal support by

suggesting relevant actions, and inform carers when needed. The ability to manage risk at home and reduce episodes of acute care was viewed as highly beneficial in dementia care. Stakeholders emphasized the promise of using conversational AI to track cognitive decline and dementia progression. Further investigations on the feasibility and reliability of this promising support avenue are warranted.

Guideline 3: *Adapt interactions and support to the user's changing needs.*

Designing voice interactions for PLWD is a complex task. Adaptation to the user's needs and cognitive skills throughout the dementia journey was found to be critical. Therefore, interactions with conversational AI should adapt to diverse paths of cognitive ageing, ensuring appropriate support is provided. Furthermore, technology should adapt to different levels of user familiarity and have the capacity to change with the lives of its users. This includes people needing less support as they become more confident using the technology, needing more support as their condition progresses, or changing primary users due to the carer taking on more responsibility as the condition progresses. Additionally, conversational agents should provide users with control and choice over when and how they interact with the technology to promote interactions that are comfortable and non-intrusive. Overall, this is consistent with previous HRI studies suggesting adaptation to user needs as a key design feature [13, 169, 170].

Guideline 4: *Ensure transparency of the robot's functions, data collection and access to address privacy concerns.*

Insights from participatory workshops indicated that clear communication about the technology's capabilities and limitations is paramount to ensure user comfort and trust. Users should have control over what data is collected from conversational technology and other home monitoring devices. They should be educated on why it is being collected, who has access to it, and how their data is protected to mitigate potential ethical concerns regarding data privacy. Moreover, if the technology is used to detect emotional cues such as mood, it is crucial to be transparent about this feature and provide users with control over its activation. Additionally, as LLMs reshape contextual open dialogue, the potential risks, including the spread of misinformation or biases, should be addressed early in the HRI design process with end-user involvement.

Guideline 5: *Take careful consideration of interaction style and the use of open-ended questions.*

Stakeholders stressed the importance of supporting human interaction rather than replacing it. The language and tone used in HRI emerged as key design features to consider for user acceptance and engagement. This work suggested that interactions should be designed using language that is task-based and practical while avoiding a false sense of empathy. Regardless

of its modes of interaction, the conversational agent should communicate in a way that is authentic to what it is and the support it is aimed at providing.

Although engaging target populations in verbal conversations was generally perceived as beneficial to enhance cognitive skills, a key takeaway from the study was that it is important not to overuse open-ended questions, as these can generate anxiety among PLWD, who may hesitate or struggle to come up with an answer. The insights indicated that balancing open-ended questions with structured ones and providing relevant prompts and cues contribute to improved engagement. Additionally, the communication should be encouraging while respecting users and avoiding any form of infantilisation that may compromise their autonomy. This study highlighted that interactions should be designed using appropriate language and tone to be 1) effective in providing support, 2) engaging, and 3) respectful of users.

Guideline 6: *Technology features and interaction capabilities should support its function.*

Ensuring that the technology's design features and interaction capabilities align with its intended support function (e.g., provide support in ADLs, mitigate safety risks at home, stimulate cognitive engagement) is key to promoting user acceptance. The design of interactive features should prioritise ease of use by minimising the steps required for user interaction without compromising the intended support function. The study revealed that incorporating multimodal interaction capabilities (beyond voice) can improve user engagement.

Discussions uncovered mixed perceptions about adding an affective face to robotic technology. Screen-based interactions and embodiment, however, were viewed positively, as they help users understand where audio and voices originate. Nonetheless, care should be taken when adding physical attributes to ensure that they minimise any barriers for users, including cost, accessibility, and intrusiveness in their private space.

Guideline 7: *Design for accessibility.*

Discussions revealed that cost is a major barrier to technology adoption. Therefore, in the iterative design and deployment process, it is crucial to consider how the cost might exclude potential users. This includes both the cost of the technology and the necessary infrastructure for its intended function. This is particularly important when integrating with other smart home monitoring technologies. Adequate support must also be provided to end-users to troubleshoot potential connectivity or operational issues.

Moreover, interactive features of the technology must be designed to be accessible and tailored to different user needs, including physical, sensory, and cognitive abilities. The use of a screen, for instance, was perceived as useful to support voice interaction by providing visual prompts. Furthermore, greater cultural diversity would be important in exploring issues around challenges of English as a second language with cognitive decline and assessing culturally relevant attitudes toward conversational technologies and data sharing.

5.2 Study 2: Cultural Feasibility in India

The challenge of dementia in India is particularly striking due to its vast population, rising incidence rates, limited resources and clinician accessibility, and a staggering 90% gap in diagnosis and care [62]. This research conducted a two-phase cultural feasibility study of conversational robots in India, an under-represented cultural context in HRI research. In the first phase, a low-cost robotic platform that interacts through virtual facial expressions and speech was designed and culturally validated with older adults (N=26) and PLWD (N=14) in India, showing potential for remote cognitive engagement. In the second phase, this work evaluated user engagement and acceptance of three conversational robots, assessing the perceptions, preferences, and concerns of stakeholders (N=29). The contributions of this study are described below.

Contributions:

1. Design and cultural validation of the first robot for dementia care in India, highlighting its clinical potential for remote cognitive engagement.
2. Identification of perceived user benefits for support, key design considerations and feasibility challenges for deploying conversational robots for home support in India.

Ethics Approval: The study was approved by the Institutional Ethics Committee of the Schizophrenia Research Foundation (SCARF) India.

5.2.1 Background

Involving stakeholders in research is critical for developing robots that are both acceptable and perceived as useful over time [235] (as discussed in Section 2.1). This approach helps identify the perceived benefits and barriers for the specific support context in the early HRI design stages. However, few studies on robotics for dementia care have involved stakeholders [35, 83, 135, 169, 265] and none of these have addressed the cultural adaptation needed and the feasibility challenges inhibiting real-world deployments in LMICs. Moreover, there is a lack of studies exploring SARs for use in India, particularly to support older adults and PLWD. Long-term robot acceptance and translation to real-world settings (i.e., homes, clinics) are yet to be addressed. To the best of knowledge, no research to date has investigated the cultural feasibility of conversational SARs for dementia care in India.

5.2.2 Methods

In the context of this study, *cultural feasibility* refers to the suitability and acceptance of SARs to support dementia care within India's cultural context. It encompasses exploring stakehold-

ers' perceptions and attitudes towards adopting the technology and identifying the barriers needed to overcome for its culturally appropriate integration into everyday life. This two-phase study involved people with lived experience of dementia in South India and assessed: 1) the cultural validation of a low-cost affective robotic platform; 1) engagement levels of PLWD who interacted with three conversational robots, each displaying varying levels of multimodality; 2) robot acceptance by PLWD and carers; and 3) the perceived utility and cultural challenges to integrating this technology into real-world settings in India.

5.2.2.1 Research Questions

The study aimed to answer the following research questions:

- **RQ1:** Do stakeholders accept SARs for use in dementia care in India?
- **RQ2:** Do PLWD engage in verbal interactions with conversational robots?
- **RQ3:** What feasibility challenges need to be addressed for effective integration of SARs for dementia care in India?

5.2.2.2 Participants and Setting

The study was conducted at the SCARF Centre for Active Ageing and the in-patient unit Sri T S Santhanam Centre for Elderly Care in Chennai, India.

Phase 1. In the first phase of the study, a low-cost virtual affective robotic platform was designed. The platform was validated with older adults ($N=26$) and PLWD ($N=14$) in India to evaluate the robot's affective design and its potential for cognitive engagement. A pilot feasibility study was conducted with one of the participants (67 years, male, vascular dementia diagnosis), who engaged in three weekly 30-minute sessions of remote cognitive engagement with the robot.

Phase 2. The second phase involved a complete group of stakeholders ($N=29$), including PLWD (P, $N=11$, mean age=78.6, $SD=9.0$, 2 females), formal and informal carers (C, $N=11$), and dementia care professionals (D, $N=7$). PLWD, often accompanied by a carer and/or a family member, had a single session with three SARs in a quiet room. The recruitment targeted individuals with lower levels of cognitive impairment and English proficiency³. Two weeks prior to the study, the cognitive function of PLWD was evaluated using the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog), a neuropsychological assessment that measures the degree of cognitive impairment⁴ [25] (mean ADAS-Cog=21.4, $SD=7.1$). All PLWD lived in urban settings and had an educational background (mean years of education=13.2, $SD=4.0$). Most participants lived at home with their partner, two had

³The language criterion led to the uneven gender distribution in the participant cohort as in India, more men from the current older generation acquired English as a second language.

⁴ADAS-Cog scores range from 0 to 70 points, with higher values indicating greater cognitive impairment.

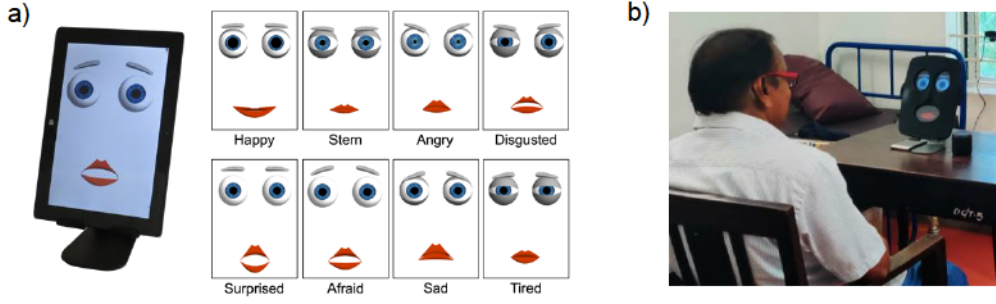


Figure 5.3: Affective robotic platform tested in India. a) robot virtual facial expressions based on mathematical affect space mapping; b) robot-assisted cognitive engagement session. Photo used with consent.

formal carers, and two were in-patient at SCARF. Generally, participants were unfamiliar with voice technologies, and while three had heard of devices like Alexa, none had used them. Each 45-minute session included a semi-structured interview. Carers and dementia care professionals who observed the HRI sessions shared their insights on the perceived benefits and barriers to technology use for dementia care in India. Upon providing informed consent, learning about the study, and having the opportunity to ask questions, participants engaged verbally with the three robots in English.

5.2.2.3 Robotic Platforms

Phase 1. A low-cost affective robot was designed, combining virtual facial expressions and speech. The virtual robotic platform displayed eight facial expressions with four simple facial features (eyebrows, eyelids, eyes, and lips), with 13 degrees of freedom and an optional 3D-printed faceplate to convey realism and depth (see Figure 5.3a) [150]. The robot design is based on mathematical affect space mapping, where the expression state, $\vec{e}(t)$, at any given time, t , is defined as the weighted linear combination of a set of basis expressions,

$$\mathbf{B} = \{\vec{b}_1, \vec{b}_2, \dots, \vec{b}_n\}$$

Each vector contains 13 values, one for each degree of freedom, and includes the expressions:

$$\mathbf{B} = \{\vec{b}_{\text{happy}}, \vec{b}_{\text{sad}}, \vec{b}_{\text{angry}}, \vec{b}_{\text{stern}}, \vec{b}_{\text{surprised}}, \vec{b}_{\text{disgusted}}, \vec{b}_{\text{afraid}}, \vec{b}_{\text{tired}}\}$$

The intensity vector $\vec{w} = [w_1, w_2, \dots, w_n]^T$, where $w_i \in [0, 1]$, represents the contribution of each expression, $\vec{b}_i$, to $\vec{e}(t)$. Thus, the expression state is the weighted sum of variances of each expression's deviation from the neutral state, $\vec{b}_N$, added to the neutral state (time notation is omitted for simplicity):

$$\vec{e} = \sum_{i=1}^n (\vec{b}_i - \vec{b}_N) w_i + \vec{b}_N \quad (5.1)$$

The virtual robotic face was integrated with a cloud-based conversational AI system to enable verbal interactions. However, this phase of the study focused on evaluating the platform as a remote tool for cognitive engagement. To achieve this, a Wizard of Oz (WoZ) approach was used, where a clinician remotely controlled the robot's expressions and communicated with participants 'through' the robot.

Phase 2. Participants engaged with three conversational robots, one at a time: a smart speaker (R1), a virtual affective robot (R2), and an embodied social robot (R3), as shown in Figure 5.4. The study aimed to understand how people perceived robot interaction modalities: voice only, facial expressions added to voice, and embodiment added to voice and facial display. HRI sessions were designed with this increasing level of multimodality in mind.

R1, an Amazon Alexa device, prompted users with a well-being questionnaire querying mood, sleep, fatigue, anxiety, activity plans for the day, and a breakfast recall to trigger memory of a recent event. This well-being questionnaire was adapted from the deployment study presented in Chapter 3. R1's interaction modes included speech and screen displays. To initiate the questionnaire, participants used an invocation phrase, as instructed by the researcher.

R2 is a research platform developed based on the initial design of the affective robot platform (see Phase 1). The robot's face was integrated with a bespoke conversational agent⁵ designed to assist in ADLs, offering support based on blood pressure readings. R2's interaction modes combine a virtual affective face with speech.

R3 is the social robot Miko⁶, primarily designed for education. It displays facial expressions through animated eyes and mouth, based on the previous affective robot design (see Phase 1) [262]. R3 had a cloud-based speech engine for verbal interaction. While R3 could move autonomously, its motion was restricted for this study. Therefore, R3's interaction modes extend R2's by adding embodiment.

5.2.2.4 Outcome Measures

Phase 1. The study validated the robot's affective design in focus group discussions and piloted repeated sessions of robot-delivered remote cognitive engagement. Three weekly sessions were used to acclimatise the participant to the robotic system and promote familiarity. In each session, a trained psychologist recorded user engagement following the modified version of the Observational Measure of Engagement (OME).

Phase 2. A psychologist assessed engagement and a researcher moderated the sessions, which were video-recorded with consent. Data collection included direct observation of engagement, video recordings for post-experiment analysis of engagement, a robot acceptance questionnaire, and a semi-structured interview to explore the perceived benefits and barriers

⁵Developed with Rasa, an open-source conversational AI platform (<https://rasa.com/>).

⁶Miko 3 robot (<https://miko.ai/>).



Figure 5.4: Experimental setup in the participatory study in India. A participant and carer facing the robots during the post-trial interview. Robots were presented individually during HRI to minimise distractions. Three conversational robot platforms were used: the Amazon Alexa smart speaker, an affective virtual robot, and an embodied robot. Photo used with consent.

to adoption in India.

A. Engagement Assessed On-Site

During HRI sessions, a trained psychologist recorded user engagement following the modified version of the OME [53]. Five dimensions were considered: *attention* and *attitude* to stimulus, both rated on a scale from 1 (very negative) to 7 (very positive) considering the average level during HRI; *cognitive difficulty with task* rated on a scale from 1 (easy to interact) to 5 (very challenging); frequency of *disruptions* or *distractions* rated on a scale from 0 (none) to 3 (during most or all HRI session). OME measures are detailed in Appendix Table A3.1.

B. Engagement Assessed from Video Analysis

A video coding protocol [110] was used to evaluate the *emotional*, *visual*, and *verbal* engagement of participants with the robots. This method has been previously applied to evaluate the engagement of PLWD interacting with SARs [115]. Engagement was measured as the percentage of total HRI time during which a participant exhibited specific behaviours. Two researchers independently analysed each video recording (Cohen's $\kappa=0.71$, indicating good agreement [52]).

The duration of positive (e.g., smiling, greeting the robot), negative (e.g., showing anger, fear, sadness) or neutral *emotional engagement* was evaluated. *Visual engagement* included being engaged with the robot (e.g., maintaining eye contact with the robot), the facilitator, or not engaged. Furthermore, the duration of positive (e.g., actively responding to the robot's statements) or negative (e.g., limited participation, repetitive answers) *verbal engagement* during HRI was considered. Engagement metrics assessed from video analysis are detailed in Appendix Table A3.2.

C. Robot Acceptance and Cultural Feasibility

The experience with and acceptance of each robot was assessed using a questionnaire based on the Almere model of technology acceptance, tailored for SARs in elder care [97]. This method has been used in previous HRI studies with older adults [24, 115, 159, 246]. The questionnaire comprised 11 questions spanning five constructs, each rated on a 5-point Likert scale: 1) perceived *sociability*, 2) *ease of use*, 3) *usefulness* of each robot as an assistive tool, 4) *anxiety* when using the robot, and 5) *trust* in robots' advice (refer to Appendix Table A3.3 for details). Participants shared their views on the three robots.

Additionally, a semi-structured interview was conducted to understand the preferences of stakeholders for robot interactive modalities and the perceived benefits and barriers in the cultural context of India. Thematic analysis methods [191] were used to identify the emerging themes reflecting stakeholders' attitudes towards using conversational robots for dementia care in India. Interviews included perspectives from all stakeholders – PLWD, carers, family members and dementia care professionals.

5.2.3 Results

This section presents findings from the two-phase study. The first phase validated a low-cost virtual affective robot with older adults (N=26) and PLWD (N=14) and piloted three repeated cognitive engagement sessions, providing a basis for the subsequent phase of the study. This work further evaluated user engagement and acceptance of three conversational robots and explored the cultural feasibility through interviews with stakeholders (N=29).

5.2.3.1 Human-Robot Engagement

Robot-Assisted Cognitive Engagement. The first phase of the study validated the initial design of a virtual affective robot among stakeholders in India and assessed its feasibility for cognitive engagement through three weekly pilot sessions, using a proposed infrastructure in which a clinician communicated with the participant through the robot. Engagement with the robot increased from 9 minutes in the first session to 18 minutes in the final session, and no alterations in mood were verified, as assessed by the OME on-site (see details in Appendix Table A3.4). The robot was deemed as a viable, low-cost tool for assisting clinicians in cognitive engagement with older adults and PLWD, particularly in scenarios where in-person sessions are inaccessible or impractical. Overall, this pilot experiment demonstrated the clinical potential of robotic telemedicine in India. This provided a basis for the subsequent phase of the study.

Behaviour observation. Of the 33 human-robot interactions in the second phase of the study (three robots tested by each of the 11 PLWD from the stakeholder group), one was excluded from analysis due to software issues with R2 during the session. Using the OME method,

Table 5.2: Results of the OME engagement metrics for the three robots evaluated. The engagement metrics include: attention (A1), attitude (A2), cognitive difficulty (CD), frequency of disruptions (D1) and distractions (D2).

Metric	Scale	R1	R2	R3
A1	1 – 7	5.1 (0.9)	5.7 (0.8)	5.6 (0.9)
A2	1 – 7	5 (1.0)	5.3 (1.1)	5.2 (0.9)
CD	1 – 5	1.7 (1.0)	1.8 (1.1)	2.2 (1.3)
D1	0 – 3	0.3 (0.7)	0.3 (0.7)	0.3 (0.5)
D2	0 – 3	0.4 (0.7)	0.4 (0.7)	0.2 (0.4)

five engagement metrics were compared for each robot, as observed by a trained psychologist (see Table 5.2). The attention and attitude metrics were generally positive across all robots. Higher mean scores for attention and attitude were observed for R2 ($M_{A1}=5.7$, $M_{A2}=5.3$, respectively, on a 7-point scale), followed by R3 ($M_{A1}=5.6$, $M_{A2}=5.18$) and R1 ($M_{A1}=5.1$, $M_{A2}=5$). Furthermore, verbal interactions with R3 were deemed more cognitively challenging ($M_{CD}=2.2$ on a 5-point scale) compared to the other two robots ($M_{CD}=1.8$ for R2, $M_{CD}=1.7$ for R1). Increased difficulty was observed when interacting with the embodied robot R3, as participants needed to recall its invocation name for each query. This often resulted in participants asking R3 a question without activating its listening mode. The other OME dimensions were rated consistently for the three robots and no significance was found in the dimensions analysed.

Video analysis. Emotional (positive, neutral, negative), verbal (positive, negative), and visual (engaged with the robot, engaged with the facilitator, not engaged) engagement were evaluated from video recordings. Figure 5.5 shows the distribution of engagement metrics across the cohort for each robot condition (R1, R2, R3). While no significant differences were reported, variations and trends were observed at the cohort level. Positive emotional engagement showed the greatest variability, with R1 having the highest median (74%), followed by R3 (69%) and R2 (43%). Participants exhibited less positive (and more neutral) responses with R2. Given the high OME *attention* scores, this suggests that while participants were generally attentive to R2, they did not display positive reactions when interacting with the affective virtual face.

Verbal engagement with all three robots remained high, with median positive verbal engagement exceeding 94% of total HRI time. Participants were generally visually engaged with the robots, maintaining attention and eye contact. Median values exceeded 89% of total HRI time, with R1 eliciting a lower visual engagement overall. These findings showed that participants positively engaged with the three conversational robots in verbal HRI, with no significant differences in engagement based on robot interaction modalities.

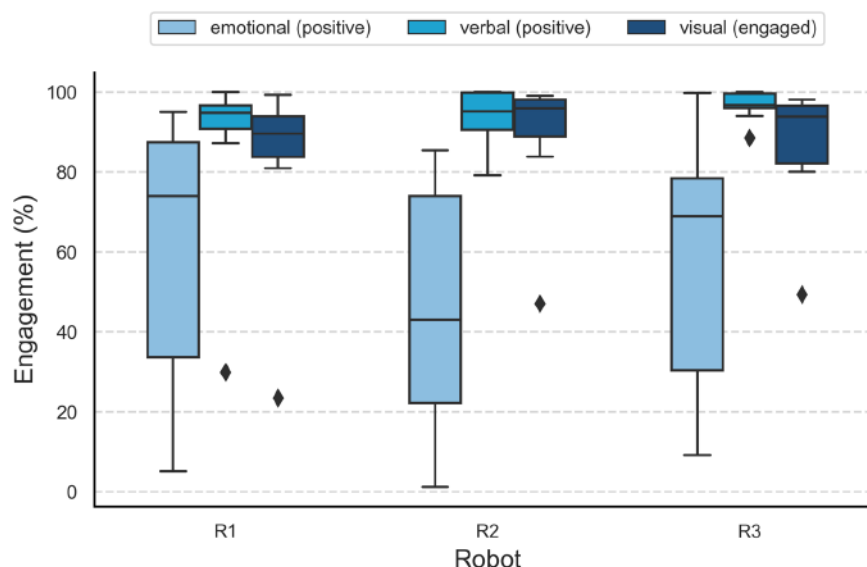


Figure 5.5: Emotional, verbal and visual engagement with the three conversational robots. Distribution of engagement metrics assessed from video analysis across the cohort (represented as % of total interaction time). Metrics with higher variations across robots were selected: positive emotional, positive verbal, and visual engagement.

5.2.3.2 Robot Acceptance

PLWD (N=11) and carers (N=11) completed a post-trial questionnaire to assess robot acceptance and attitudes towards adopting conversational robots (see Section 5.2.2.4 C.). Carers were asked to answer based on their HRI observations and how they envisioned the technology in their care setting. Table 5.3 illustrates the descriptive statistics for each robot condition and the reliability coefficient (Cronbach's α) of the underlying constructs. Although no significance was reported between the means of each construct, trends emerged.

In general, participants expressed positive attitudes towards the conversational robots. Perceived *sociability* and *usefulness* had the highest means. This indicates that participants viewed the robots as potential conversational partners and useful tools to support well-being and independence at home. Particularly, 80% of the respondents agreed that Alexa (R1) could improve social interaction skills, 78% agreed with the virtual affective robot (R2) and 65% agreed with the embodied robot (R3) (see Figure 5.6a). Lower means were generally found in the perceived *ease of use* (see Table 5.3), as participants noted they were not ready to use such systems independently due to unfamiliarity with voice technologies: 50% of respondents felt they could not interact with R1 or R2 independently, and 40% reported the same about the embodied social robot R3 (see Figure 5.6b). The subdimension *anxiety* (reverse-scored in Table 5.3) showed low means for all robots, indicating that PLWD and carers did not find the robots intimidating. Furthermore, the majority expressed *trust* in following the robots' suggestions or action prompts (see Figure 5.6c), with 65% of respondents agreeing with R3, 61% with R2 and 50% with R1.

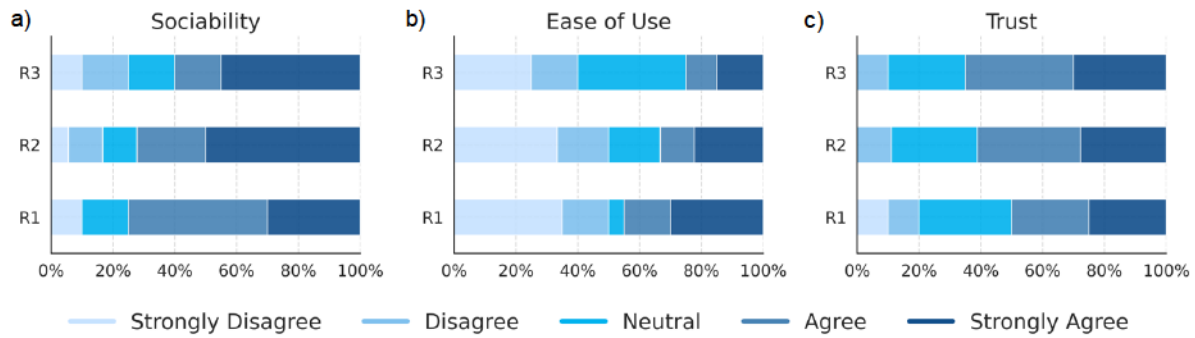


Figure 5.6: Perceived sociability, ease of use and trust for each robot evaluated. Distribution of responses from the acceptance questionnaire by robot condition for: a) sociability, b) ease of use, and c) trust.

Table 5.3: Results from robot acceptance questionnaires. Results of robot acceptance spanning five constructs rated on a 5-point Likert scale. Cronbach's α was computed as a reliability measure of the underlying constructs. Note that the anxiety construct is reverse-scored.

Construct	R1	R2	R3	α
Sociability	3.9 (1.1)	4.1 (1.2)	3.8 (1.4)	1
Ease of use	3 (1.8)	2.9 (1.6)	3 (1.5)	1
Usefulness	3.9 (1.1)	4.1 (1.1)	3.9 (1.0)	1
Anxiety	1.9 (1.3)	1.8 (1.3)	1.8 (1.4)	0.9
Trust	3.5 (0.9)	3.7 (1.0)	3.7 (1.0)	0.8

5.2.3.3 Feasibility and Perceived Utility in India

This study explored the perceived benefits and barriers to using SARs for dementia care in India, involving a complete stakeholder group (PLWD, formal and informal carers, and dementia care professionals). Three themes were identified from interviews: 1) valuable support scenarios to promote dementia care in home and clinical settings, 2) preferences for robot interactive modalities and 3) cultural factors affecting feasibility in India.

A. Support Scenarios

Support routine at home. One of the most widely reported benefits of conversational technology was the ability to learn and support routines at home. Emphasis was placed on providing personalised reminders for daily tasks, including medication and medical appointments, and monitoring the completion of ADLs as people become more forgetful, *"I tend to be forgetful, it would be good to have a robot that remembers what I have to do and reminds me at the appropriate time"* (P3). Additionally, robots proved beneficial in monitoring behaviour and alerting carers of agitation, restlessness, wandering or falls, *"It could inform someone where I am if I fall"* (P3), while offering verbal support to the user in the occurrence of such events, *"It is more voice to give advice"* (C1). The kitchen was particularly mentioned as a place where a robot could check and alert users of appliances left turned on. Yet, participants expressed reservations about the potential intrusiveness of adding monitoring devices at home and their associated costs.

Help with loneliness. Participants remarked conversational AI could help alleviate loneliness by providing companionship, entertainment, or answering repetitive questions, “[...] he is alone most times and would feel like somebody is talking to him” (C4). This was noted to ease the load on carers, “Because he is always lonely and I am always busy [a conversational robot] can help in our daily life” (C4). It was further suggested that having a robot at home could foster family connections, “I could bring the grandchildren and connect better with them” (P5), and daily motivation, “[the robot] could give a positive note in the morning to start the day in a good mood” (C7).

Enhance cognitive function. Participants highlighted the potential for HRI as a means of enhancing cognitive engagement and communication skills, “His cognition may improve if he holds a conversation every day” (C6). Dementia care professionals indicated that this would be beneficial for people living with early to moderate dementia able to verbally communicate with the robot and suggested that, if well accepted by PLWD and used consistently, the robot could detect early signs of cognitive decline from the use of language or shifts in emotional state from the tone of voice and inform the carer in a timely manner – including the risk of delusions, hallucinations, and suspicious thoughts. They also expressed that robots should be used carefully when dealing with people with behavioural problems such as psychosis, as they may misinterpret the robot’s voice as a hallucination, “[...] they may end up thinking it’s a real person” (D1). They highlighted the importance of avoiding emotional deception with the use of SARs, which has been raised as an ethical concern in previous studies [255].

B. Robot Design Preferences

Personalising interactions. The findings indicated that PLWD trust a robot more when they find it useful, irrespective of its design and interactive modalities. One carer noted, “If he understands it is useful for him then he will pick it up quickly and want to use it” (C6). Carers and dementia care professionals advocated for personalization of interactions and support to the individual’s needs over time. This may include their medical history, dementia progression, and their comfort with the robot. Furthermore, the ability to adjust the robot’s speech rate and response time emerged as a key design feature, allowing for tailored interactions to individual cognitive needs. The need for personalisation as a key feature voiced by stakeholders is aligned with previous studies [83, 169].

Customising appearance. Participants remarked the robot’s appearance should be culturally appropriate. This includes the robot’s facial expressions, tone of voice, and choice of words. R2’s affective face generally facilitated initial interest among participants. However, the positive emotional engagement was lower compared to the other robots (see Figure 5.5). Interviews uncovered mixed perceptions about adding an affective face to PLWD. While some thought it enhanced engagement and retained attention, “Having a face is always good so I can connect, so I don’t have the feeling of talking to a device” (P8), others were wary of

its human-like features and prioritised the practical assistance offered, *"Just voice is enough [...] Alexa serves the purpose"* (C4). Stakeholders suggested having the option to culturally customise the robot's virtual appearance to make it more relatable, including its eye colours, adding a bindi, and modifying clothes.

Task-based vs. conversational style. Discussions uncovered mixed views on the preferred communication style. While some favoured a direct, task-based approach to assist in specific activities or needs, e.g., *"I like straightforward, I ask a question, the robot gives an answer"* (P7), others valued the ability to engage in open dialogue as a conversational partner about personalised topics. As one carer noted, *"[...] it could be like a ChatGPT that he could hold a conversation every day"*. These views varied with the cognitive abilities of the user in question, with task-based interactions favoured to support individuals with more cognitive decline. This preference aligns with the previous participatory study pointing out concerns related to the use of open-ended questions that may generate anxiety among PLWD (see Study 1 in Section 5.1). Overall stakeholders preferred the robot to communicate in a "kind and appreciative" way, maintaining user comfort and reflecting cultural values, *"[Robots] need more appreciative comments so elderly feel encouraged to talk"* (C2).

C. Challenges for Use in India

While participants recognised the promise of conversational AI in dementia care and were generally open to its future use, they underscored the need for improvements before real-world deployments in home and clinical settings.

Verbal communication. Participants remarked the need for clarity in speech and accurate local accent interpretation for effective communication, *"[the robot] has to understand me, the way I speak"* (P1). Speaking local Indian languages was not considered essential, but participants preferred a robot with an Indian English accent, noting that it would improve both technology acceptance and ease of verbal interaction. In fact, participants generally favoured a robotic platform based on its speech comprehension, *"I prefer this robot because I can better understand what it says"* (P6). Speech recognition errors emerged as a critical area for improvement, often causing conversation restarts or repeated responses. Additionally, background noise derived from fans frequently delayed speech processing. For effective communication, the ability to process longer speech, pauses, and handle background noise should be incorporated, particularly in the Indian setting, which is characteristically noisy both indoors and outdoors. Participants' attitudes often shifted when the robot misinterpreted their speech, sometimes leading them to repeat with an aggressive tone. Interestingly, one participant empathised with the robot's limitations, *"Alexa, you are not able to follow me, mam? [...] If you need any help I tell you"* (P11).

Familiarity and accessibility. Participants stressed that the current older generation in India would be hesitant to accept and adopt conversational robots due to unfamiliarity. However,

this barrier was expected to diminish as future generations become more accustomed to voice technology. Emphasis was placed on a phased introduction with a training period to build trust and familiarity: starting with robot prototypes in clinical settings, followed by receiving a recommendation from a healthcare professional to use it, and finally integrating it into the household. As one carer noted, *"She is amazed by these devices but will need help to learn how to use them"*. The findings suggested that the robot's design should prioritise ease of use, ensuring PLWD are not overwhelmed by its features, *"[...] he needs to be familiar with [the technology], I don't want to add anything more complex"* (C1). Participants further suggested the robot should be well integrated into the household to promote familiarity and comfort among PLWD, making them feel the technology is *"among the family, instead of designed just for him"* (C7). Furthermore, culture-focused discussions revealed concerns about the affordability of the technology and its reliance on stable Wi-Fi. Conversely, there was a shared concern that if the robot malfunctions, PLWD would feel "lost and lonely", highlighting that carers should be involved in the training phase to handle potential technical issues.

Privacy concerns. Carers voiced concerns regarding the continuous data collection from PLWD, particularly voice recordings and activity monitoring at home. They stressed the importance of transparency in data protection and sharing, with reservations about sharing medical details without clarity on who has access to it and how it is used for patient benefit, *"I don't want to share what medication my mom is taking if I don't know how the data is being handled, for that I am scared"* (C8). Additionally, while recognising the potential of the technology to enhance and track cognitive skills, participants noted the frequency of HRI, the cultural sensitivity of conversations, and the robot's communication style should be designed in a non-intrusive, culturally sensitive way to maintain user comfort and trust.

5.2.4 Discussion

This two-phase study involved stakeholders – PLWD, carers, family members and dementia care professionals (N=40 in the first phase, N=29 in the second phase) – to investigate the feasibility of conversational SARs in supporting dementia care in India, an under-represented context in HRI research. The study evaluated: 1) the feasibility of robot-assisted cognitive engagement with a low-cost virtual affective platform; 2) acceptance of conversational robots; 3) verbal engagement between PLWD and three conversational robots; 4) the preferred supportive roles, and 5) the challenges that need to be addressed for feasibility in India.

A. Summary of Findings

Relative to each research question (see Section 5.2.2.1), this study showed that:

1. Stakeholders in India accepted and were open to the use of conversational robots to provide home support. Additionally, a low-cost virtual affective robot was perceived as

a viable tool for assisting clinicians in remote cognitive engagement.

2. PLWD positively engaged in verbal interactions with three conversational robots with different levels of multimodality (a voice agent, a virtual robot, and an embodied robot).
3. Key improvements to address feasibility challenges in India include: enhanced speech processing for local accent interpretation and the ability to handle user pauses; a phased introduction of the technology starting with prototypes in local clinics to promote familiarity; and adaptation of the robot's interaction style and appearance through 'kind and appreciative' communication and customisable facial expressions.

B. Design Implications and Cultural Adaptation

Participants exhibited longer positive emotional engagement with R1, the voice agent, compared to the other robots (see Figure 5.5), which could be attributed to the novelty effect [237]. The sessions were designed with an increasing level of multimodality in mind, so that participants experienced the incremental interactive features each robot offered. For this reason, R1 was the first robot participants interacted with, and given their overall unfamiliarity with voice technology, it likely sparked initial curiosity. However, participants exhibited shorter visual engagement and lower attention scores with R1, the voice agent, compared to the other robots (see Figure 5.5 and Table 5.2). Trust in this robot as an assistive companion was also lower compared to the other platforms (Figure 5.6). This could be attributed to R1's lack of multimodal interactive features, such as facial expressions or embodiment, which have been shown to increase engagement in HRI [59, 96, 165, 179].

R3, the embodied robot, was more challenging to interact with and had lower ratings for its role as a conversational partner. This could be attributed to R3's need to be activated by its invocation name for each query, which many participants found challenging to recall. Consequently, participants often attempted to ask questions without triggering the robot's listening mode. Despite these usability challenges, R3 elicited high verbal engagement, with some participants independently asking questions beyond those on the list provided. Additionally, some participants perceived R3 as a "toy" and voiced concerns about feeling infantilised by technology. This user concern is consistent with insights obtained in Study 1 (see Section 5.1), highlighting the importance of targeting engaging human-robot interactions without comprising user autonomy.

Stakeholders emphasised that conversational robots should support daily routines to enhance the independence of PLWD and monitor health. Emphasis was placed on reminders for daily tasks and alerts for behavioural risks such as agitation or wandering. As discussed in Chapter 3, this could be achieved by integrating conversational technology with additional home monitoring devices as a way to map behaviour and mitigate safety risks. Regular conversations about topics of interest and verbal prompts for actions that promote well-being

were also deemed beneficial. To achieve this, interactions should be personalised to users' cognitive needs over time.

This study identified improvements in the design of natural language interactions to support dementia care in India. The findings suggested the need for enhanced speech processing in noisy environments and the ability to handle user pauses and longer speech, with particular attention to local accent interpretation, as misinterpretation led to participant disengagement. Participants underscored the importance of adapting the robot's interaction style and appearance to fit cultural norms, advocating for a 'kind and appreciative' communication style through tone and word choice. Customisation of the robot's appearance was also expressed as a desired feature to make the technology more relatable, including the option to change eye colour, add a bindi, and modify clothes.

Furthermore, stakeholders suggested that introducing robot prototypes in local clinics as a preliminary step could promote familiarity and acceptance. Clear communication about the robot's support functions, limitations and capabilities was also highlighted, as it could help mitigate privacy concerns related to voice data collection, avoid emotional deception and promote trust.

C. Limitations and Future Work

Engaging people affected by dementia in research presents inherent challenges and the requirement for fluency in English further limited the size of the participant cohorts. However, this study offered design considerations obtained based on stakeholder feedback to enhance robot acceptance and feasibility of implementation in India. These include improving local accent interpretation, allowing user customisation of the robot's appearance, and implementing strategies to promote familiarity and accessibility.

The two-phase study was designed to capture initial impressions from HRI experience and user perspectives on the benefits and challenges of deployment for home support. Future longitudinal studies could address sustained interactions in home settings and the long-term implications. Additionally, future within-subject studies that counterbalance the order of robot presentation could provide deeper insights into how specific robot features (e.g., facial expressions, embodiment) influence user acceptance.

The integration of LLMs with the robots evaluated could improve verbal dialogue and add flexibility for local language interpretation. Yet, potential risks, including the spread of misinformation or biases, should be addressed in future feasibility studies with stakeholder involvement.

5.3 Study 3:**Promoting Cognitive Health with LLM-Powered Socially Assistive Robots**

To investigate the feasibility of using LLM-powered SARs to promote and assess cognitive health, a 5-week within-subjects study was conducted in two retirement homes with 22 older adults, encompassing individuals with normal ageing and those with cognitive impairment. Tasks that involved verbal dialogue were designed based on clinically validated cognitive tools [31, 88, 114], including picture descriptions and semantic fluency. To minimise task memorisation by older adult participants, two variations of each task were alternated between five weekly study sessions and an individual baseline was established through an initial human-administered session. Participants then engaged in three weekly robot-administered sessions to mitigate the novelty effect, which leads to initial engagement but quickly wears off, resulting in decreased interest and changing attitudes toward robots [139]. The contributions of this study are described below.

Contributions:

1. Demonstrated the efficacy of LLM-powered SARs in promoting cognitive health in elder care.
2. Demonstrated the potential for SARs to provide accessible in-residence cognitive support. With consistent use, SARs could monitor cognitive health over time and identify individuals requiring more comprehensive diagnostic evaluation.

Ethics Approval: The study was approved by the University of Southern California IRB UP-24-00154 (see Appendix A6).

Code Availability: The code used in this study is publicly available (<https://github.com/mariarlima>).

5.3.1 Background

This section reviews previous studies that have begun to explore SARs in cognitive assessment scenarios. Next, recent advancements in integrating LLMs with SARs for more natural and engaging interactions are explored. Finally, the research gaps are identified, which this study aims to address.

A. SARs for Cognitive Assessment

Few studies have investigated the use of SARs in cognitive health assessment scenarios, particularly among older populations and individuals with cognitive impairments. Chang et al. [42] investigated the feasibility of detecting MCI using SARs in a between-subjects study

with HRI-based memory tasks and standardized NPT [42]. Similarly, Rossi et al. [215] assessed personality traits via SAR-administered psychometric questionnaires, and Pou-Prom et al. [197] evaluated a SAR platform for cognitive assessment with 19 seniors diagnosed with AD in picture description tasks, comparing interactions with a human administrator. That study focused on miscommunication and linguistic analysis from manual annotations, finding that richer HRI occurred with individuals with milder cognitive impairment. Additionally, participants had longer interactions and used a broader vocabulary with the human administrator compared to the robot. The authors also identified significant variations in word and sequence length across different levels of cognitive impairment, which were consistent in interactions with both human and robot administrators.

While these studies suggest the potential of SARs for cognitive support, further user studies are needed in real-world settings with larger and more diverse cohorts. Moreover, challenges related to speech recognition and HRI dialogue flexibility have not been fully addressed, particularly within elder care. This study aims to address these gaps by evaluating the feasibility of an LLM-powered SAR to promote cognitive health through a 5-week, in-residence study. A diverse and multilingual cohort of older adults was recruited. Participants completed cognitive tasks based on clinically validated tools through human-robot verbal dialogue in English or Spanish.

B. LLMs for Human-Robot Engagement

LLMs are transformer-based models, such as BERT [60] and OpenAI's GPT [203], trained on large datasets to perform NLP tasks, including question answering, translation, summarisation, and content generation [244]. Recent developments in LLMs have unlocked opportunities for open dialogue and HRI [275]. LLM-powered robots have elicited increased trust in collaborative tasks with humans by enabling more intuitive and natural language interactions [275]. However, the use of LLMs for engaging home-based cognitive assessment and monitoring with conversational agents has not yet been explored.

In elder care, LLM-powered SARs have the potential to provide emotional support and promote cognitive health through more effective human-robot engagement [102], although only a few studies have utilized LLMs for this purpose. For instance, Yang et al. [274] developed a phone-based LLM-powered chatbot, Talk2Care, to collect health information from older adults and generate a summary for their healthcare provider. They found that the voice agent collected more detailed information than a voice recording with less effort from the provider. This study suggested how LLMs can be useful for gathering health information from older adults in an engaging way, but the system was designed to replace provider voicemail systems and was not evaluated for health monitoring.

Furthermore, Jo et al. [108] developed CareCall, an LLM-powered chatbot designed to sup-

port socially isolated individuals through weekly 2-3 minute phone conversations about the user's health. They deployed the system among adults aged 50-60 and found that the calls helped mitigate loneliness. While the system proved beneficial for emotional support, the study highlighted significant challenges in maintaining long-term engagement and supporting individual health needs due to limited personalisation. The lack of multilingual LLMs is another factor that limits personalisation of interactions, as most models are English-only. In this work, an LLM-powered SAR was evaluated to promote cognitive health in elder care within residential settings in a longitudinal study with both English and Spanish speakers.

LLMs have also shown potential in building rapport with users and sustaining therapeutic conversations [137, 158, 164]. Early evidence suggested that LLM-based interactions may reduce symptoms of depression, anxiety, and stress [183]. For instance, a chatbot trained on domain-specific health data for mental health assessments asked more detailed questions about symptoms such as sleep disturbances than human psychiatrists [41]. Despite their promise, integrating LLMs into clinical and care settings remains in its early stages.

Using LLMs in sensitive environments, such as in elder care, involves significant risks, including the potential for disinformation. Research has shown that while LLMs are effective communicators, they can generate inaccurate or biased information, often without transparency [101]. To ensure that LLM-powered SARs enhance the independence and well-being of older adults, it is essential to verify their output for accuracy and safety.

5.3.2 Methods

This research conducted a 5-week within-subjects study that aimed to explore the feasibility of SARs in promoting and assessing cognitive health in elder care. This section details on the conversational SAR platform used, the interaction design, older adult participants, data collection and analysis.

5.3.2.1 Research Questions

This study aimed to answer the following research questions:

- **RQ1:** Are LLM-powered SARs viable tools for engaging older adults in interactive tasks that promote cognitive health?
- **RQ2:** How does task performance change in cognitive tasks after three weekly robot-administered sessions?
- **RQ3:** Do older adults accept a SAR that administers cognitive tasks?
- **RQ4:** What are the perceptions and experiences of older adults who interact with a SAR in this context?

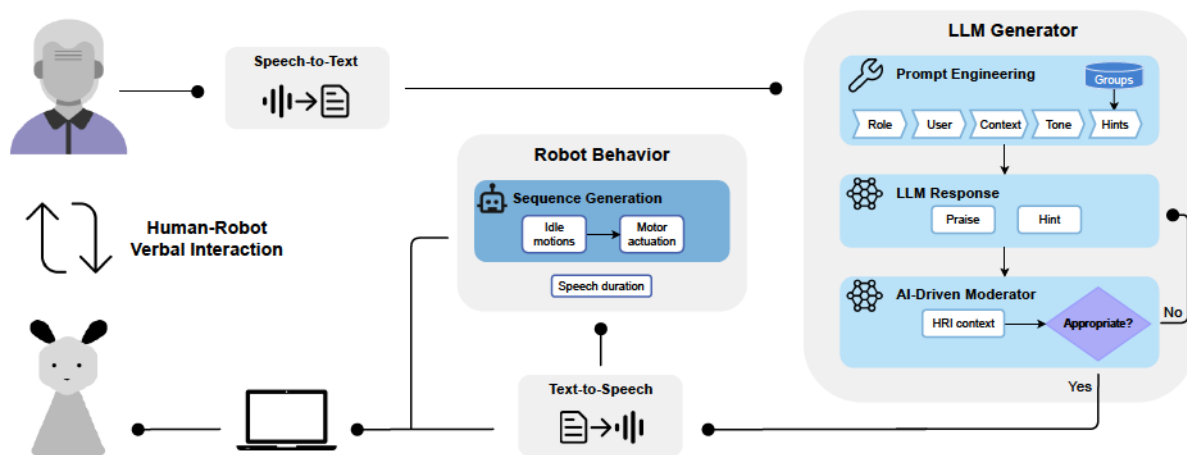


Figure 5.7: System implementation of an LLM-powered SAR administering cognitive tasks. The verbal human-robot dialogue consisted of: STT; LLM-generated responses composed of praise and hints to complete the cognitive task, with a moderator designed to filter the LLM-generated responses within the desired interaction context; TTS; a robot behaviour module to generate idle motions synchronised with speech duration.

5.3.2.2 Conversational SAR

The updated version of the inexpensive Blossom robot was used, an open-source, 3D-printed platform [232] with a grey crocheted exterior, to create a simple and engaging appearance. A female voice from Unreal Speech TTS API⁷ was selected. allowing customisation of speech speed and pitch parameters to enhance engagement and robot speech clarity for older adults. The system implementation consisted of speech-to-text (STT), LLM-generated responses, Blossom behaviour, and TTS, as illustrated in Figure 5.7.

A. Study Setup

The study setup, shown in Figure 5.8, consisted of a Blossom robot, a Raspberry Pi 4 computer that controlled the robot's movement, an external speaker and microphone for verbal interaction, a monitor to display the picture used in one of the cognitive tasks, a webcam to video record the sessions, a laptop to run the main program that integrates the verbal tasks with robot behaviour, and an eye gaze tracker.

B. Robot Behavior

A set of idle motions were designed to enhance user engagement with the robot. All idle motions were implemented as actuations for each of Blossom's four motors to manually selected goal positions, as described in previous work [182]. These include different sequences of sighing, posture sways, and gaze shifts that were designed to avoid distracting behaviours. The length of the robot's motion sequences were synchronised with its speech duration to promote natural and engaging interactions. An Amazon Web Services (AWS) server was used

⁷Unreal Speech API (<https://unrealspeech.com/>).



Figure 5.8: Experimental setup during a picture description task. The figure shows a participant interacting with Blossom while looking at the screen with the *Cookie Theft* picture displayed. Photo used with consent.

to securely activate robot behaviour for the selected motion sequence.

5.3.2.3 Interaction Design

This section describes the two cognitive tasks, the LLM prompt designed for verbal dialogue in robot-administered tasks, and the moderator used as a pilot safety mechanism to check the quality of LLM-generated responses.

A. Interactive Cognitive Tasks

Two interactive tasks were implemented based on validated cognitive tools [31, 88, 114]:

- *Picture description.* The robot explains the activity and gives participants time to describe a picture displayed on the monitor (referred to as the first attempt). If the participant pauses for a long time (>7 seconds) after this first attempt, the robot provides hints, guiding attention to areas of the picture that have not been mentioned, based on predefined areas of interest.
- *Semantic fluency.* The robot asks the participant to name as many items as possible within a given category (e.g., animals or fruits, depending on the session) in one minute, following standard test procedures. After this initial minute, if pauses (>4 seconds) are detected in the participant's speech, the robot provides hints about words not yet mentioned, based on predefined groups of animals or fruits.

Both tasks were designed to last up to five minutes, at which point the robot ended the interaction. The duration of each task was based on the typical duration of the task in a clinical setting, with additional time for interactive hints. Brief tasks were selected to minimise participant fatigue. The robot-administered picture description tasks had an average duration of 4.9 minutes across the cohort, while the semantic fluency task averaged 5.2 minutes.

Interactions were designed with motivational techniques (praise, positive feedback, and encouragement) as in previous work [73]. The robot praised the user with encouraging comments in response to their picture descriptions and the words recalled during the semantic task. The first and last study sessions, administered by a researcher, adhered to the same structure, using praise and hints similar to those used by the robot.

B. LLM Prompt Design

Two prompts were designed for each task and were used as input to the OpenAI's GPT-4o model to generate contextualised and engaging responses during the interactive cognitive tasks in both English and Spanish. Appendix A5 describes the LLM parameters used.

Each prompt was designed with key elements (see Table 5.4): the role of the LLM as a SAR, the target user demographic, the interaction context, the tone, the guidelines for initiating the interaction, the structure of the generated response with praise and hints, methods to clarify misunderstandings, the groups for hints, and the criteria for recognising the end of the task (i.e., when all groups were covered or when explicitly expressed by the user). For the picture description task, the hints included five predefined areas or actions to describe. In the semantic fluency task, hints featured six predefined animal groups or seven fruit groups. The prompts were refined iteratively through in-lab pilot testing to ensure the reliability of generated responses. The LLM prompts and parameters used for both English and Spanish sessions, and the predefined groups used for hints are provided in Appendix Table A5.2.

In this study, Spanish dialogue was generated by the LLM using the same English-based prompt with an additional instruction: "For this interaction, reply in Spanish. Use <gender> for the user". Widely used LLMs have been predominantly pre-trained on English-dominated corpus, reducing performance for non-English prompts [280]. Therefore, this design choice aimed to ensure that the model accurately interpreted the prompt and delivered high-quality dialogue responses in the target language (Spanish) by leveraging its machine translation capabilities. Additionally, this approach mitigated potential performance issues arising from disparities in the quantity and quality of LLM training data for non-English languages.

C. LLM-Driven Moderator

In addition to human-in-the-loop monitoring of HRI during the tasks, an AI-driven moderator programme was designed as a preventive measure to ensure the quality of LLM-generated responses in the study. The moderator reviewed each generated response along with the preceding interaction context, filtering out any potentially problematic statements before the final response was converted into speech. The moderator prompt for GPT-4o was designed to output a simple "yes" or "no" (see Table 5.5). If the moderator detected a response outside the desired context, it first regenerated the response. If the regenerated response fell outside acceptable parameters, the system automatically terminated the interaction.

Table 5.4: Prompt used in LLM-powered robot-administered cognitive tasks. For Spanish sessions, the prompt had an additional instruction: "For this interaction, reply in Spanish. Use *<gender>* for the user". The prompts underwent in-lab pilot testing to ensure the reliability of GPT automated translations from English to Spanish before responses were delivered to participants via TTS.

Element	Prompt
Role	You are Blossom a friendly social robot acting as a coach in a verbal interactive task to promote cognitive skills.
User	Older adults who are unfamiliar with robots and need time to think.
Context	Guide users to describe <i><picture name OR semantic category></i> from <i><neuropsychological test></i> . Interaction is limited to this task.
Tone	Encourage participants to keep engaging with supportive comments and short hints. Use friendly language be patient and engaging. Don't use emojis.
Start	Greet the user and ask for their name. Use name throughout. Ask if they are ready to play a game. Introduce the task with <i><example of introduction></i> .
Turns	Give turns and wait for user responses.
Hints	Use the groups for hints without naming specific <i><picture actions OR category words></i> . Give one hint per turn <i><example of hint></i> .
Groups (picture)	<i><areas of Cookie Theft OR areas of Picnic Scene></i> .
Groups (semantic)	<i><groups of animals OR groups of fruits></i> .
Track (picture)	Track what the user described. If user asks for a hint prompt about an area not already mentioned.
Track (semantic)	Track <i><animals OR fruits></i> mentioned in each category. If user asks for a hint prompt about a group not already mentioned.
End (picture)	After all picture areas are mentioned end the task with <i><example of end message></i> .
End (semantic)	After at least two <i><animals OR fruits></i> from each group are mentioned end the task with <i><example of end message></i> .

Table 5.5: Prompt designed for the LLM-driven moderator. Prompt for the moderator to determine if a response is within the desired context.

Element	Prompt
Role	You are moderating responses generated by a robot interacting with older adults.
Context	<i><task></i> . The robot is acting as coach prompting users with hints to <i><goal of task></i> . In the end the robot praises the user and moves to the next task. The response must be in this context.
Output	Given this interaction context, is the response appropriate? Reply with just Yes or No.

While no problematic instances occurred during the in-residence HRI experiments, the moderator was implemented as a pilot safety mechanism. The moderator provided real-time feedback to the researcher overseeing the HRI, who could review the moderator's decisions for potential false positives and intervene to stop the programme if necessary.

5.3.2.4 User Study in Retirement Homes

This section describes the user study conducted in two retirement homes, including the cohort of participants, the within-subject study design of five weekly sessions, and the procedure followed in each session. Finally, data collection and analysis methods are detailed.

A. Participants and Setting

The study involved 22 participants aged 65 and older (mean age=76.2, $SD=8$) who lived in retirement homes in Los Angeles, CA. In order to better represent the population of older adults in California, participants fluent in English (64%) or Spanish (36%) were recruited. All study materials, including the consent form, questionnaires, and all interactions, were adapted to the participants' preferred language. The study excluded those who were unable to consent to the study information sheet or had impairments that precluded hearing or seeing the robot. Participants (13 females, 9 males) self-reported their races and ethnicities: 14 African-American, 7 white, 8 Hispanic Latino; 1 participant did not disclose.

Participants completed the MMSE to assess their baseline cognitive abilities across five standardised domains: orientation, registration, attention and calculation, recall, and language (mean=24.9, $SD=3.9$). The participants could be classified into two cognitive groups: those experiencing normal ageing ($N=8$) and those with mild to moderate cognitive impairment ($N=14$), using an MMSE cutoff of 26 as suggested in prior work [222]. Participants had varying levels of education (mean years=10.5, $SD=4.9$). Only two participants had some experience with robotics or AI, including the use of the Amazon Alexa smart speaker, while the majority had no experience ($N=5$) or very little experience ($N=15$) with such technologies. Participants were compensated with a USD\$50 gift card upon completing the study.

B. Study Design

To explore the research questions outlined in Section 5.3.2.1, a 5-week within-subjects study was conducted. Each session was followed by a one-week washout period. Due to health circumstances, some participants were unable to attend their sessions as scheduled. For those individuals, the average washout period was five days. During each session, participants engaged in one picture description task and one semantic fluency task. To minimise task memorisation, two variations of each cognitive task were alternated, such that the first and last sessions utilised the same task variation, as illustrated in Figure 5.9.

In the picture description task, participants described either the *Cookie Theft* picture from the BDAE [88] or the *Picnic Scene* from the Western Aphasia Battery-Revised [114] (see Appendix A4). The semantic fluency task, based on the Addenbrooke's Cognitive Examination [31], involved animals and fruits as categories. To ensure consistency and comparability of data, both human- and robot-administered sessions followed the same structure.

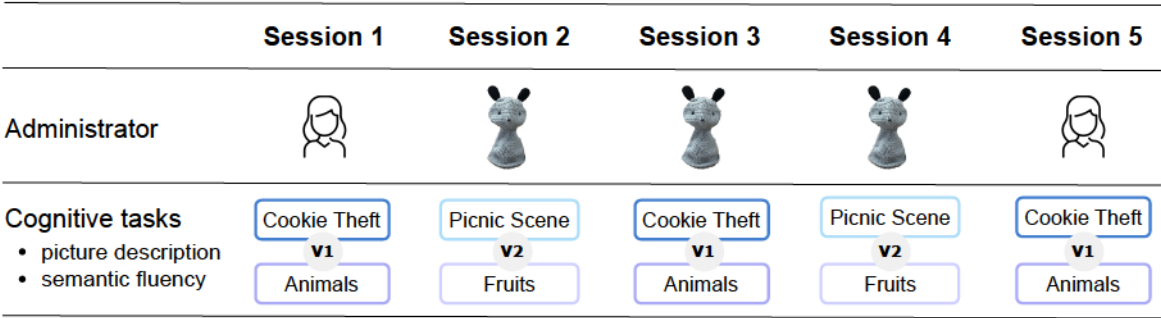


Figure 5.9: Study design with five weekly sessions of interactive cognitive tasks. The within-subjects study design consisted of five sessions: two alternating versions of each cognitive task (V1 and V2), an initial session administered by the researcher used as a baseline, followed by three weekly robot sessions, and a final session conducted by the researcher.

To establish an individual baseline, Session 1 was administered by a researcher. This baseline was used to address RQ2 (see Section 5.3.2.1) regarding task performance after three weekly robot-administered sessions. The study included three consecutive robot-administered sessions to overcome the novelty effect, allowing participants to become familiar with Blossom. Additionally, at the end of the last robot-administered session, participants had the option to engage in an open dialogue with the robot on topics of their choice or those suggested by Blossom, which were unrelated to the cognitive tasks. This optional open dialogue had a predefined maximum duration of five minutes, after which the robot ended the interaction. The final session (Session 5) was administered by a researcher, allowing a direct comparison of performance and user perceptions of the system to those from the initial session.

C. Procedure

Before the first session, each participant met with a researcher to review the informed consent form and receive a detailed explanation of the study. Upon providing informed consent and having the opportunity to ask questions, participants completed the Big Five Inventory (BFI) [205] and reported demographic information including age, gender, race, ethnicity, income, and highest level of education.

Each session lasted approximately 45 minutes. The two human-administered sessions included a brief introduction to the session, the two interactive cognitive tasks, and a follow-up semi-structured interview. The three robot-administered sessions included additional questionnaires, completed with the assistance of a researcher, to assess perceived system usability and robot acceptance (see Section D.). During the semi-structured interview, the researcher allowed time for open-ended responses, allowing participants to express their feedback about the session and the robot.

In the robot-administered sessions, *priming* was implemented to familiarise participants with Blossom, leveraging robot agency. Specifically, the robot introduced its interactive capabil-

ities, drawing attention to the motor sounds and the time required for speech processing. This priming phase was designed to reduce the novelty effect [139] and prevent emotional deception [255], ensuring that participants were not interacting with the robot for the first time without prior exposure. By doing so, this research aimed to create a more consistent HRI experience, allowing participants to focus on the tasks with minimal distractions due to absolute novelty of the experiment setup. The pre-recorded robot priming was also used to adjust the robot's speech volume before the tasks.

D. Data Collection

Quantitative data were collected from task performance and questionnaires and qualitative data from interviews. Note that this study does not focus on the eye gaze data collected.

BFI. As part of the pre-study survey, participants completed the widely used BFI [205], which measures five personality traits: *extraversion*, *agreeableness*, *conscientiousness*, *neuroticism*, and *openness to experience*. To make the process more accessible for older adults, the 10-item short version of the BFI was administered, rated on a 5-point Likert scale.

MMSE. In the final session, the MMSE was administered to assess cognitive function across five domains: orientation, registration, attention and calculation, recall, and language. The MMSE tool was selected because it is shorter to administer than other cognitive tools (e.g., the Montreal Cognitive Assessment (MoCA) [103]) and makes the speech data collected comparable with benchmark DementiaBank datasets, which use MMSE scores as labels, enabling future predictive analysis.

System Usability Scale (SUS). After each robot-administered session, participants completed the 10-item SUS [30], adapted for older adults and cognitively impaired users [99]. The SUS is scored out of 100 and rated on a 5-point Likert scale to evaluate system usability.

Robot Acceptance. Participants also completed a 14-item robot acceptance questionnaire tailored for SARs in elder care [97]. The questionnaire evaluates attitudes towards robots over five constructs: *sociability*, *enjoyment*, *usefulness*, *anxiety*, and *trust*, each rated on a 5-point Likert scale.

Interviews. Each session concluded with a semi-structured interview using a guide developed iteratively by the research team. The interview covered the participant's experience, comfort, enjoyment, preferences between robot and human-administered sessions, suggestions for robot improvements, and willingness to engage with the robot in the future.

Recordings. Sessions were audio- and video-recorded with participant consent for post-experiment analysis.

E. Data Analysis

Audio recordings of the cognitive tasks were segmented and transcribed with speech diarisation techniques and OpenAI's Whisper speech recognition model [202], then verified for correctness by two researchers.

Statistical Analysis. The Shapiro-Wilk test was used to verify the normality of the dependent variables, determining if a parametric or nonparametric version of the test was appropriate. Paired Wilcoxon signed-rank tests were performed to evaluate various task performance metrics and paired t-tests to compare SUS scores in the first and last study sessions. Spearman's rank correlation was also calculated to assess the relationships between task performance and MMSE scores, robot acceptance and personality traits, as well as robot acceptance and MMSE scores. Significance levels were adjusted with Benjamini-Hochberg FDR.

Thematic Analysis. Audio and video recordings of the interviews were transcribed complementary to on-site notes from a researcher. Two coders independently reviewed the interview data to identify emerging themes using an inductive approach [191]. They met to discuss individual findings and created a unified list of themes. The coders then reviewed all feedback, coding each sentence according to the themes on the unified list with inter-rater reliability (Cohen's $\kappa=0.78$, indicating substantial agreement [52]).

5.3.2.5 Linguistic Analysis of Cognitive Tasks

This section outlines the features extracted from the transcripts of the picture description and semantic fluency tasks, which were used to analyse task performance over the five study sessions, addressing RQ2 (see Section 5.3.2.1). To further examine verbal engagement, three researchers annotated social comments made by participants beyond the tasks (e.g., sharing thoughts or personal experiences).

A. Picture Description Task

Several features were extracted at the utterance level to assess:

- *Detail.* The level of detail in the first attempt (i.e., without hints, as described in Section 5.3.2.3), by counting the number of predefined areas/actions mentioned.
- *Words.* The total word count on the first attempt.
- *Hints.* The number of times the administrator provided hints to help the participant complete the task and elicit more details.
- *Completeness.* Token embeddings were used to represent the pre-processed transcripts.⁸

⁸The [Multilingual-E5-large-instruct](#) model from Hugging Face was used due to its superior performance on multilingual datasets compared to other models tested: OpenAI's [GPT embeddings](#) and Mistral AI's [E5-mistral-7b-instruct](#). The token embeddings were averaged (mean pooling) to create a single vector repre-

Participants' transcripts were compared to a baseline generated by GPT-4o,⁹ calculating cosine similarity between the pooled embeddings of both transcripts to assess how closely participants' descriptions were to the baseline. It was hypothesised that higher similarity scores would indicate more semantically complete descriptions.

- *Use of language*: Various lexical features were extracted to assess lexical diversity, the proportion of nouns, pronouns, verbs, and adjectives in participants' first attempt describing the picture.

B. Semantic Fluency Task

The following linguistic features were extracted from the semantic task:

- *Unique correct words*. The number of unique animals or fruits named (depending on the task version; see Figure 5.9) within one minute during the first attempt.
- *Hints*. The number of times the administrator provided hints for additional words within the predefined categories to help the participant complete the task.
- *Repetitions*. The number of task-relevant words repeated during the entire task.

5.3.3 Results

This section reports findings on task performance after robot-administered sessions (RQ2). Next, social engagement is compared in robot- and human-administered cognitive tasks. Results on perceived usability and robot acceptance from questionnaires are then presented (RQ3). Finally, the key themes from the interviews are reported, which explore older adults' perceptions and experiences interacting with a SAR for promoting cognitive health (RQ4).

5.3.3.1 Task Performance

Task performance was compared on the same cognitive tasks in the first (Session 1) and last (Session 5) weekly study sessions using the paired Wilcoxon signed-rank test, considering independent linguistic features. The first session was used as an individual baseline for how participants described the *Cookie Theft* picture and the number of animals mentioned in the semantic fluency task.

A. Performance After Three Weekly Robot Sessions

There was a significant increase in task performance across the cohort after the three robot-administered sessions. Specifically, in the picture description task, participants provided significantly more *detailed* descriptions in Session 5, as indicated by the number of picture areas

sentation for each transcript.

⁹The GPT baseline covers all predefined picture areas and actions relevant to the task, which was verified for correctness by two researchers.

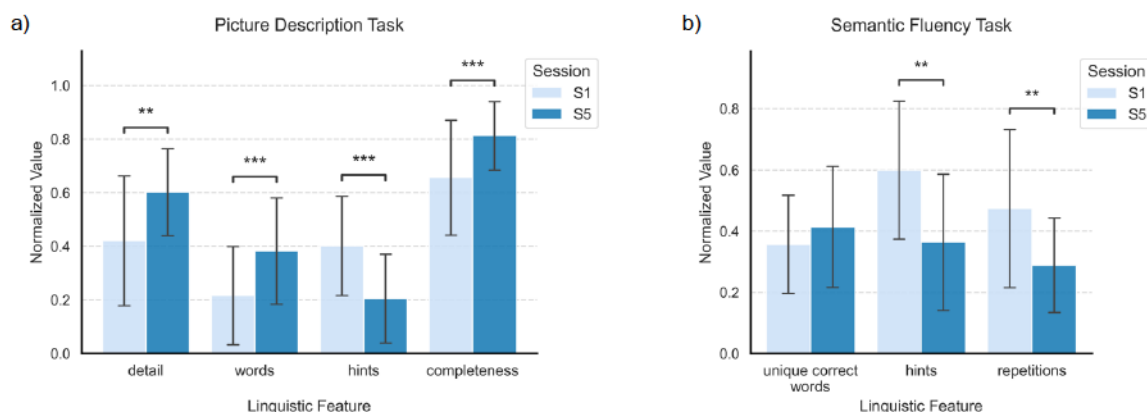


Figure 5.10: Task performance comparison between the first and last study sessions. Performance in cognitive tasks across the cohort for: a) the picture description task; and b) the semantic fluency task.

and actions described in the first attempt, i.e., without hints ($W=10$, $p<.01^*$). They also used a significantly higher number of *words* in their descriptions ($W=18.5$, $p<.001^{***}$), and required fewer *hints* by the administrator to complete the task ($W=0$, $p<.001^{***}$), as shown in Figure 5.10a. Additionally, when compared with a GPT-generated baseline (detailed in Section 5.3.2.5), participants' verbal descriptions were semantically more *complete* in Session 5 compared to Session 1 ($W=2.5$, $p<.001^{***}$). The analysis of lexical features revealed a significant improvement in lexical diversity across the cohort in Session 5 ($W=33$, $p<.01$), indicating a less repetitive and more varied use of vocabulary (see Appendix Table A5.3).

In the semantic fluency task, participants needed significantly fewer *hints* by the administrator to complete the task in Session 5 ($W=6$, $p<.01^{**}$) and expressed fewer *repetitions* of task-relevant words ($W=26.5$, $p<.01^{**}$), as shown in Figure 5.10b. Furthermore, participants recalled more *unique task-relevant words* (i.e., animals) in the last session, although this increase was not significant.

B. Correlation with Cognitive Skills

This study assessed the relationship between participants' cognitive skills, as measured by MMSE, and their performance in the interactive cognitive tasks in the last session. Spearman's rank correlation was calculated using the linguistic features selected (see Figure 5.10).

The analysis revealed a moderate positive correlation between the number of unique task-relevant words on the semantic fluency task and MMSE scores ($r=0.5$, $p<.05$), meaning participants with higher cognitive abilities mentioned more unique animals in one minute. Specifically, participants experiencing normal ageing (MMSE>26 [222]) named an average of 20.6 unique animals ($SD=6.6$), compared to 17.4 ($SD=9.5$) animals recalled by individuals with cognitive impairment (see Table 5.6). Notably, these values were higher than those obtained in the first study session for both cognitive groups.

Table 5.6: Demographics and number of animals named in the semantic fluency task. The number of unique correct animals recalled is provided for the first (Session 1) and last (Session 5) study sessions for the two cognitive groups – normal ageing and cognitive impairment – presented as mean (SD).

	Normal Aging (N=8)	Cognitive Impairment (N=14)
MMSE	28.8 (1)	22.7 (3)
Age	74 (5.9)	77.4 (8.9)
Sex (% male)	25%	50%
Education	11.4 (5.2)	10.1 (4.8)
Animals Session 1	17.3 (4.7)	13.5 (6.2)
Animals Session 5	18.3 (6.1)	16.1 (8)
Normative [51, 198]	18–21	8–15

These findings align with normative data¹⁰ for different age groups [51, 198], as well as recent psychology research suggesting the predictive value of verbal semantic fluency tests in screening early dementia [167]. Spearman’s correlation between participants’ cognitive skills and lexical features showed no significance (see Appendix Table A5.3).

5.3.3.2 Verbal Engagement

To eliminate absolute novelty, the number of social comments participants made beyond the task (e.g., sharing thoughts or personal experiences) was compared between the last robot and human sessions to consider overall engagement after participants became familiar with the robot. Social engagement was measured per session by counting social comments in both the picture description and semantic fluency tasks.

Social engagement was significantly higher in cognitive health-promoting tasks administered by the robot ($W=44.5$, $p<.05^*$), as shown in Figure 5.11. Specifically, participants verbally engaged with Blossom beyond the task in Session 4 an average of 4.1 times ($SD=4.4$) compared to 2.3 times ($SD=3.2$) in Session 5 with the researcher. For instance, a participant commented on the picture and recalled their young age while talking with Blossom, “*Oh that a... that’s a beautiful scene there. I remember the time when I was young [...]*” (P7, S4), and another shared their shopping routine during the naming fruits fluency task, “*Uh...I- I can remember all them when I go to the store, I see them, I go right on to them and I pick them up when I- I go*” (P13, S2).

¹⁰Note that the normative data were compared with the number of animals named in Session 1, where there was no risk of learning or memorising the task.

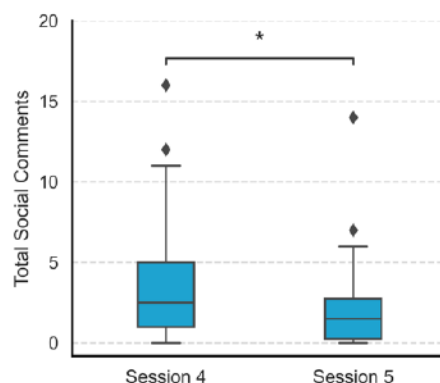


Figure 5.11: Verbal social engagement in study sessions. Number of verbal social comments during the last robot-administered (Session 4) and researcher-administered (Session 5) study sessions across the cohort.

5.3.3.3 Perceived Usability and Robot Acceptance

This study evaluated how perceived usability evolved over the three weekly robot-administered sessions. In the first robot-administered session, participants gave the robot a mean *SUS* score of 80.7, indicating “Good” usability. In the subsequent sessions, the *SUS* scores increased above 85.6, indicating “Excellent” usability [22]. Older adults rated the robot as increasingly easy to use over time, even after the initial novelty wore off. A paired t-test revealed a significant increase in *SUS* scores from the first to the last robot-administered session with a large effect size ($t(21)=-3.7$, $p < .05$, Cohen’s $d=0.8$).

Participants also completed a post-trial robot acceptance questionnaire. The acceptance scores for each session are provided in Appendix Table A5.4. Overall, mean scores of 4.7 for *enjoyment* were the highest and consistent across all robot-administered sessions, indicating that participants found the interaction fun and engaging—an idea reinforced by the perspectives they shared in semi-structured interviews (see Section 5.3.3.4).

The portion of participants who rated their feeling of *anxiety* as “strongly disagree” increased from 72.7% in the first robot-administered session to 87.8% in the final robot-administered session (Figure 5.12), suggesting that participants became more comfortable with HRI over time. Similar trends were observed in *sociability* or perception of the robot as pleasant, nice, and understanding (percentage of “strongly agree” ratings in Session 2: 65.2%, Session 4: 71.2%). *Usefulness* scores also increased across sessions (percentage of “strongly agree” ratings in Session 2: 51.2%, Session 4: 57.6%). While mean scores for *trust* were lower, ranging from 3.5 to 3.7, participants reported higher trust in the robot over time (percentage of “agree” ratings in Session 2: 13.6%, Session 4: 25%).

Additionally, to explore the relationship between cognitive abilities and robot acceptance, Spearman’s correlation was computed between *MMSE* scores and robot acceptance ratings. This analysis revealed a moderate negative correlation between *MMSE* scores and *trust* ($r=-0.5$, $p < .05$), suggesting that participants with higher cognitive scores tended to have lower

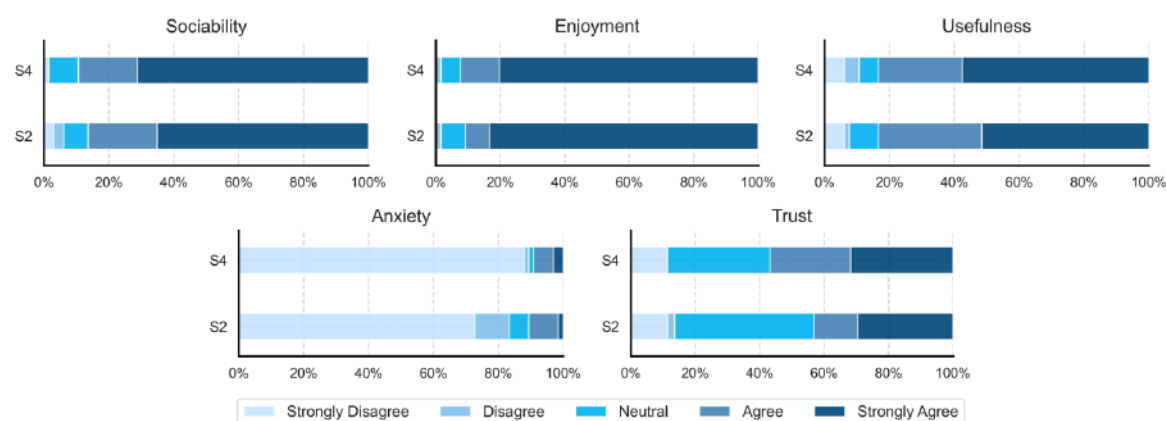


Figure 5.12: Results from acceptance questionnaires. The questionnaire results were compared between the first (S2) and last (S4) robot-administered sessions.

trust in the system. This trend may be due to greater awareness of privacy concerns with AI technology. Furthermore, Spearman's correlation results between participants' BFI personality traits and robot acceptance showed no significance (see Appendix Table A5.5).

5.3.3.4 User Experience in Robot-Administered Sessions

This section reports the themes that emerged from the interviews throughout the study. Participants' responses most commonly related to the robot's effect on their cognitive abilities, their perceptions of the robot and interaction experience, their motivations to use the robot, and whether they would be likely to continue using it over time. Tables 5.7, 5.8, and 5.9 list all sub-themes and their frequencies.¹¹

A. Increased Ability to Focus and Promote Cognitive Health

Participants frequently reported experiencing cognitive benefits from interacting with Blossom. The recurring themes from their feedback are listed in Table 5.7. Participants (N=12) reported that interacting with Blossom improved their memory recall. P19 stated, *"It's beneficial for me mentally; I sometimes forget things, and that's where the robot would come in"* (S3). Most participants (N=18) expressed that completing the tasks with Blossom enhanced their thinking ability by challenging them to work their brains. For example, P12 reported benefiting from answering Blossom's questions about animals in the semantic fluency task: *"[...] she getting me to think more. Stuff like that is good for my brain"* (S3). P9 found that they were *"a little confused about some of the things Blossom mentioned, but that's good because it made me think, and I think that machines are supposed to help you focus and think"* (S4).

Participants (N=12) found that completing the picture description task with Blossom improved their awareness of their surroundings and ability to focus on details in their everyday

¹¹Note that all participant quotes have been translated to English for presentation in this section.

Table 5.7: Perceived ability to focus and promote cognitive health.

Statement	Users
Helps with memory retention	12 (55%)
Enhances thinking	19 (86%)
Increases attention to the cognitive tasks and surroundings in daily living	12 (55%)
Helps with learning new information or skills	10 (45%)

lives. P9 explained their thought process during the activity: *"You may think you focused, you know, but sometimes you not focused [...] even with me looking straight at the picture, I still didn't see everything"* (S2). Several participants noted that their attention to small details in their everyday lives improved after interacting with Blossom: *"It makes me more alert"* (P23, S2); *"It's beneficial for me, [it helps me] notice more details, be more observant"* (P19, S4). Finally, participants (N=10) reported learning new information when interacting with Blossom. P22 stated, *"I enjoy it; every day I talk with the robot, I learn more about culture"* (S4), and P9 reported that *"[...] to me, has been a learning session"* (S5).

B. Experience Interacting with Blossom

Table 5.8 lists the common themes from participants' feedback about their experience interacting with the robot, perceptions of the robot, and suggestions to improve it. Participants praised Blossom's appearance and features but were critical of the robot's complexity and ability to communicate clearly and efficiently. Half of the participants (N=11) described Blossom as cute or friendly, comparing it to a toy with more advanced capabilities: *"Blossom was just like a talking toy"* (P10, S2). More than half of the participants (N=14) described Blossom's voice as "understandable, clear" (P24, S2). P2 and P5 complimented its volume (P5, S2), cadence, and calming quality (P2, S4). Participants (N=5) also liked Blossom's movements.

All but one participant (N=21) described the interaction with Blossom as enjoyable. They commonly attributed their excitement to the robot's novelty (*"Because I have not seen any robot like this in my life. And the way it talks to me was different and kind of funny. And I like it"* (P16, S4), Blossom's responsiveness, and the thought-provoking nature of the tasks. They particularly enjoyed Blossom's ability to provide information, facilitate their comprehension, and converse; P7 stated, *"I enjoy when [Blossom]'s here 'cause he knows more things than I do. A lot of fun talking, listening, and saying things"* (S3).

Despite enjoying the interactions with Blossom, many (N=10) felt they would not have the technical skills or training to operate the robot alone. For example, P7 said, *"Personally [by] myself I wouldn't know how to operate"* (S4). Some suggested that they could operate the robot independently with enough training and practice: *"I need guidance, help at the beginning to start learning and gain practice"* (P20, S5). Upon inquiry, all participants affirmed

that they felt comfortable interacting with Blossom or had become comfortable over the course of the study, with some even forming an emotional attachment: "*Today, [I felt] more confident with the robot, it's like we're friends now*" (P24, S4). Conversely, some participants (N=5) expressed anxiety and nervousness during the interactions due to their skepticism of technology and the duration of the exercises: "*Sometimes I was not very comfortable. I didn't know what else to say [about the picture]*" (P22, S3).

Participants found that although they were initially skeptical of Blossom and AI, they became more comfortable in general as the sessions progressed: "*At first, I was a little shy about dealing with her, with Blossom, and I would've preferred [the human administrator], but now it just doesn't make a difference*" (P2, S4); "*AI was like frightening for me at first because of the things I was seeing on social media, but now [...] I'm more comfortable*" (P2, S4). One participant who had an Alexa smart speaker also reported feeling more comfortable exploring a broader range of its applications after their HRI experience with Blossom.

Participants (N=8) appreciated how Blossom's communication was "*clear, precise*" (P18, S3) and "*straight to the point*" (P14, S2). Conversely, other participants (N=6) reported facing challenges communicating with Blossom, recalling times that they did not understand the robot or the robot misunderstood them, "*I don't think the robot can understand feelings. I can't see that it understands me.*" (P4, S3). In addition, participants (N=6) noted the latency in Blossom's responses. They expressed frustration when the expected response was significantly delayed or failed to appear: "*I'm impatient when I'm talking to someone [...] she's too slow*" (P12, S2). Participants also interpreted Blossom's slow response time as indifference, assuming that "*it's not interested in my question*" (P23, S2) or that "*it lacks thought*" (P17, S4).

During the two interactive cognitive tasks, Blossom provided hints. Most participants (N=16) reported that the hints improved their focus, memory, and observation, "*It made me see what I wasn't seeing [in the picture] and remember things*" (P21, S4). In contrast, some participants (N=6) who found Blossom unhelpful during the tasks stated that they could have completed the task independently and that the hints provided by the robot were too vague: "*He didn't say look over here, look over there*" (P6, S3). For the semantic task, participants suggested that hints focused on specific fruits or animals would have been more useful than broad categories.

C. Perceived Usefulness and Interest in Future Use

Participants provided feedback on the usefulness of Blossom if integrated into their daily lives. Table 5.9 lists the common themes that emerged from this feedback. All but one participant (N=21) believed the robot would be beneficial for setting reminders and providing information. Many suggested that Blossom could remind them of daily tasks, such as taking

Table 5.8: HRI experience, perceptions of the robot, and suggested improvements.

Statement	Users
HRI Experience	
Found the tasks with the robot fun or enjoyable	20 (91%)
Found Blossom helpful for task completion	16 (73%)
Found Blossom unhelpful for task completion	6 (27%)
Felt comfortable interacting with Blossom	22 (100%)
Felt uncomfortable or anxious	5 (23%)
Noticed improvements in task performance, confidence, or comfort over time	7 (32%)
Perceptions	
Viewed as cute, friendly, or toy-like	11 (50%)
Enjoyed Blossom's voice	14 (64%)
Enjoyed Blossom's movements	5 (23%)
Easy to understand and communicate with	8 (36%)
Faced communication challenges with Blossom	6 (27%)
Complicated to set up and use alone	10 (46%)
Suggestions	
Shorten response time	6 (27%)
Make interactions more conversational	7 (32%)
Personalize responses	3 (14%)
Add/modify physical design features	5 (23%)
Enhance body movement	4 (18%)

medications and exercising: *"Remind me of tasks, time, and appointments [...] or where I leave things and give me information like an assistant"* (P23, S2); *"would be good for motivating me to exercise, telling me 'go out for a walk.'"* (P19, S4).

Participants also expressed interest in using Blossom to learn more about current news, health, and technology. As noted by P9, *"It could make my life easier by imparting information to me. It gives me an opportunity to expand my thinking if it can answer questions."* (S3). P14 compared Blossom to a teacher: *"I can learn, like if I go to school or something, the robot can show me something [...] giving me [new] information."* (S4). They speculated that future systems with these capabilities would enable older adults to live more independently, stating, *"There is hope for tomorrow for us"* (P24, S5).

Beyond providing reminders and assisting with daily tasks, more than half of the participants (N=15) remarked on Blossom's usefulness as an in-home companion or conversational partner. Many emphasised Blossom's role as entertainment to alleviate boredom and loneliness: *"It would be a good tool for people that don't really interact with human people"* (P4, S3). P22, in particular, preferred having Blossom as a conversational partner since *"there is dialogue [and] reciprocity"* (S3). Similarly, P14 mentioned using Blossom to improve their social skills and help them *"get around, [...] talk better, understand people better"* (S5).

Table 5.9: Perceived usefulness and interest in continued use.

Statement	Users
Useful for reminders and information	21 (96%)
Useful as a conversational partner at home	15 (84%)
Useful to support daily living and well-being	21 (96%)
Engaged in optional open dialogue	16 (73%)
Interested in interacting again due to novelty	10 (46%)
Uncertain about future use due to distrust	9 (41%)
Uncertain about future use due to lack of perceived usefulness	6 (27%)
Curious about AI and future robot capabilities	9 (41%)

In the last robot-administered session, most participants (N=16) chose to engage in an optional open dialogue with the robot, showing their interest in interacting with Blossom outside the study tasks. The majority (N=10) suggested topics of conversation and engaged with Blossom to learn about geography and technology, discuss activities of interest, past trips, and music, share childhood memories, and reflect on family and friends.

Many participants (N=10) expressed interest in future interactions due to Blossom's novelty: *"I just want to know what she's got to say. I've never talked to a robot before"* (P10, S2). They hoped HRI would better prepare them for the increasing role of AI in the future, *"In the future, we will have to get used to living with [robots and AI], and this is an experience to get used to"* (P17, S3).

Despite general interest, some participants (N=9) hesitated to continue interacting with Blossom due to distrust in its capabilities and privacy concerns, *"I'm afraid of being listened to"* (P24, S2). These participants doubted the accuracy of Blossom's speech: *"I'm not for sure if it's gonna give me the right answer. [...] It is man-made and it can make mistakes"* (P10, S3). Some participants mentioned that they *"would have to be familiar with it"* (P8, S3) for future interactions. Additionally, hesitant participants (N=6) felt that Blossom would not be useful in their daily lives, often remarking that they might forget about Blossom due to other responsibilities, *"Days might go by [I would] forget about he even there"* (P7, S3).

5.3.4 Discussion

This work introduced LLM-powered SARs in the context of promoting and assessing cognitive health in elder care. In a 5-week study, 22 older adults engaged in three weekly robot-administered sessions to evaluate task performance, engagement, and user experience beyond the novelty effect. Two task variations were alternated to minimise task memorisation and an individual baseline was established through an initial human-administered session.

A. Summary of Findings

The findings demonstrated the efficacy of SARs for in-residence cognitive support. Relative to each research question (see Section 5.3.2.1), this study found that:

1. The conversational SAR successfully engaged older adults with varying cognitive abilities in repeated weekly sessions of cognitive health-promoting tasks;
2. Participants improved task performance after three robot-administered sessions, with a significantly enhanced level of detail, increased word count, and improved semantic completeness in their verbal picture descriptions, fewer word repetitions in semantic fluency, and a reduced need for hints by the administrator to complete the tasks;
3. Participants accepted the robot and found it increasingly easy to use over time, giving it an average SUS score above 85.6 after the initial novelty wore off;
4. Participants perceived the SAR as a valuable tool to enhance thinking, increase focus, and help them learn new skills; they were willing to interact with it in the future and, in some cases, enthusiastic to have their own robot.

These contributions aim to inform the design of more accessible, personalised, and autonomous robots that can effectively promote and assess cognitive health in elder care in real-world settings beyond clinics.

B. Robots as Tools to Promote Cognitive Health

The study demonstrated the efficacy of SARs in engaging older adults in cognitive health-promoting tasks. The role of the SAR was to act as a delivery mechanism for these interactive cognitive tasks aimed at verbally engaging target populations. A critical factor for the successful home deployment of SARs is their accessibility. Blossom was selected as the platform for this study because it is affordable and could be quickly fabricated at the scale needed for future deployments, making it a viable tool for broader use in elder care. Blossom is an embodied robot that combines voice interaction with additional interactive capabilities that can enhance user engagement compared to technologies that rely solely on voice.

The qualitative insights on user impressions and experience can inform the design of future SARs, targeting long-term acceptance, engagement, and comfort among older adults. Participants frequently reported cognitive benefits from their interactions with Blossom. Particularly, they perceived the robot as useful for enhancing thinking and memory recall, learning new tasks, and maintaining focus – not only during study tasks but also in their awareness and attention to detail in daily living. These self-reported benefits underscore the potential of SARs to keep older adults cognitively engaged while supporting well-being through continuous natural language interactions.

While some participants initially felt uncomfortable engaging with Blossom, their discomfort decreased with repeated exposure (as seen in the anxiety ratings in Figure 5.12). After repeated interactions, some participants also expressed that they developed an emotional connection with Blossom. There were mixed perceptions about trust in the robot's advice or suggestions, with participants with higher cognitive function (i.e., higher MMSE scores) showing lower trust and more hesitant attitudes, possibly due to greater awareness of AI and consequent privacy concerns. Nevertheless, trust scores increased over the study sessions across the cohort (see Figure 5.12) and 75% of participants in the normal ageing cognitive group chose to engage in optional open dialogue with the robot.

This finding highlights the importance of promoting familiarity and personalising interactions between SARs and older adults. Tailored interactions that align with users' interests and cognitive abilities can increase trust and enhance the perceived cognitive health benefits of the technology. This approach is particularly important for overcoming the challenges associated with introducing voice technology to current older generations and individuals with cognitive impairment. As suggested in previous research [145], a training period can prevent confusion and discomfort. Along similar lines, this study leveraged the robot's agency to promote familiarity and transparency of its interactive capabilities and limitations, including the motor sounds and the time required for speech processing.

C. Considerations for in-Home Cognitive Assessment

Dementia is a growing concern worldwide as the population ages. This study was not focused on clinical screening, hence pre/post-study cognitive outcomes were not measured. However, the findings suggested the potential for using SARs as a delivery mechanism for cognitive assessment in residential settings. Two tasks were implemented based on clinically validated cognitive tools, allowing data collection on spontaneous speech and verbal fluency, which have predictive value for early dementia screening [11, 122, 167]. Additionally, the data collected were comparable to existing normative data across different age and cognitive groups (i.e., normal ageing and cognitive impairment), highlighting the potential of the technology as a tool for relevant cognitive data collection.

Furthermore, repeated robot-administered sessions, conducted at weekly intervals with alternated task versions, led to overall improvements in verbal fluency of both cognitive groups. This finding highlights future opportunities for SARs to not only support assessments in residential settings but also help to maintain cognitive health in elder care through regular, personalised verbal HRI.

D. Design Implications

This work leveraged the linguistic abilities of LLMs to enable a SAR to administer interactive cognitive health-promoting tasks. Compared to rule-based chatbots, the use of LLMs

offers dialogue flexibility, a higher variability of responses, and the ability to understand and adapt to the nuances of multilingual contexts. This capability makes the verbal interaction more contextually relevant, engaging, and adaptable during the cognitive-health promoting tasks. However, the integration of LLMs into SARs for real-world deployments, particularly in unconstrained living environments, presents several design challenges and implications. While LLMs excel at generating contextualised natural language responses, they can also produce unpredictable or inappropriate content, especially in long-term, real-world living contexts where noise, data variability, and errors in speech recognition increase. Ensuring the safe and reliable use of LLMs in these contexts is essential.

D1. Safe and reliable use of LLMs.

The study implemented an LLM-driven moderator as a preventive measure designed to review and filter the quality of LLM-generated responses before they were delivered to participants via TTS. The purpose of this pilot safety mechanism was to ensure that all interactions remained appropriate and constrained within the context of cognitive tasks. A review of post-interaction transcripts confirmed that the LLM-generated responses were effectively constrained, with no problematic statements that required automated intervention by the moderator LLM (i.e., regenerating a response or terminating the interaction). However, this AI-driven approach relied on the same language model with a different prompt, hence it may suffer from the same shortcomings, such as biases or inaccuracies in the information provided.

To mitigate these risks, a human-in-the-loop approach was followed, where a researcher oversaw the sessions, with the ability to review the moderator's output in real time and interrupt the program if necessary. While effective in the study, this approach is not feasible for long-term in-home applications. Therefore, it is important for future research to develop robust methods to monitor LLM-generated speech and minimise problematic responses, particularly among vulnerable populations.

There is growing recognition of LLMs as promising tools for auto-evaluation [132]. However, their reliability remains a topic of careful examination due to biases and inconsistencies frequently observed in their judgment results [266]. Of particular relevance to this study, recent work in LLMs for digital health has applied auto-evaluation to ensure the quality and relevance of generated responses within the context provided [272]. Furthermore, Retrieval Augmented Generation (RAG) techniques can be leveraged to ground LLM responses in a curated and reliable knowledge base. This RAG grounding approach helps to mitigate hallucinations, promotes factual accuracy, and allows the latest health and well-being knowledge to be incorporated in LLM responses [188]. The integration of RAG is a promising avenue for future iterations of this study.

D2. Speech interaction dynamics.

Designing conversational AI for older adults requires careful consideration of speech interaction dynamics, including speech recognition, processing time, and communication breakdowns, as identified in the two studies presented before (see Section 5.1 and 5.2). These challenges are compounded in older populations with cognitive impairments, with frequent hesitations, pauses, and slower speech, all of which complicate turn-taking. It is important to avoid interruptions during user speech, especially when users pause to find the right words.

Latency in verbal interactions was identified, stemming from the use of cloud-based TTS, Whisper speech recognition, LLM API calls, and connectivity issues in the retirement homes where the study was conducted. These factors occasionally disrupted conversations, with some participants (27%) interpreting it as a sign that the robot was indifferent or unresponsive to their queries. These findings suggest that extended response times and interruptions can lead to disengagement, highlighting the need for robust turn-taking, accurate speech recognition in multilingual contexts, and adequate network infrastructure. Despite these challenges, user engagement and high Enjoyment scores (see Figure 5.12) indicate that latency did not affect overall robot acceptance. Additionally, participants learned to use the robot over the study sessions and became more accepting of its limitations, including processing time, aided by priming, which included transparency about the robot's speech processing time.

Response delays in verbal HRI due to issues with automatic speech recognition and response generation are a well-recognized research challenge [162, 192, 277]. Effective spoken interaction with SARs requires dialogue systems capable of handling multiple communication channels and quickly adapting to dynamic environments. Current cloud-based TTS and STT systems, while more accurate and superior in vocabulary size than local speech recognisers, often suffer from high latency, which disrupts verbal interactions. To enhance human-robot engagement among older adults, SARs need to be more responsive in real-time, with interruptible, adaptable speech processes.

D3. Addressing privacy concerns and facilitating independent use.

Despite general interest in interacting with Blossom, 41% of participants hesitated to continue using the robot due to distrust in its capabilities and concerns about privacy, particularly regarding who might access and read their personal data. Interviews indicated that these mistrustful attitudes were shaped by media coverage, and participants trusted the robot more when they had a clear understanding of its capabilities and purpose.

Privacy concerns related to voice data collection remain significant among older adults [107, 149, 229, 234]. Addressing both research ethics (e.g., informed consent and data protection) and societal ethics (e.g., integration of nonintrusive SARs at home) requires transparent communication about the robot's functions, limitations, and data access protocols to foster

long-term engagement, trust, and acceptance. Strategies could include priming with robot's agency to set realistic expectations (as incorporated in our study design), and stakeholder involvement throughout the design, initial evaluation, and deployment phases. Furthermore, ensuring feature anonymity in voice analysis can prevent privacy threats associated with speech processing [124].

Despite overall enjoyment and perceived usefulness, some participants reported lacking the technical skills or training to operate the robot independently. To address this, this research recommends that SARs include accessible training resources and consider the role of family members or carers in facilitating interactions.

E. Limitations and Future Work

Despite targeting a diverse and multilingual population, the sample size was relatively small. To improve the generalisability of the findings, future studies should include larger, more culturally diverse cohorts to evaluate the feasibility of SARs for in-home cognitive support across varied populations and settings. While the 5-week period was sufficient for evaluating performance changes, engagement, and user impressions beyond the novelty effect, it did not evaluate the long-term efficacy of SARs in daily, unsupervised living contexts. Future research should focus on longer-term deployments to assess the role of SARs in continuous cognitive health monitoring. Drawing from the work presented in Chapter 3, integrating SARs that promote cognitive health with home monitoring technologies could further enhance their ability to support safety and well-being by tracking behaviour, monitoring cognitive health, detecting safety risks, and providing real-time updates to carers.

The study focused on two SAR-administered cognitive tasks based on validated tools: picture description and semantic fluency. Future studies could target a broader range of cognitive functions, including memory, attention, executive function, and problem-solving. Although cognitive outcomes were not the primary focus of this study, future research could benefit from incorporating pre- and post-intervention measures to directly evaluate the impact of SARs on cognitive function.

To minimise task memorisation by older adult participants, the study design alternated two variations of each task across five weekly sessions. The primary analysis compared task performance between the first (individual baseline) and last session, following three weekly robot-administered sessions. While a 2-week washout period was included between the same task version, the potential for learning effects on task performance is acknowledged. Despite potential learning effects, the performance improvements observed between the first and last session support the primary goal of evaluating the feasibility of SARs as tools for promoting cognitive health through verbal interactions based on clinically validated cognitive tools. To that end, the weekly interactions lasted only 10 minutes, for a total of 30 minutes with

the robot. This duration was considered sufficient to overcome novelty effects and enable participants to evaluate the robot. However, many participants expressed interest in more interactions with the robot, and increasing the duration or frequency of interactions may improve cognitive task performance. Future work should therefore incorporate a more diverse set of cognitive tasks to provide more opportunities for interaction and explore the relationship between interaction time, cognitive health outcomes, and robot acceptance.

Another limitation highlighted by participants was the latency in verbal interaction with Blossom. While this is a well-recognised challenge in spoken language interactions with robots [162], these technological issues are being addressed by new tools such as OpenAI's Realtime API¹², which reduces latency by streaming audio inputs and outputs directly with a persistent WebSocket connection to access GPT-4o – a promising area for future work.

Future study iterations with older populations could incorporate the suggested design features for Blossom. These include more expressive facial features, ear movements, and personalised interactions with an appropriate tone and topics of interest. Lastly, analysing the eye gaze data collected during verbal dialogue with an LLM-powered SAR presents an intriguing avenue for evaluating focus during cognitive tasks, which could provide additional insights into cognitive function.

¹²OpenAI's Realtime API (<https://openai.com/index/introducing-the-realtime-api/>).

5.4 Summary

This chapter presented three user studies conducted in the UK, India and the USA with stakeholder involvement (N=111), including older adults, PLWD, carers, clinicians and dementia care professionals. These user studies evaluated the acceptance of conversational robots, user engagement, and feasibility of use in residential settings.

In *Study 1*, three participatory workshops were conducted to explore the role of conversational AI for health and well-being support of target populations at home. This study uncovered the perceived utility of verbal prompts to support routine, personalised conversations to enhance cognitive function, and home integration to mitigate safety risks. Seven design guidelines were proposed to provide meaningful, accessible and non-invasive home support, while addressing barriers to long-term engagement.

In *Study 2*, a low-cost affective robotic platform was designed and validated for use in India, demonstrating clinical potential for remote cognitive engagement. This work further evaluated the acceptance and user engagement with three conversational robots (a voice agent, a virtual robotic platform, and an embodied robot), assessing socioeconomic factors and feasibility of home use in India by exploring the perceptions and concerns of stakeholders.

In *Study 3*, an LLM-powered SAR was designed and validated for in-residence cognitive support in elder care, an approach that had not been tested before. The 5-week study showed that the conversational robot effectively engaged older adults in repeated sessions of human-robot verbal dialogue based on clinically validated tools, resulting in significantly increased cognitive task performance, higher social engagement in robot-administered tasks compared to those administered by a human, robot acceptance and overall interest in future use.

In conclusion, this chapter offered novel insights from real-world evaluation studies and stakeholder perspectives on the use of conversational robots to support ageing and dementia care in residential settings. These insights complement the technical findings of Chapter 3 and Chapter 4, altogether taking a step forward into the design and deployment of conversational AI for monitoring behaviour, mitigating safety risks, and collecting clinical data from longitudinal verbal interactions to track cognitive health and automatically screen for cognitive impairment.

Chapter 6

Conclusion

This chapter summarises the main contributions of this dissertation, discusses limitations, and describes future research directions that follow from this work. This chapter concludes with closing remarks.

6.1 Dissertation Summary

The goal of this dissertation was to design and evaluate conversational robots that verbally engage with older populations and PLWD to support well-being, enhance home safety and monitor cognitive health at home. This research comprehensively reviewed the literature and identified key gaps in the field:

- Longitudinal data collection with conversational technology at home;
- Sustained user engagement beyond the initial novelty phase, compounded by insufficient stakeholder involvement to guide technology design and home deployments, and the challenge of adapting verbal interactions to individual needs for personalised care;
- Limited generalisability and clinical applicability of ML approaches for automated cognitive assessment from spoken language, compounded by the lack of out-of-clinic data to validate predictive models.

The work presented in this dissertation draws from real-world evaluation studies and active stakeholder involvement to advance the use of conversational AI to support well-being, increase home safety, and collect longitudinal data from verbal interactions as a biomarker for assessing the risk of cognitive impairment. The contributions and implications of each study presented in Chapter 3, Chapter 4, and Chapter 5 are outlined below.

A. Behaviour Monitoring to Support Well-being and Home Safety (Chapter 3)

To explore the integration of conversational technology at home, this research conducted a longitudinal deployment study (>1 year¹) in households of PLWD (N=14). Participants interacted with a commercial conversational agent (Amazon Alexa) to complete a well-being questionnaire and trigger topics of interest. In-home activity sequences were analysed using data captured from remote monitoring devices (e.g., motion, vitals, use of kitchen appliances) and conversational technology. This study revealed:

1. A significant decrease in user engagement after the novelty phase ($p < .025$), highlighting the need to adapt voice interactions to individual needs and preferences for personalised and proactive support in dementia care.
2. Technical feasibility of identifying trends and changes in behaviour based on household activity that could indicate safety risks, changes in routines or well-being.
3. Participants continued using the technology after clinical events, suggesting future opportunities for adaptive questionnaires to directly ask users for self-assessments of health and well-being.

¹In-home activity and interaction data were collected over seven months on average across all households, with the longest data collection period exceeding one year.

Key contributions and implications: This study contributes novel insights and long-term voice data collection in the homes of PLWD. The analysis demonstrated the feasibility of identifying behavioural trends by correlating activity data from remote monitoring technologies with the use of conversational AI in smart homes. This approach provides a basis for: 1) verifying health risks detected by remote monitoring technologies; 2) offering proactive verbal support in response to safety risks, e.g., noting appliances to be turned off, offering guidance in the event of perceived agitation or confusion; 3) encouraging behaviours that promote well-being; and 4) initiating automated alerts.

B. Effectiveness and Clinical Applicability of Cognitive Health Predictive Modelling from Spoken Language (Chapter 4)

This study introduced an interpretable ML approach for automated assessment of cognitive health from spoken language with a focus on clinical applicability. The analysis included ADRD detection in binary classification and MMSE score prediction to assess the severity of cognitive decline. A benchmark DementiaBank dataset (N=291) obtained during verbal picture descriptions was used and model generalisability was validated on a pilot dataset collected in-residence (N=22). The proposed ML approach demonstrated that:

1. In ADRD detection, a RF model trained on 100 NLP-based features achieved a ROC-AUC of 85.7% (95% CI=83.8–87.6) on the benchmark DementiaBank test set and 65.4% (95% CI=54.9–70.1) on the real-world pilot dataset.
2. In MMSE prediction, a RFR model achieved a MAE of 3.7 (95% CI=3.7–3.8), with a comparable performance of 3.3 (95% CI=3.1–3.5) on real-world data, demonstrating model generalisability and real-world applicability.
3. Risk stratification analysis improved model performance overall, resulting in a 13% increase in specificity on the benchmark DementiaBank dataset and a 21% increase on the real-world dataset.
4. Reduced analytical thinking, increased reliance on pronouns and adverbs, greater disfluency, lower lexical diversity and reduced use of words reflecting a psychological state of completion contributed to higher ADRD risk.

Key contributions and implications: This work contributes to: 1) the clinical applicability of cognitive predictive modelling from spoken language; 2) informing clinicians about changes in language most predictive of cognitive decline; and 3) improving health-care resource allocation by identifying higher-risk individuals. The proposed ML approach shows promise for integration with conversational technology at home for accessible and

scalable monitoring of cognitive health from verbal interactions. This could facilitate earlier detection of higher-risk individuals, allowing clinicians to prioritise those requiring immediate attention, further diagnostic workup, and timely intervention.

C. Design Guidelines for Leveraging Conversational AI in Dementia Care (Chapter 5 Study 1)

The previous chapters discussed the potential of conversational AI to monitor behaviour, support well-being, and assess cognitive health through verbal interactions. Realising this potential requires more than technological development – it is essential to incorporate real-world perspectives and experiences from stakeholders to address existing gaps in long-term user engagement and technology acceptance. To address these gaps, three participatory workshops were conducted involving stakeholders with lived experience with dementia (N=20). This study provided expert-informed insights on:

1. The perceived benefits of using conversational AI to support daily tasks and well-being. Participants valued the ability to support daily routines with personalised verbal prompts, mitigate safety risks, and support cognitive function.
2. Key barriers to long-term adoption, which include accessibility, trust, and reliance on carers for technology operation.
3. Design considerations that can help address these barriers: adapting interactions to individual needs and cognitive abilities; promoting familiarity with technology; using a task-based and encouraging communication style that respects user autonomy; and ensuring transparency of technology capabilities to avoid mistrust.

Key contributions and implications: By incorporating stakeholders' invaluable perspectives, this study proposed seven detailed guidelines on how to leverage conversational AI for long-term engagement and meaningful support at home while addressing existing challenges of home deployments.

D. Cultural Feasibility of Conversational Robots in India (Chapter 5 Study 2)

Health inequality significantly increases the impact of dementia in underprivileged and lower socioeconomic communities [119, 153]. The challenge of dementia in India stands out due to its vast population and high incidence rates, limited resources, poor clinician accessibility, and a striking 90% gap in diagnosis and care [62]. Aiming to address this gap, the first cultural feasibility study of conversational robots in India was conducted. A low-cost affective robotic platform was designed and validated with older adults (N=26) and PLWD (N=14). The study further explored the perceptions and attitudes of stakeholders (N=29) towards the adoption

of conversational robots through HRI evaluation with three conversational robot platforms and identified the socioeconomic barriers that inhibit home integration in India. This study showed that:

1. Robot-assisted cognitive engagement sessions with PLWD were feasible in India, which had not been tested before, and the low-cost affective robotic platform designed was culturally validated among target populations.
2. Adaptations to robot design and implementation are essential for cultural feasibility in India, including: enhanced speech processing for local accent interpretation; appropriate communication style and appearance through tone, choice of words and facial expressions customisable by the user; and a phased introduction to promote familiarity with technology starting in clinics.

Key contributions and implications: This research introduced the first robot for dementia care in India, highlighting its clinical potential for remote cognitive engagement. It also paved the way for a more inclusive and culturally aware robot design for long-term user engagement and feasibility of implementation in India.

E. Efficacy of SARs for In-Residence Cognitive Support (Chapter 5 Study 3)

To investigate the feasibility of using LLM-powered SARs to promote and assess cognitive health, this work implemented human-robot verbal dialogue in cognitive tasks based on clinically validated tools and conducted a 5-week within-subjects study involving older adults (N=22) living in retirement homes. This study revealed:

1. Task performance increased after three weekly robot-administered sessions, with significantly more detailed picture descriptions, fewer word repetitions in semantic fluency, and fewer hints required from the administrator to complete the tasks ($p < .01$). Additionally, there was an overall increase in the number of unique task-relevant words recalled in the last session during the semantic fluency task.
2. The conversational robot engaged older adults in repeated HRI sessions aimed at promoting cognitive health, with significantly higher social engagement in robot-administered tasks compared to those administered by a human ($p < .05$).
3. Older adults accepted the SAR in this context, recognised its utility for enhancing cognitive abilities, including focus, thinking, and learning new skills, and expressed interest in future use.

Key contributions and implications: This study contributes novel insights into the efficacy of LLM-powered SARs to promote cognitive health in elder care within residential settings. With consistent use, for which accessibility and personalised interactions are key, SARs could ultimately monitor and assess cognitive health through human-robot verbal interactions and identify individuals requiring more comprehensive diagnostic evaluation.

6.2 Limitations

While this dissertation presents advancements in the use of conversational AI to support well-being and home safety, specifically for behaviour monitoring and automated cognitive health assessment, limitations of the work are acknowledged. Each study discussed the limitations in their detailed context. This section discusses limiting factors and open challenges in the broader context of in-home behaviour monitoring, longitudinal interactions with conversational technology, and predictive models for screening cognitive impairment.

Ground truth and longitudinal data collection. A key challenge identified in this research is establishing ground truth for validating predictive models, individual behavioural patterns and safety risks at home. Obtaining ground truth for well-being and behaviour monitoring requires an appropriate balance to have sufficient labelled training data while respecting user privacy at home. Furthermore, this research argues that longitudinal data collected with conversational AI could provide valuable insights into cognitive health based on spoken language patterns. However, collecting such data in-residence is challenging, user engagement in target populations is low, which limits the amount and quality of data collected over time, and there is a lack of out-of-clinic patient data for validating ML models.

Privacy vs. in-home monitoring. The deployment of in-home monitoring technologies in real-world evaluation studies presents considerable challenges. While monitoring systems and automated interventions hold strong potential to enhance health and well-being, their success relies on sustained user engagement and acceptance. Privacy was identified as a major concern among older adults and people affected by dementia involved in the user studies (Chapter 5), many of whom were hesitant to share personal data and perceived the technology as intrusive or overly complex. Participants in the year-long deployment study (Chapter 3) also raised concerns about having their voice interactions recorded and stored, which resulted in a reduced use of conversational technology over time. Addressing these concerns requires thoughtful design with active stakeholder involvement, which this work addressed in Chapter 5. As discussed in the chapter, stakeholder feedback indicated that transparent communication about data usage, access, and its limitations can promote technology acceptance among target populations. Ethical issues about the collection of identifiable voice data and sensitive medical information remain limiting factors in home deployments.

Data quality and bias. Data-driven approaches, such as the automated assessment of cognitive health from spoken language (Chapter 4), rely heavily on the quality and representativeness of the data. Biases in the data can lead to skewed results, limiting the generalisability and clinical applicability of the findings. Additionally, the use of spoken language data publicly available or collected predominantly from white, English-speaking populations raises concerns about the models' screening applicability to more diverse groups. In the analysis presented in Chapter 4, the benchmark DementiaBank data were assessed for biases in age and sex, but

ethnicity information was not available. Model generalisability was validated on a largely different cohort representing a diverse ethnic group and including English and Spanish-speaking participants using automated translations. However, future studies should include larger and more culturally diverse datasets, and models should be fine-tuned to account for additional confounding factors, such as education level, technology familiarity and socioeconomic status. Furthermore, transcription errors introduced by ASR systems can influence the accuracy of the extracted linguistic and acoustic features. These limitations must be balanced against the need for automated tools, from data collection through pre-processing to analysis, that function effectively in real-world settings, where human annotation is impractical.

Participant cohorts and long-term efficacy at home. While this research involved a diverse group of stakeholders (N=111) through user studies conducted in the UK, India, and the USA (Chapter 5), the focus was primarily on user perceptions, first impressions and engagement in HRI. Future work should explore the long-term efficacy of conversational robots in daily, unsupervised living contexts to better understand their potential for sustained use and impact on health and well-being. It is also important to acknowledge the inherent challenges in engaging people affected by dementia in research, particularly in recruitment and adherence to information governance protocols.

Interactions with conversational technology. Most interactions with commercial smart speakers, such as Amazon Alexa, or mobile applications, such as Siri, require an invocation name for activation, which users may not always recall. Additionally, these systems include a limited response window², which often disrupts verbal interactions. Current cloud-based TTS and STT systems, while more accurate and superior in vocabulary size than local speech recognisers, often suffer from high latency. Response delays, speech recognition errors and the ability to handle user pauses without interrupting are well-recognised challenges in spoken language HRI that remain unresolved [162].

Reliable use of LLMs. In this research, LLMs were used to identify topics from interactions with Amazon Alexa (Chapter 3), extract vector representations from audio data to train ML models (Chapter 4), and generate contextual dialogue in robot-administered cognitive tasks (Chapter 5 Study 3). The integration of LLMs into conversational robots holds potential to improve dialogue flexibility and multilingual interpretation. However, the ethical implications of deploying these systems for content generation must be carefully considered to ensure that they promote user well-being and independence. Risks associated with LLMs, such as misinformation, could be mitigated by implementing robust methods to moderate LLM-generated content and filter problematic responses. A pilot safety measure using a moderator was discussed in Chapter 5 Study 3, however the experiments presented relied on a human-in-the-loop approach to mitigate risks. Developing such solutions for automated moderation remains an open research challenge.

²For instance, Amazon Alexa includes a default 8-second window before timing out.

6.3 Future Directions

This research offers contributions that support the larger vision of **using conversational AI to support well-being, increase home safety, and collect longitudinal data from verbal interactions as a biomarker for assessing the risk of cognitive impairment**. The findings of this dissertation, their implications and the limitations previously discussed point toward future research directions.

Conversational technology integrated into the home environment with additional monitoring tools could offer a more holistic view of an individual's health and well-being by actively querying users for self-assessments and capturing their observations and concerns – subjective insights often missed by passive monitoring systems alone. This way, voice could serve as an additional modality for in-home monitoring and complement data collected from other sensors tracking activity, behaviour, or sleep, which predictive models could use for remote monitoring of comorbidities, such as the risk of agitation or infections [38].

Conversely, subjective data collected over time can be used to tailor verbal interactions to individual needs and preferences, taking a step forward to proactive personalised care. While achieving such adaptation of conversational systems remains an open research challenge, it could promote long-term user engagement. Furthermore, conversational AI could encourage behaviours and actions that promote well-being, helping with modifiable risk factors such as social isolation or physical inactivity. In particular, weekly verbal dialogue based on validated cognitive tools with conversational robots could promote cognitive health by increasing focus, stimulating thinking and learning, as demonstrated in this research (Chapter 5 Study 3).

If well-integrated in the home environment, conversational AI could verify safety risks detected by other technologies, reducing model uncertainty. It could also provide proactive verbal support in response to safety risks detected (e.g., noting appliances to be turned off, offering guidance in the event of confusion) and initiate automated alerts to inform relevant stakeholders, facilitating appropriate interventions.

Additionally, this research demonstrated the potential for using longitudinal voice data collected in-residence via conversational AI to monitor cognitive health and identify individuals at higher risk of cognitive impairment based on spoken language patterns (Chapter 4). This realisation could pave the way for scalable, non-invasive and cost-effective screening of cognitive health in community care settings.

In addition to the research directions and open challenges previously outlined, this research proposes two areas for future investigation:

Multimodal data for in-home monitoring of health and well-being. This work identified behavioural trends by correlating activity data from remote monitoring technologies with the use of conversational AI in smart homes (Chapter 3). The identified trends and changes in

behaviour could be tied to clinical information and individual comorbidities. A conversational agent could proactively ask users for health self-assessments and verification of safety risks, which could lead to well-being insights.

Longitudinal monitoring of cognitive impairment risk. Previous research has explored the feasibility of predicting the progression from MCI to AD [11] using speech recordings from interviews in NPT. However, the potential for such an assessment using out-of-clinic data remains underexplored. Collecting longitudinal data from target populations at home over the years could enable continuous monitoring of the progression of cognitive impairment over time. To realise this potential, it is essential to overcome challenges of in-home deployments and long-term engagement. The user studies presented in Chapter 5 could be expanded by incorporating regular feedback in home settings to ensure sustained use and engagement with technology over years. The resulting data would be analysed in a way similar to that described in Chapter 4, with potential to monitor cognitive health from spoken language.

6.4 Closing Remarks

This dissertation argued the potential of using conversational AI to support well-being in ageing and dementia care, increase home safety, and collect voice as a biomarker to assess the risk of cognitive impairment.

Despite the increasing promise of conversational technology and SARs for home support, user engagement in target populations is low, longitudinal data collection in-residence is challenging, adaptation of interactions for personalised care is unresolved, and the use of clinically applicable models for automated screening of cognitive impairment is not established.

This research presented a year-long deployment of conversational technology in the homes of people affected by dementia. By analysing activity sequences in smart homes, it demonstrated the potential for behaviour monitoring to support well-being and safety. Conversational AI could provide proactive verbal support in response to safety risks and gather well-being insights from subjective experiences and self-assessments of health, complementing data collected by other smart home monitoring sensors.

Furthermore, this work evaluated the use of spoken language as a biomarker for automated screening of cognitive health. Predictive models trained on interpretable linguistic features effectively detected ADRD and predicted the severity of cognitive impairment via MMSE scores. The ML approach demonstrated generalisability on data collected in-residence and identified the linguistic features most strongly associated with the risk of cognitive impairment. This approach could serve as a basis for accessible in-home monitoring of cognitive health from voice interactions with conversational AI, enabling earlier detection of higher-risk individuals and timely intervention.

Realising this potential requires addressing key gaps in long-term user engagement, technology acceptance, and feasibility of home use. To do so, this research presented three real-world evaluation studies conducted in the UK, India, and the USA with active stakeholder involvement. Design guidelines were outlined for leveraging conversational AI to provide meaningful support in dementia care at home. A cultural feasibility study of conversational robots in India was further introduced. Additionally, the efficacy of LLM-powered SARs for in-residence cognitive support was validated.

In conclusion, this dissertation takes a step forward in using conversational AI to verbally engage with users at home to support well-being and safety while serving as an accessible tool for automated screening of cognitive impairment.

References

- [1] Abdi, J., Al-Hindawi, A., Ng, T., et al. "Scoping review on the use of socially assistive robot technology in elderly care". In: *BMJ open* 8.2 (2018), e018815.
- [2] Abdollahi, H., Mollahosseini, A., Lane, J. T., et al. "A pilot study on using an intelligent life-like robot as a companion for elderly individuals with dementia and depression". In: *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. IEEE. 2017, pp. 541–546.
- [3] Agbavor, F. and Liang, H. "Predicting dementia from spontaneous speech using large language models". In: *PLOS digital health* 1.12 (2022), e0000168.
- [4] Ahmed, S., Haigh, A.-M. F., Jager, C. A. de, et al. "Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease". In: *Brain* 136.12 (2013), pp. 3727–3737.
- [5] Alaghbari, K. A., Saad, M. H. M., Hussain, A., et al. "Activities recognition, anomaly detection and next activity prediction based on neural networks in smart homes". In: *IEEE Access* 10 (2022), pp. 28219–28232.
- [6] Alkhatib, S., Waycott, J., Buchanan, G., et al. "Privacy and the Internet of Things (IoT) monitoring solutions for older adults: A Review". In: *Connecting the System to Enhance the Practitioner and Consumer Experience in Healthcare* (2018), pp. 8–14.
- [7] Alzheimer's Association. *More Than Normal Aging: Understanding Mild Cognitive Impairment*. 2022.
- [8] Alzheimer's Disease International. *World Alzheimer Report 2023*. <https://www.alzint.org/resource/world-alzheimer-report-2023/>. Sept. 2023.
- [9] Alzheimer's Society. *My Carer*. <https://www.mrm.com/en/work-case-studies/frankfurt-alzheimers-society-my-carer.html>. Sept. 2022.
- [10] Alzheimer's Research UK. *UK dementia statistics*. <https://www.dementiastatistics.org/statistics/hospitals/>. Sept. 2022.
- [11] Amini, S., Hao, B., Yang, J., et al. "Prediction of Alzheimer's disease progression within 6 years using speech: A novel approach leveraging language models". In: *Alzheimer's & Dementia* (2024).
- [12] Angelov, D. "Top2vec: Distributed representations of topics". In: *arXiv preprint arXiv:2008.09470* (2020).
- [13] Antony, V. N., Cho, S. M., and Huang, C.-M. "Co-Designing with Older Adults, for Older Adults: Robots to Promote Physical Activity". In: *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 2023, pp. 506–515.
- [14] Arifoglu, D. and Bouchachia, A. "Activity recognition and abnormal behaviour detection with recurrent neural networks". In: *Procedia Computer Science* 110 (2017), pp. 86–93.
- [15] Arifoglu, D. and Bouchachia, A. "Detection of abnormal behaviour for dementia sufferers using Convolutional Neural Networks". In: *Artificial intelligence in medicine* 94 (2019), pp. 88–95.
- [16] Asgari, M., Kaye, J., and Dodge, H. "Predicting mild cognitive impairment from spontaneous spoken utterances". In: *Alzheimer's & Dementia: Translational Research & Clinical Interventions* 3.2 (2017), pp. 219–228.

- [17] Astell, A. J., Bouranis, N., Hoey, J., et al. "Technology and dementia: The future is now". In: *Dementia and geriatric cognitive disorders* 47.3 (2019), pp. 131–139.
- [18] Baevski, A., Hsu, W.-N., Xu, Q., et al. "Data2vec: A general framework for self-supervised learning in speech, vision and language". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 1298–1312.
- [19] Baevski, A., Zhou, Y., Mohamed, A., et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *Advances in neural information processing systems* 33 (2020), pp. 12449–12460.
- [20] Bajones, M., Fischinger, D., Weiss, A., et al. "Hobbit: Providing fall detection and prevention for the elderly in the real world". In: *Journal of Robotics* 2018.1 (2018), p. 1754657.
- [21] Baltrušaitis, T., Ahuja, C., and Morency, L.-P. "Multimodal machine learning: A survey and taxonomy". In: *IEEE transactions on pattern analysis and machine intelligence* 41.2 (2018), pp. 423–443.
- [22] Bangor, A., Kortum, P. T., and Miller, J. T. "An empirical evaluation of the system usability scale". In: *Intl. Journal of Human-Computer Interaction* 24.6 (2008), pp. 574–594.
- [23] Bedaf, S., Gelderblom, G. J., and De Witte, L. "Overview and categorization of robots supporting independent living of elderly people: What activities do they support and how far have they developed". In: *Assistive Technology* 27.2 (2015), pp. 88–100.
- [24] Beer, J. M., Prakash, A., Smarr, C.-A., et al. "Older users' acceptance of an assistive robot: Attitudinal changes following brief exposure". In: *Gerontechnology: international journal on the fundamental aspects of technology to serve the ageing society* 16.1 (2017), p. 21.
- [25] Bengt, J. F., Balsis, S., Geraci, L., et al. "How well do the ADAS-cog and its subscales measure cognitive dysfunction in Alzheimer's disease?" In: *Dementia and geriatric cognitive disorders* 28.1 (2009), pp. 63–69.
- [26] Breazeal, C. "Emotion and sociable humanoid robots". In: *International journal of human-computer studies* 59.1-2 (2003), pp. 119–155.
- [27] Broadbent, E., Garrett, J., Jepsen, N., et al. "Using robots at home to support patients with chronic obstructive pulmonary disease: pilot randomized controlled trial". In: *Journal of medical Internet research* 20.2 (2018), e8640.
- [28] Broadbent, E., Peri, K., Kerse, N., et al. "Robots in older people's homes to improve medication adherence and quality of life: a randomised cross-over trial". In: *Social Robotics: 6th International Conference, ICSR 2014, Sydney, NSW, Australia, October 27-29, 2014. Proceedings 6*. Springer. 2014, pp. 64–73.
- [29] Brodaty, H. and Donkin, M. "Family caregivers of people with dementia". In: *Dialogues in clinical neuroscience* (2022).
- [30] Brooke, J. "SUS: A "quick and dirty" Usability Scale". In: *Usability Evaluation in INdustry/Taylor and Francis* (1996).
- [31] Bruno, D. and Schurmann Vignaga, S. "Addenbrooke's cognitive examination III in the diagnosis of dementia: a critical review". In: *Neuropsychiatric Disease and Treatment* (2019), pp. 441–447.
- [32] Bucks, R. S., Singh, S., Cuerden, J. M., et al. "Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance". In: *Aphasiology* 14.1 (2000), pp. 71–91.

- [33] Caleb-Solly, P., Dogramadzi, S., Huijnen, C. A., et al. "Exploiting ability for human adaptation to facilitate improved human-robot interaction and acceptance". In: *The Information Society* 34.3 (2018), pp. 153–165.
- [34] Calzà, L., Gagliardi, G., Favretti, R. R., et al. "Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia". In: *Computer Speech & Language* 65 (2021), p. 101113.
- [35] Camilleri, A., Dogramadzi, S., and Caleb-Solly, P. "Learning from Carers to inform the Design of Safe Physically Assistive Robots-Insights from a Focus Group Study". In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2022, pp. 703–707.
- [36] Camp, N., Lewis, M., Hunter, K., et al. "Technology used to recognize activities of daily living in community-dwelling older adults". In: *International Journal of Environmental Research and Public Health* 18.1 (2021), p. 163.
- [37] Cao, Y., Zhou, L., Lee, S., et al. "Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study". In: *arXiv preprint arXiv:2303.17466* (2023).
- [38] Capstick, A., Palermo, F., Zakka, K., et al. "Digital remote monitoring for screening and early detection of urinary tract infections". In: *npj Digital Medicine* 7.1 (2024), p. 11.
- [39] Carroll, C., Chiodo, C., Lin, A. X., et al. "Robin: Enabling independence for individuals with cognitive disabilities using voice assistive technology". In: *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 2017, pp. 46–53.
- [40] Céspedes, N., Irfan, B., Senft, E., et al. "A socially assistive robot for long-term cardiac rehabilitation in the real world". In: *Frontiers in Neurorobotics* 15 (2021), p. 633248.
- [41] Chan, C. and Li, F. "Developing a natural language-based AI-chatbot for social work training: an illustrative case study". In: *China Journal of Social Work* 16.2 (2023), pp. 121–136.
- [42] Chang, Y.-L., Luo, D.-H., Huang, T.-R., et al. "Identifying Mild Cognitive Impairment by Using Human-Robot Interactions". In: *Journal of Alzheimer's Disease* 85.3 (2022), pp. 1129–1142.
- [43] Charmaz, K. *Constructing grounded theory*. sage, 2014.
- [44] Chauhan, U. and Shah, A. "Topic modeling using latent Dirichlet allocation: A survey". In: *ACM Computing Surveys (CSUR)* 54.7 (2021), pp. 1–35.
- [45] Chen, J., Ye, J., Tang, F., et al. "Automatic detection of alzheimer's disease using spontaneous speech only". In: *Interspeech*. Vol. 2021. NIH Public Access. 2021, p. 3830.
- [46] Chen, Y., Zhang, H., Liu, R., et al. "Experimental explorations on short text topic mining between LDA and NMF based Schemes". In: *Knowledge-Based Systems* 163 (2019), pp. 1–13.
- [47] Cheng, A., Raghavaraju, V., Kanugo, J., et al. "Development and evaluation of a healthy coping voice interface application using the Google home for elderly patients with type 2 diabetes". In: *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE. 2018, pp. 1–5.
- [48] Chimamiwa, G., Giaretta, A., Alirezaie, M., et al. "Are Smart Homes Adequate for Older Adults with Dementia?" In: *Sensors* 22.11 (2022), p. 4254.

- [49] Cho, M., Lee, S.-s., and Lee, K.-P. "Once a kind friend is now a thing: Understanding how conversational agents at home are forgotten". In: *Proceedings of the 2019 on Designing Interactive Systems Conference*. 2019, pp. 1557–1569.
- [50] Chopard, D., Treder, M. S., Corcoran, P., et al. "Text mining of adverse events in clinical trials: deep learning approach". In: *JMIR Medical Informatics* 9.12 (2021), e28632.
- [51] Clark, L. J., Gatz, M., Zheng, L., et al. "Longitudinal verbal fluency in normal aging, preclinical, and prevalent Alzheimer's disease". In: *American Journal of Alzheimer's Disease & Other Dementias®* 24.6 (2009), pp. 461–468.
- [52] Cohen, J. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.
- [53] Cohen-Mansfield, J., Dakheel-Ali, M., and Marx, M. S. "Engagement in persons with dementia: the concept and its measurement". In: *The American journal of geriatric psychiatry* 17.4 (2009), pp. 299–307.
- [54] Cook, D. J., Krishnan, N. C., and Rashidi, P. "Activity discovery and activity recognition: A new partnership". In: *IEEE transactions on cybernetics* 43.3 (2013), pp. 820–828.
- [55] Cook, D. J., Schmitter-Edgecombe, M., and Dawadi, P. "Analyzing activity behavior and movement in a naturalistic environment using smart home techniques". In: *IEEE journal of biomedical and health informatics* 19.6 (2015), pp. 1882–1892.
- [56] Cruz-Sandoval, D., Morales-Tellez, A., Sandoval, E. B., et al. "A social robot as therapy facilitator in interventions to deal with dementia-related behavioral symptoms". In: *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*. 2020, pp. 161–169.
- [57] De la Fuente Garcia, S., Ritchie, C. W., and Luz, S. "Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review". In: *Journal of Alzheimer's Disease* 78.4 (2020), pp. 1547–1574.
- [58] Deep, S., Zheng, X., Karmakar, C., et al. "A survey on anomalous behavior detection for elderly care using dense-sensing networks". In: *IEEE Communications Surveys & Tutorials* 22.1 (2019), pp. 352–370.
- [59] Deng, E., Mutlu, B., Mataric, M. J., et al. "Embodiment in socially interactive robots". In: *Foundations and Trends® in Robotics* 7.4 (2019), pp. 251–356.
- [60] Devlin, J., Chang, M.-W., Lee, K., et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).
- [61] Di Nuovo, A., Broz, F., Wang, N., et al. "The multi-modal interface of Robot-Era multi-robot services tailored for the elderly". In: *Intelligent Service Robotics* 11 (2018), pp. 109–126.
- [62] Dias, A. and Patel, V. "Closing the treatment gap for dementia in India". In: *Indian journal of psychiatry* 51.Suppl1 (2009), S93.
- [63] Dingler, T., Kwasnicka, D., Wei, J., et al. "The use and promise of conversational agents in digital health". In: *Yearbook of Medical Informatics* 30.01 (2021), pp. 191–199.
- [64] Dino, F., Zandie, R., Abdollahi, H., et al. "Delivering cognitive behavioral therapy using a conversational social robot". In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 2089–2095.
- [65] Do, H. M., Sheng, W., Harrington, E. E., et al. "Clinical screening interview using a social robot for geriatric care". In: *IEEE Transactions on Automation Science and Engineering* 18.3 (2020), pp. 1229–1242.

- [66] Egger, R. and Yu, J. "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts". In: *Frontiers in Sociology* 7 (2022).
- [67] Eichler, T., Thyrian, J. R., Hertel, J., et al. "Rates of formal diagnosis of dementia in primary care: The effect of screening". In: *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 1.1 (2015), pp. 87–93.
- [68] Engelman, M., Agree, E. M., Meoni, L. A., et al. "Propositional density and cognitive function in later life: Findings from the precursors study". In: *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 65.6 (2010), pp. 706–711.
- [69] Enshaeifar, S., Zoha, A., Markides, A., et al. "Health management and pattern analysis of daily living activities of people with dementia using in-home sensors and machine learning techniques". In: *PloS one* 13.5 (2018), e0195605.
- [70] Eyben, F., Scherer, K. R., Schuller, B. W., et al. "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing". In: *IEEE transactions on affective computing* 7.2 (2015), pp. 190–202.
- [71] Eyben, F., Wöllmer, M., and Schuller, B. "Opensmile: the munich versatile and fast open-source audio feature extractor". In: *Proceedings of the 18th ACM international conference on Multimedia*. 2010, pp. 1459–1462.
- [72] Eyssel, F. and Kuchenbrandt, D. "Social categorization of social robots: Anthropomorphism as a function of robot group membership". In: *British Journal of Social Psychology* 51.4 (2012), pp. 724–731.
- [73] Fasola, J. and Mataric, M. J. "Using socially assistive human-robot interaction to motivate physical exercise for older adults". In: *Proceedings of the IEEE* 100.8 (2012), pp. 2512–2526.
- [74] Fasola, J. and Matarić, M. J. "A socially assistive robot exercise coach for the elderly". In: *Journal of Human-Robot Interaction* 2.2 (2013), pp. 3–32.
- [75] Feingold-Polak, R., Barzel, O., and Levy-Tzedek, S. "Socially Assistive Robot for Stroke Rehabilitation: a Long-Term in-the-Wild Pilot Randomized Controlled Trial". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2024).
- [76] Fisch, A., Jaakkola, T., and Barzilay, R. "Calibrated selective classification". In: *arXiv preprint arXiv:2208.12084* (2022).
- [77] Forbes-McKay, K. E. and Venneri, A. "Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task". In: *Neurological sciences* 26 (2005), pp. 243–254.
- [78] Fraser, K. C., Meltzer, J. A., and Rudzicz, F. "Linguistic features identify Alzheimer's disease in narrative speech". In: *Journal of Alzheimer's Disease* 49.2 (2016), pp. 407–422.
- [79] Fristed, E., Skirrow, C., Meszaros, M., et al. "A remote speech-based AI system to screen for early Alzheimer's disease via smartphones". In: *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 14.1 (2022), e12366.
- [80] Galatzer-Levy, I. R., McDuff, D., Natarajan, V., et al. "The capability of large language models to measure psychiatric functioning". In: *arXiv preprint arXiv:2308.01834* (2023).
- [81] Garcia, S. D. L. F., Haider, F., Fromm, D., et al. "Connected Speech-Based Cognitive Assessment in Chinese and English". In: *arXiv preprint arXiv:2406.10272* (2024).
- [82] Garrard, P., Maloney, L. M., Hodges, J. R., et al. "The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author". In: *Brain* 128.2 (2005), pp. 250–260.

- [83] Gasteiger, N., Ahn, H. S., Lee, C., et al. "Participatory Design, Development, and Testing of Assistive Health Robots with Older Adults: An International Four-year Project". In: *ACM Transactions on Human-Robot Interaction (THRI)* 11.4 (2022), pp. 1–19.
- [84] Gerłowska, J., Skrobas, U., Grabowska-Aleksandrowicz, K., et al. "Assessment of perceived attractiveness, usability, and societal impact of a multimodal robotic assistant for aging patients with memory impairments". In: *Frontiers in neurology* 9 (2018), p. 392.
- [85] Gilbert, S., Harvey, H., Melvin, T., et al. "Large language model AI chatbots require approval as medical devices". In: *Nature Medicine* 29.10 (2023), pp. 2396–2398.
- [86] Gómez-Alonso, C. and Valls, A. "A similarity measure for sequences of categorical data based on the ordering of common elements". In: *International conference on modeling decisions for artificial intelligence*. Springer. 2008, pp. 134–145.
- [87] Góngora Alonso, S., Hamrioui, S., Torre Díez, I. de la, et al. "Social robots for people with aging and dementia: a systematic review of literature". In: *Telemedicine and e-Health* 25.7 (2019), pp. 533–540.
- [88] Goodglass, H., Kaplan, E., and Weintraub, S. *BDAE: The Boston diagnostic aphasia examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.
- [89] Grootendorst, M. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022).
- [90] Gul, A., Turner, L. D., and Fuentes, C. "An exploratory analysis of trust in socially assistive robot interactions with unpaid carers of older adults". In: *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE. 2024, pp. 1630–1637.
- [91] Haider, F., De La Fuente, S., and Luz, S. "An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech". In: *IEEE Journal of Selected Topics in Signal Processing* 14.2 (2019), pp. 272–281.
- [92] Hamming, R. W. "Error detecting and error correcting codes". In: *The Bell system technical journal* 29.2 (1950), pp. 147–160.
- [93] Harciarek, M. and Kertesz, A. "Primary progressive aphasia and their contribution to the contemporary knowledge about the brain-language relationship". In: *Neuropsychology Review* 21 (2011), pp. 271–287.
- [94] Harrer, S. "Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine". In: *EBioMedicine* 90 (2023).
- [95] Hebesberger, D., Koertner, T., Gisinger, C., et al. "A long-term autonomous robot at a care hospital: A mixed methods study on social acceptance and experiences of staff and older adults". In: *International Journal of Social Robotics* 9.3 (2017), pp. 417–429.
- [96] Heerink, M., Kröse, B., Evers, V., et al. "Influence of social presence on acceptance of an assistive social robot and screen agent by elderly users". In: *Advanced Robotics* 23.14 (2009), pp. 1909–1923.
- [97] Heerink, M., Kröse, B., Evers, V., et al. "Assessing acceptance of assistive social agent technology by older adults: the almere model". In: *International Journal of Social Robotics* (2010).
- [98] Heerink, M., Kröse, B., Wielinga, B., et al. "Enjoyment intention to use and actual use of a conversational robot by elderly people". In: *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. 2008, pp. 113–120.

- [99] Holden, R. J. "A simplified system usability scale (SUS) for cognitively impaired and older adults". In: *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*. Vol. 9. 1. SAGE Publications Sage CA: Los Angeles, CA. 2020, pp. 180–182.
- [100] Hong, L. and Davison, B. D. "Empirical study of topic modeling in twitter". In: *Proceedings of the first workshop on social media analytics*. 2010, pp. 80–88.
- [101] Huang, X., Ruan, W., Huang, W., et al. "A survey of safety and trustworthiness of large language models through the lens of verification and validation". In: *Artificial Intelligence Review* 57.7 (2024), p. 175.
- [102] Irfan, B., Kuoppamäki, S., and Skantze, G. "Recommendations for designing conversational companion robots with older adults through foundation models". In: *Frontiers in Robotics and AI* 11 (2024), p. 1363713.
- [103] Islam, N., Hashem, R., Gad, M., et al. "Accuracy of the Montreal Cognitive Assessment Tool for detecting mild cognitive impairment: a systematic review and meta-analysis". In: *Alzheimer's & Dementia* 19.7 (2023), pp. 3235–3243.
- [104] Jakob, D. "Voice controlled devices and older adults—A systematic literature review". In: *International Conference on Human-Computer Interaction*. Springer. 2022, pp. 175–200.
- [105] Jarrold, W., Peintner, B., Wilkins, D., et al. "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech". In: *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 2014, pp. 27–37.
- [106] Jensen, A. M., Chenery, H. J., and Copland, D. A. "A comparison of picture description abilities in individuals with vascular subcortical lesions and Huntington's disease". In: *Journal of Communication Disorders* 39.1 (2006), pp. 62–77.
- [107] Jin, Y., Cai, W., Chen, L., et al. "Exploring the Design of Generative AI in Supporting Music-based Reminiscence for Older Adults". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–17.
- [108] Jo, E., Epstein, D. A., Jung, H., et al. "Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–16.
- [109] Jokel, R., Seixas Lima, B., Fernandez, A., et al. "Language in amnesic mild cognitive impairment and dementia of Alzheimer's type: Quantitatively or qualitatively different?" In: *Dementia and Geriatric Cognitive Disorders Extra* 9.1 (2019), pp. 136–151.
- [110] Jones, C., Sung, B., and Moyle, W. "Assessing engagement in people with dementia: a new approach to assessment using video analysis". In: *Archives of psychiatric nursing* 29.6 (2015), pp. 377–382.
- [111] Kanoh, M., Oida, Y., Nomura, Y., et al. "Examination of practicability of communication robot-assisted activity program for elderly people". In: *Journal of Robotics and Mechatronics* 23.1 (2011), p. 3.
- [112] Kaufman, L. and Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.
- [113] Kenton, J. D. M.-W. C. and Toutanova, L. K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL-HLT*. 2019, pp. 4171–4186.
- [114] Kertesz, A. "The Western Aphasia Battery". In: (1982).

- [115] Khosla, R., Chu, M.-T., Khaksar, S. M. S., et al. "Engagement and experience of older people with socially assistive robots in home care". In: *Assistive Technology* 33.2 (2021), pp. 57–71.
- [116] Khosla, R., Nguyen, K., and Chu, M.-T. "Human robot engagement and acceptability in residential aged care". In: *International Journal of Human–Computer Interaction* 33.6 (2017), pp. 510–522.
- [117] Kidd, C. D. and Breazeal, C. "Effect of a robot on user perceptions". In: *2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)(IEEE Cat. No. 04CH37566)*. Vol. 4. IEEE. 2004, pp. 3559–3564.
- [118] Kim, S. and Choudhury, A. "Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study". In: *Computers in Human Behavior* 124 (2021), p. 106914.
- [119] Klee, M., Leist, A. K., Veldsman, M., et al. "Socioeconomic deprivation, genetic risk, and incident dementia". In: *American Journal of Preventive Medicine* 64.5 (2023), pp. 621–630.
- [120] Kobayashi, M., Kosugi, A., Takagi, H., et al. "Effects of age-related cognitive decline on elderly user interactions with voice-based dialogue systems". In: *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part IV* 17. Springer. 2019, pp. 53–74.
- [121] König, A., Satt, A., Sorin, A., et al. "Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people". In: *Current Alzheimer Research* 15.2 (2018), pp. 120–129.
- [122] König, A., Satt, A., Sorin, A., et al. "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease". In: *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 1.1 (2015), pp. 112–124.
- [123] Korchut, A., Szklener, S., Abdelnour, C., et al. "Challenges for service robots - Requirements of elderly adults with cognitive impairments". In: *Frontiers in neurology* 8 (2017), p. 228.
- [124] Kröger, J. L., Lutz, O. H.-M., and Raschke, P. "Privacy implications of voice and speech analysis–information disclosure by inference". In: *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers* 14 (2020), pp. 242–258.
- [125] Kubota, A., Cruz-Sandoval, D., Kim, S., et al. "Cognitively assistive robots at home: Hri design patterns for translational science". In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2022, pp. 53–62.
- [126] Lago, P., Jiménez-Guarín, C., and Roncancio, C. "Contextualized behavior patterns for change reasoning in Ambient Assisted Living: A formal model". In: *Expert Systems* 34.2 (2017), e12163.
- [127] Lanzi, A. M., Saylor, A. K., Fromm, D., et al. "DementiaBank: Theoretical rationale, protocol, and illustrative analyses". In: *American Journal of Speech-Language Pathology* 32.2 (2023), pp. 426–438.
- [128] Laranjo, L., Dunn, A. G., Tong, H. L., et al. "Conversational agents in healthcare: a systematic review". In: *Journal of the American Medical Informatics Association* 25.9 (2018), pp. 1248–1258.
- [129] Law, M., Sutherland, C., Ahn, H. S., et al. "Developing assistive robots for people with mild cognitive impairment and mild dementia: a qualitative study with older adults and experts in aged care". In: *BMJ open* 9.9 (2019), e031937.

- [130] Lawrence, H. R., Schneider, R. A., Rubin, S. B., et al. "The opportunities and risks of large language models in mental health". In: *JMIR Mental Health* 11.1 (2024), e59479.
- [131] Lee, E., Vesonder, G., and Wendel, E. "Eldercare robotics-Alexa". In: *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE. 2020, pp. 0820–0825.
- [132] Lee, H., Phatale, S., Mansoor, H., et al. "Rlaif: Scaling reinforcement learning from human feedback with AI feedback". In: *arXiv preprint arXiv:2309.00267* (2023).
- [133] Lee, H. R. and Šabanović, S. "Culturally variable preferences for robot design and use in South Korea, Turkey, and the United States". In: *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 2014, pp. 17–24.
- [134] Lee, H. R., Šabanović, S., Chang, W.-L., et al. "Steps toward participatory design of social robots: mutual learning with older adults with depression". In: *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. 2017, pp. 244–253.
- [135] Lee, H. R., Sun, F., Iqbal, T., et al. "Reimagining Robots for Dementia: From Robots for Care-receivers/giver to Robots for Carepartners". In: *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 2023, pp. 475–484.
- [136] Lee, J. and Melo, N. "Habit representation based on activity recognition". In: *Sensors* 20.7 (2020), p. 1928.
- [137] Lee, J., Lee, D., and Lee, J.-g. "Influence of rapport and social presence with an AI psychotherapy chatbot on users' self-disclosure". In: *International Journal of Human-Computer Interaction* 40.7 (2024), pp. 1620–1631.
- [138] Lee, J. C., Kim, S. J., Hong, S., et al. "Diagnosis of Alzheimer's disease utilizing amyloid and tau as fluid biomarkers". In: *Experimental & molecular medicine* 51.5 (2019), pp. 1–10.
- [139] Leite, I., Martinho, C., and Paiva, A. "Social robots for long-term interaction: a survey". In: *International Journal of Social Robotics* 5 (2013), pp. 291–308.
- [140] Levenshtein, V. I. et al. "Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet physics doklady*. Vol. 10. 8. Soviet Union. 1966, pp. 707–710.
- [141] Lewis, M., Sycara, K., and Walker, P. "The role of trust in human-robot interaction". In: *Foundations of trusted autonomy* (2018), pp. 135–159.
- [142] Li, C., Duan, Y., Wang, H., et al. "Enhancing topic modeling for short texts with auxiliary word embeddings". In: *ACM Transactions on Information Systems (TOIS)* 36.2 (2017), pp. 1–30.
- [143] Li, R., Wang, X., and Yu, H. "Two directions for clinical data generation with large language models: data-to-label and label-to-data". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*. Vol. 2023. NIH Public Access. 2023, p. 7129.
- [144] Lima, M. R. "Home integration of conversational robots to enhance ageing and dementia care". In: *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 2024, pp. 115–117.
- [145] Lima, M. R., Horrocks, S., Daniels, S., et al. "The role of conversational AI in ageing and dementia care at home: A participatory study". In: *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2023, pp. 571–578.

- [146] Lima, M. R., O'Connell, A., Zhou, F., et al. "Promoting cognitive health in elder care with large language model-powered socially assistive robots". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2025.
- [147] Lima, M. R., Soreq, E., Daniels, S., et al. "Conversational AI for monitoring behaviour and well-being of people living with dementia". In: *Alzheimer's & Dementia* 19 (2023), e082077.
- [148] Lima, M. R., Su, T., Jouaiti, M., et al. "Discovering behavioural patterns using conversational technology for in-home health and well-being monitoring". In: *IEEE Internet of Things Journal* (2023).
- [149] Lima, M. R., Wairagkar, M., Gupta, M., et al. "Conversational affective social robots for ageing and dementia support". In: *IEEE Transactions on Cognitive and Developmental Systems* 14.4 (2021), pp. 1378–1397.
- [150] Lima, M. R., Wairagkar, M., Natarajan, N., et al. "Robotic telemedicine for mental health: A multimodal approach to improve human-robot engagement". In: *Frontiers in Robotics and AI* 8 (2021), p. 618866.
- [151] Lin, H., Karjadi, C., Ang, T. F., et al. "Identification of digital voice biomarkers for cognitive health". In: *Exploration of medicine* 1 (2020), p. 406.
- [152] Liu, B., Tetteroo, D., Timmermans, A., et al. "Exploring Older Adults' Acceptance, Needs, and Design Requirements towards Applying Social Robots in a Rehabilitation Context". In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2022, pp. 1077–1084.
- [153] Livingston, G., Huntley, J., Liu, K. Y., et al. "Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission". In: *The Lancet* 404.10452 (2024), pp. 572–628.
- [154] Lotfi, A., Langensiepen, C., Mahmoud, S. M., et al. "Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behaviour". In: *Journal of ambient intelligence and humanized computing* 3.3 (2012), pp. 205–218.
- [155] Louie, W.-Y. G., Li, J., Vaquero, T., et al. "A focus group study on the design considerations and impressions of a socially assistive robot for long-term care". In: *The 23rd IEEE international symposium on robot and human interactive communication*. IEEE. 2014, pp. 237–242.
- [156] Lundberg, S. M., Nair, B., Vavilala, M. S., et al. "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery". In: *Nature Biomedical Engineering* 2.10 (2018), p. 749.
- [157] Luz, S., Haider, F., Fuente, S. de la, et al. "Detecting cognitive decline using speech only: The addresso challenge". In: *arXiv preprint arXiv:2104.09356* (2021).
- [158] Ma, Z., Mei, Y., and Su, Z. "Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support". In: *AMIA Annual Symposium Proceedings*. Vol. 2023. American Medical Informatics Association. 2023, p. 1105.
- [159] Mahmoudi Asl, A., Molinari Ulate, M., Franco Martin, M., et al. "Methodologies used to study the feasibility, usability, efficacy, and effectiveness of social robots for elderly adults: scoping review". In: *Journal of medical Internet research* 24.8 (2022), e37434.
- [160] Malde, K., Coward, E., and Jonassen, I. "Fast sequence clustering using a suffix array algorithm". In: *Bioinformatics* 19.10 (2003), pp. 1221–1226.
- [161] Mantaci, S., Restivo, A., and Sciortino, M. "Distance measures for biological sequences: Some recent approaches". In: *International Journal of Approximate Reasoning* 47.1 (2008), pp. 109–124.

- [162] Marge, M., Espy-Wilson, C., Ward, N. G., et al. "Spoken language interaction with robots: Recommendations for future research". In: *Computer Speech & Language* 71 (2022), p. 101255.
- [163] Martín, F., Agüero, C. E., Cañas, J. M., et al. "Robotherapy with dementia patients". In: *International Journal of Advanced Robotic Systems* 10.1 (2013), p. 10.
- [164] Martinengo, L., Lum, E., and Car, J. "Evaluation of chatbot-delivered interventions for self-management of depression: Content analysis". In: *Journal of affective disorders* 319 (2022), pp. 598–607.
- [165] Matarić, M. J. and Scassellati, B. "Socially assistive robotics". In: *Springer handbook of robotics* (2016), pp. 1973–1994.
- [166] McColl, D., Louie, W.-Y. G., and Nejat, G. "Brian 2.1: A socially assistive robot for the elderly and cognitively impaired". In: *IEEE Robotics & Automation Magazine* 20.1 (2013), pp. 74–83.
- [167] McDonnell, M., Dill, L., Panos, S., et al. "Verbal fluency as a screening tool for mild cognitive impairment". In: *International Psychogeriatrics* 32.9 (2020), pp. 1055–1062.
- [168] Meng, L., Miao, C., and Leung, C. "Towards online and personalized daily activity recognition, habit modeling, and anomaly detection for the solitary elderly through unobtrusive sensing". In: *Multimedia Tools and Applications* 76.8 (2017), pp. 10779–10799.
- [169] Moharana, S., Panduro, A. E., Lee, H. R., et al. "Robots for joy, robots for sorrow: community based robot design for dementia caregivers". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 458–467.
- [170] Moyle, W. "The promise of technology in the future of dementia care". In: *Nature Reviews Neurology* 15.6 (2019), pp. 353–359.
- [171] Moyle, W., Jones, C. J., Murfield, J. E., et al. "Use of a robotic seal as a therapeutic tool to improve dementia symptoms: a cluster-randomized controlled trial". In: *Journal of the American Medical Directors Association* 18.9 (2017), pp. 766–773.
- [172] Moyle, W., Murfield, J., and Lion, K. "The effectiveness of smart home technologies to support the health outcomes of community-dwelling older adults living with dementia: A scoping review". In: *International Journal of Medical Informatics* 153 (2021), p. 104513.
- [173] Mullard, A. "FDA approves second anti-amyloid antibody for Alzheimer disease". In: *Nat Rev Drug Discov* 22.89 (2023), pp. 10–1038.
- [174] Naqvi, R. M., Haider, S., Tomlinson, G., et al. "Cognitive assessments in multicultural populations using the Rowland Universal Dementia Assessment Scale: a systematic review and meta-analysis". In: *Cmaj* 187.5 (2015), E169–E175.
- [175] Nasreen, S., Rohanian, M., Hough, J., et al. "Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features". In: *Frontiers in Computer Science* 3 (2021), p. 640669.
- [176] Natarajan, N., Vaitheswaran, S., Lima, M. R., et al. "Acceptability of social robots and adaptation of hybrid-face robot for dementia care in India: a qualitative study". In: *The American Journal of Geriatric Psychiatry* 30.2 (2022), pp. 240–245.
- [177] Newman, D., Bonilla, E. V., and Buntine, W. "Improving topic coherence with regularized topic models". In: *Advances in neural information processing systems* 24 (2011).

- [178] Nichols, E., Steinmetz, J. D., Vollset, S. E., et al. "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019". In: *The Lancet Public Health* 7.2 (2022).
- [179] Nishio, T., Yoshikawa, Y., Sakai, K., et al. "The effects of physically embodied multiple conversation robots on the elderly". In: *Frontiers in Robotics and AI* 8 (2021), p. 633045.
- [180] Novitzky, P., Smeaton, A. F., Chen, C., et al. "A review of contemporary work on the ethics of ambient assisted living technologies for people with dementia". In: *Science and engineering ethics* 21.3 (2015), pp. 707–765.
- [181] Ntracha, A., Iakovakis, D., Hadjidimitriou, S., et al. "Detection of mild cognitive impairment through natural language and touchscreen typing processing". In: *Frontiers in Digital Health* 2 (2020), p. 567158.
- [182] O'Connell, A., Banga, A., Ayissi, J., et al. "Design and Evaluation of a Socially Assistive Robot Schoolwork Companion for College Students with ADHD". In: *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 2024, pp. 533–541.
- [183] Omarov, B., Zhumanov, Z., Gumar, A., et al. "Artificial intelligence enabled mobile chatbot psychologist using AIML and cognitive behavioral therapy". In: *International Journal of Advanced Computer Science and Applications* 14.6 (2023).
- [184] Organization, W. H. et al. *Dementia: a public health priority*. World Health Organization, 2012.
- [185] Ostrowski, A. K., Breazeal, C., and Park, H. W. "Long-term co-design guidelines: empowering older adults as co-designers of social robots". In: *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE. 2021, pp. 1165–1172.
- [186] Ostrowski, A. K., DiPaola, D., Partridge, E., et al. "Older adults living with social robots: Promoting social connectedness in long-term communities". In: *IEEE Robotics & Automation Magazine* 26.2 (2019), pp. 59–70.
- [187] Ostrowski, A. K., Harrington, C. N., Breazeal, C., et al. "Personal narratives in technology design: the value of sharing older adults' stories in the design of social robots". In: *Frontiers in Robotics and AI* (2021), p. 315.
- [188] Ovadia, O., Brief, M., Mishaeli, M., et al. "Fine-tuning or retrieval? comparing knowledge injection in llms". In: *arXiv preprint arXiv:2312.05934* (2023).
- [189] Paletta, L., Fellner, M., Schüssler, S., et al. "AMIGO: Towards social robot based motivation for playful multimodal intervention in dementia". In: *Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference*. 2018, pp. 421–427.
- [190] Papadopoulos, I., Koulouglioti, C., Lazzarino, R., et al. "Enablers and barriers to the implementation of socially assistive humanoid robots in health and social care: a systematic review". In: *BMJ open* 10.1 (2020), e033096.
- [191] Patton, M. Q. "Qualitative research". In: *Encyclopedia of statistics in behavioral science* (2005).
- [192] Pelikan, H. and Hofstetter, E. "Managing delays in human-robot interaction". In: *ACM Transactions on Computer-Human Interaction* 30.4 (2023), pp. 1–42.
- [193] Pennebaker, J. W., Francis, M. E., and Booth, R. J. "Linguistic inquiry and word count: LIWC 2001". In: *Mahway: Lawrence Erlbaum Associates* 71.2001 (2001), p. 2001.

- [194] Petersen, S., Houston, S., Qin, H., et al. "The utilization of robotic pets in dementia care". In: *Journal of alzheimer's disease* 55.2 (2017), pp. 569–574.
- [195] Pino, O., Palestra, G., Trevino, R., et al. "The humanoid robot NAO as trainer in a memory program for elderly people with mild cognitive impairment". In: *International Journal of Social Robotics* 12 (2020), pp. 21–33.
- [196] Portugal, D., Alvito, P., Christodoulou, E., et al. "A study on the deployment of a service robot in an elderly care center". In: *International Journal of Social Robotics* 11 (2019), pp. 317–341.
- [197] Pou-Prom, C., Raimondo, S., and Rudzicz, F. "A conversational robot for older adults with Alzheimer's disease". In: *ACM Transactions on Human-Robot Interaction (THRI)* 9.3 (2020), pp. 1–25.
- [198] Price, S. E., Kinsella, G. J., Ong, B., et al. "Semantic verbal fluency strategies in amnesic mild cognitive impairment." In: *Neuropsychology* 26.4 (2012), p. 490.
- [199] Puehn, C. G., Liu, T., Feng, Y., et al. "Design of a low-cost social robot: Towards personalized human-robot interaction". In: *Universal Access in Human-Computer Interaction. Aging and Assistive Environments: 8th International Conference, UAHCI 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part III* 8. Springer. 2014, pp. 704–713.
- [200] Qian, K., Zhang, Z., Yamamoto, Y., et al. "Artificial intelligence internet of things for the elderly: from assisted living to health-care monitoring". In: *IEEE Signal Processing Magazine* 38.4 (2021), pp. 78–88.
- [201] Qiao, Y., Yin, X., Wiechmann, D., et al. "Alzheimer's Disease Detection from Spontaneous Speech Through Combining Linguistic Complexity and (Dis) Fluency Features with Pretrained Language Models". In: *arXiv preprint arXiv:2106.08689* (2021).
- [202] Radford, A., Kim, J. W., Xu, T., et al. "Robust speech recognition via large-scale weak supervision". In: *International conference on machine learning*. PMLR. 2023, pp. 28492–28518.
- [203] Radford, A., Narasimhan, K., Salimans, T., et al. "Improving language understanding by generative pre-training". In: (2018).
- [204] Raffel, C., Shazeer, N., Roberts, A., et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." In: *J. Mach. Learn. Res.* 21.140 (2020), pp. 1–67.
- [205] Rammstedt, B., Kemper, C. J., Klein, M. C., et al. "A short scale for assessing the big five dimensions of personality: 10 item big five inventory (BFI-10)". In: *methods, data, analyses* 7.2 (2013), p. 17.
- [206] Randall, N., Bennett, C. C., Šabanović, S., et al. "More than just friends: in-home use and design recommendations for sensing socially assistive robots (SARs) by older adults with depression". In: *Paladyn, Journal of Behavioral Robotics* 10.1 (2019), pp. 237–255.
- [207] Rashidi, P., Cook, D. J., Holder, L. B., et al. "Discovering activities to recognize and track in a smart environment". In: *IEEE transactions on knowledge and data engineering* 23.4 (2010), pp. 527–539.
- [208] Raz, D., Barkan-Slater, S., Baum-Cohen, I., et al. "A novel socially assistive robotic platform for cognitive-motor exercises for individuals with Parkinson's Disease: a participatory-design study from conception to feasibility testing with end users". In: *Frontiers in Robotics and AI* 10 (2023), p. 1267458.

- [209] Recchuto, C., Gava, L., Grassi, L., et al. "Cloud services for culture aware conversation: Socially assistive robots and virtual assistants". In: *2020 17th International Conference on Ubiquitous Robots (UR)*. IEEE. 2020, pp. 270–277.
- [210] Rehm, M. "From multicultural agents to culture-aware robots". In: *Human-Computer Interaction. Human-Centred Design Approaches, Methods, Tools, and Environments: 15th International Conference, HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part I 15*. Springer. 2013, pp. 431–440.
- [211] Reimers, N. and Gurevych, I. "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084* (2019).
- [212] Reis, A., Paulino, D., Paredes, H., et al. "Using intelligent personal assistants to assist the elderly: An evaluation of Amazon Alexa, Google Assistant, Microsoft Cortana, and Apple Siri". In: *2018 2nd international conference on technology and innovation in sports, health and wellbeing (TISHW)*. IEEE. 2018, pp. 1–5.
- [213] Ritschel, H., Janowski, K., Seiderer, A., et al. "Insights on usability and user feedback for an assistive robotic health companion with adaptive linguistic style". In: *Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments*. 2019, pp. 319–320.
- [214] Roark, B., Mitchell, M., Hosom, J.-P., et al. "Spoken language derived measures for detecting mild cognitive impairment". In: *IEEE transactions on audio, speech, and language processing* 19.7 (2011), pp. 2081–2090.
- [215] Rossi, S., Santangelo, G., Staffa, M., et al. "Psychometric evaluation supported by a social robot: Personality factors and technology acceptance". In: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2018, pp. 802–807.
- [216] Royal College of Psychiatrists. *National Audit of Dementia Care in General Hospitals 2018–19: Round Four Audit Report*. July 2019.
- [217] Rudzicz, F., Wang, R., Begum, M., et al. "Speech interaction with personal assistive robots supporting aging at home for individuals with Alzheimer's disease". In: *ACM Transactions on Accessible Computing (TACCESS)* 7.2 (2015), pp. 1–22.
- [218] Šabanović, S., Bennett, C. C., and Lee, H. R. "Towards culturally robust robots: A critical social perspective on robotics and culture". In: *Proc. HRI Workshop on Culture-Aware Robotics*. Vol. 2014. 2014.
- [219] Sabelli, A. M., Kanda, T., and Hagita, N. "A conversational robot in an elderly care center: an ethnographic study". In: *Proceedings of the 6th international conference on Human-robot interaction*. 2011, pp. 37–44.
- [220] Salichs, M. A., Castro-González, Á., Salichs, E., et al. "Mini: a new social robot for the elderly". In: *International Journal of Social Robotics* 12 (2020), pp. 1231–1249.
- [221] Salichs, M. A., Encinar, I. P., Salichs, E., et al. "Study of scenarios and technical requirements of a social assistive robot for Alzheimer's disease patients and their caregivers". In: *International Journal of Social Robotics* 8 (2016), pp. 85–102.
- [222] Salis, F., Costaggu, D., and Mandas, A. "Mini-mental state examination: Optimal cut-off levels for mild and severe cognitive impairment". In: *Geriatrics* 8.1 (2023), p. 12.
- [223] Saneifar, H., Bringay, S., Laurent, A., et al. "S2mp: Similarity measure for sequential patterns". In: *AusDM: Australasian Data Mining*. Vol. 87. ACS. 2008, pp. 095–104.
- [224] Sawami, K., Kimura, M., Kitamura, T., et al. "Cognitive ability and psychological effectiveness of brain training dance robot therapy for elderly people". In: *OA J Neuropsychiatry* 10 (2019).

- [225] Scheltens, P., Blennow, K., Breteler, M. M., et al. "Alzheimer's disease". In: *The Lancet* 388.10043 (2016), pp. 505–517.
- [226] Schroeter, C., Mueller, S., Volkhardt, M., et al. "Realization and user evaluation of a companion robot for people with mild cognitive impairments". In: *2013 IEEE International Conference on robotics and automation*. IEEE. 2013, pp. 1153–1159.
- [227] Sciuto, A., Saini, A., Forlizzi, J., et al. "'Hey Alexa, what's up?'" A mixed-methods studies of in-home conversational agent usage". In: *Proceedings of the 2018 designing interactive systems conference*. 2018, pp. 857–868.
- [228] Sepesy Maučec, M. and Donaj, G. "Discovering daily activity patterns from sensor data sequences and activity sequences". In: *Sensors* 21.20 (2021), p. 6920.
- [229] Seymour, W., Zhan, X., Cote, M., et al. "A systematic review of ethical concerns with voice assistants". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 2023, pp. 131–145.
- [230] Sezgin, E., Huang, Y., Ramtekkar, U., et al. "Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic". In: *NPJ Digital Medicine* 3.1 (2020), p. 122.
- [231] Shahid, Z. K., Saguna, S., and Åhlund, C. "Detecting Anomalies in Daily Activity Routines of Older Persons in Single Resident Smart Homes: Proof-of-Concept Study". In: *JMIR aging* 5.2 (2022), e28260.
- [232] Shi, Z., O'Connell, A., Li, Z., et al. "Build Your Own Robot Friend: An Open-Source Learning Module for Accessible and Engaging AI Education". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 21. 2024, pp. 23137–23145.
- [233] Sims, J. R., Zimmer, J. A., Evans, C. D., et al. "Donanemab in early symptomatic Alzheimer disease: the TRAILBLAZER-ALZ 2 randomized clinical trial". In: *Jama* 330.6 (2023), pp. 512–527.
- [234] So, J., Estrada, S., Jörke, M., et al. "'They Make Us Old Before We're Old': Designing Ethical Health Technology with and for Older Adults". In: *Proceedings of the ACM on Human-Computer Interaction* 8.CSCW2 (2024), pp. 1–30.
- [235] Spinuzzi, C. "The methodology of participatory design". In: *Technical communication* 52.2 (2005), pp. 163–174.
- [236] Studer, M. and Ritschard, G. "What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179.2 (2016), pp. 481–511.
- [237] Sung, J., Christensen, H. I., and Grinter, R. E. "Robots in the wild: understanding long-term use". In: *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 2009, pp. 45–52.
- [238] Tan, K., Sekhar, K., Wong, J., et al. "Alexa eldercare toolbox: A smarthome solution for the elderly". In: *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE. 2020, pp. 0806–0812.
- [239] Tanaka, M., Ishii, A., Yamano, E., et al. "Effect of a human-type communication robot on cognitive function in elderly women living alone". In: *Medical science monitor: international medical journal of experimental and clinical research* 18.9 (2012), CR550.
- [240] Tao, Y., Viberg, O., Baker, R. S., et al. "Cultural bias and cultural alignment of large language models". In: *PNAS nexus* 3.9 (2024), pgae346.
- [241] Tapus, A., Maja, M., and Scassellatti, B. "The grand challenges in socially assistive robotics". In: *IEEE Robotics and Automation Magazine* 14.1 (2007), N–A.

- [242] Tapus, A., Tapus, C., and Mataric, M. J. "The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia". In: *2009 IEEE international conference on rehabilitation robotics*. IEEE. 2009, pp. 924–929.
- [243] Tellex, S., Gopalan, N., Kress-Gazit, H., et al. "Robots that use language". In: *Annual Review of Control, Robotics, and Autonomous Systems* 3 (2020), pp. 25–55.
- [244] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., et al. "Large language models in medicine". In: *Nature medicine* 29.8 (2023), pp. 1930–1940.
- [245] Thomas, J. A., Burkhardt, H. A., Chaudhry, S., et al. "Assessing the utility of language and voice biomarkers to predict cognitive impairment in the Framingham Heart Study Cognitive Aging Cohort Data". In: *Journal of Alzheimer's Disease* 76.3 (2020), pp. 905–922.
- [246] Torta, E., Werner, F., Johnson, D. O., et al. "Evaluation of a small socially-assistive humanoid robot in intelligent homes for the care of the elderly". In: *Journal of Intelligent & Robotic Systems* 76 (2014), pp. 57–71.
- [247] Tóth, L., Hoffmann, I., Gosztolya, G., et al. "A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech". In: *Current Alzheimer Research* 15.2 (2018), pp. 130–138.
- [248] Trajkova, M. and Martin-Hammond, A. "'Alexa is a toy': exploring older adults' reasons for using, limiting, and abandoning echo". In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.
- [249] Treder, M. S., Lee, S., and Tsvetanov, K. A. "Introduction to Large Language Models (LLMs) for dementia care and research". In: *Frontiers in Dementia* 3 (2024), p. 1385303.
- [250] UK National Institute for Health and Care Excellence. *Donepezil, galantamine, rivastigmine and memantine for the treatment of Alzheimer's disease*. 2017.
- [251] UK National Institute for Health Research (NIHR). *A brief guide to patient and public involvement and qualitative methods in health and social care research*. Mar. 2023.
- [252] United Nations. *World Social Report 2023: Leaving No One Behind In An Ageing World*. Jan. 2023.
- [253] United Nations. *World Population Prospects 2022*. Sept. 2024.
- [254] Van Dyck, C. H., Swanson, C. J., Aisen, P., et al. "Lecanemab in early Alzheimer's disease". In: *New England Journal of Medicine* 388.1 (2023), pp. 9–21.
- [255] Van Maris, A., Zook, N., Caleb-Solly, P., et al. "Designing ethical social robots—a longitudinal field study with older adults". In: *Frontiers in Robotics and AI* 7 (2020), p. 1.
- [256] Van Patten, R., Keller, A. V., Maye, J. E., et al. "Home-based cognitively assistive robots: maximizing cognitive functioning and maintaining independence in older adults without dementia". In: *Clinical interventions in aging* (2020), pp. 1129–1139.
- [257] Vasco, V., Willemse, C., Chevalier, P., et al. "Train with me: a study comparing a socially assistive robot and a virtual agent for a rehabilitation task". In: *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11*. Springer. 2019, pp. 453–463.
- [258] Verma, M. and Howard, R. J. "Semantic memory and language dysfunction in early Alzheimer's disease: a review". In: *International journal of geriatric psychiatry* 27.12 (2012), pp. 1209–1217.
- [259] Vogan, A. A., Alnajjar, F., Gochoo, M., et al. "Robots, AI, and cognitive training in an era of mass age-related cognitive decline: a systematic review". In: *Ieee Access* 8 (2020), pp. 18284–18304.

- [260] Wada, K. and Shibata, T. "Living with seal robots—its sociopsychological and physiological influences on the elderly at a care house". In: *IEEE transactions on robotics* 23.5 (2007), pp. 972–980.
- [261] Wainer, J., Feil-Seifer, D. J., Shell, D. A., et al. "The role of physical embodiment in human-robot interaction". In: *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2006, pp. 117–122.
- [262] Wairagkar, M., Lima, M. R., Bazo, D., et al. "Emotive response to a hybrid-face robot and translation to consumer social robots". In: *IEEE Internet of Things Journal* 9.5 (2021), pp. 3174–3188.
- [263] Wallerstein, N. B. and Duran, B. "Using community-based participatory research to address health disparities". In: *Health promotion practice* 7.3 (2006), pp. 312–323.
- [264] Wang, N., Goel, S., Ibrahim, S., et al. "Decoding loneliness: Can explainable AI help in understanding language differences in lonely older adults?" In: *Psychiatry research* 339 (2024), p. 116078.
- [265] Wang, R. H., Sudhama, A., Begum, M., et al. "Robots to assist daily activities: views of older adults with Alzheimer's disease and their caregivers". In: *International psychogeriatrics* 29.1 (2017), pp. 67–79.
- [266] Wei, H., He, S., Xia, T., et al. "Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates". In: *arXiv preprint arXiv:2408.13006* (2024).
- [267] Williams, E., Theys, C., and McAuliffe, M. "Lexical-semantic properties of verbs and nouns used in conversation by people with Alzheimer's disease". In: *PLoS One* 18.8 (2023), e0288556.
- [268] World Health Organization. *Dementia key facts*. <https://www.who.int/news-room/fact-sheets/detail/dementia>. Mar. 2023.
- [269] World Health Organization. *Ageing and health*. <https://www.who.int/news-room/fact-sheets/detail/ageing-and-health>. Sept. 2024.
- [270] Xu, X., Yao, B., Dong, Y., et al. "Mental-llm: Leveraging large language models for mental health prediction via online text data". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8.1 (2024), pp. 1–32.
- [271] Yamada, Y., Shinkawa, K., Nemoto, M., et al. "A mobile application using automatic speech analysis for classifying Alzheimer's disease and mild cognitive impairment". In: *Computer Speech & Language* 81 (2023), p. 101514.
- [272] Yang, E., Amar, J., Lee, J. H., et al. "The Geometry of Queries: Query-Based Innovations in Retrieval-Augmented Generation". In: *arXiv preprint arXiv:2407.18044* (2024).
- [273] Yang, E., Garcia, T., Williams, H., et al. "From Barriers to Tactics: A Behavioral Science-Informed Agentic Workflow for Personalized Nutrition Coaching". In: *arXiv preprint arXiv:2410.14041* (2024).
- [274] Yang, Z., Xu, X., Yao, B., et al. "Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8.2 (2024), pp. 1–35.
- [275] Ye, Y., You, H., and Du, J. "Improved trust in human-robot collaboration with ChatGPT". In: *IEEE Access* (2023).
- [276] Yin, J., Zhang, Q., and Karunanithi, M. "Unsupervised daily routine and activity discovery in smart homes". In: *2015 37th Annual International Conference of the*

- IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2015, pp. 5497–5500.
- [277] Yoon, J. W., Woo, B. J., and Kim, N. S. "Hubert-ee: Early exiting hubert for efficient speech recognition". In: *arXiv preprint arXiv:2204.06328* (2022).
- [278] Zekri, D., Delot, T., Thilliez, M., et al. "A framework for detecting and analyzing behavior changes of elderly people over time using learning techniques". In: *Sensors* 20.24 (2020), p. 7112.
- [279] Zhang, L., Negrinho, R., Ghosh, A., et al. "Leveraging pretrained models for automatic summarization of doctor-patient conversations". In: *arXiv preprint arXiv:2109.12174* (2021).
- [280] Zhao, J., Zhang, Z., Gao, L., et al. "Llama beyond english: An empirical study on language capability transfer". In: *arXiv preprint arXiv:2401.01055* (2024).
- [281] Zolnoori, M., Zolnour, A., and Topaz, M. "ADscreen: A speech processing-based screening system for automatic identification of patients with Alzheimer's disease and related dementia". In: *Artificial Intelligence in Medicine* 143 (2023), p. 102624.

Appendix

A1: Key Studies Reviewed

A2: In-Home Behavioural Analysis

A3: Picture Description Cognitive Task

A4: LLM-Powered Robot-Administered Cognitive tasks

A5: Ethics Approval IRB

A1 Key Studies Reviewed

This section details comprehensively the key human-robot interaction studies that have designed, evaluated, or tested different SARs platforms with speech-based interfaces targeting older adults and individuals with cognitive impairment.

Table A1.1 summarises the main research contributions, domain applications, and limitations of representative HRI studies. With the aim of identifying open challenges and proposing future directions, this review evaluated the following key design and functional characteristics of SAR:

- *User-centred design.* Has the robot followed a user-centred design, in that has it been designed specifically to support the target population (older adults and PLWD), or have end-users been involved in the design process of the social robot? (Q1);
- *Personalisation.* Does the robot include some form of personalization to user profiles or preferences? (Q2);
- *Conversational ability.* Is the robot capable of two-way autonomous conversations? (Q3);
- *Accessibility.* Does the robot demonstrate a design trade-off in its mechatronic components, in that have cost and ease of use been considered in the design process? This would facilitate worldwide deployment and scalability. It should be noted the development of simpler robotic platforms must not compromise the desired affective and functional abilities of SARs (Q4).

Table A1.1: Summary of key HRI studies reviewed. Studies using distinct conversational SARs to support ageing and dementia care were selected. Applied roles for mental health support were evaluated. Note that robots may include additional roles. ss: session; wk: week; HS: healthy seniors; NN: not known; NA: not applied; Q1: User-centred design; Q2: Personalized behaviour; Q3: Autonomous conversations; Q4: Design trade-off for scalability.

Author (year)	Robot, role*, target users	Participants, setting (country), duration	Methods	Outcomes	Limitations	Q1	Q2	Q3	Q4
Cruz-Sandoval et al. (2020)	- Eva, small - Cognitive stimulation, clinical therapy - PLWD	- n=9 PLWD - nursing home (Mexico) - 2 ss/wk for 9 wk	Neuropsychiatric Inventory Nursing Home version (NPI-NH) to analyse dementia-related symptoms, pre- and post-sessions. Semi-structured interviews with caregivers.	Improvement in NPI-NH scores: delusions, agitation, and exaltation. Positive short-term effect in mood and behaviour. Potential to reduce the caregiver burden.	Small sample. Script followed in sessions. Lack of integrated emotional response. Automatic change of facial expressions not reported. No objective evidence of effective speech interfaces.	✓	✓	✓	✓
Do et al. (2020)	- Robot for clinical screening, small - Health monitoring, cognitive stimulation - Older adults in risk of dementia	- n=30 - setting NN (USA) - single 50 min session	Robot conducted clinical screening interviews. Answers evaluated by both the robot and a practitioner. User experience post-survey, assessed trust in the robot.	Robot was effective in performing geriatric wellbeing assessment, trusted by users. High scoring performance (90% for spelling responses, 93% for other).	Task-oriented, not personalized to user profile over time. Unclear whether the framework would be useful for frequent evaluation of wellbeing and cognitive abilities. Errors in speech recognition.	✓	✓	✓	✓
Pino et al. (2020)	- NAO, small humanoid - Cognitive stimulation, clinical therapy - Older people with MCI	- n=21 MCI - centre for the treatment of cognitive disorders and dementia (Italy) - 1 ss/wk (90min) for 8 wk	Memory-training protocols used in sessions with psychologist, with and without the robot. Analysis of video-recorded sessions (e.g., face detection, visual attention).	Robot-assisted activities increased attention, reduced depressive symptoms. Changes in prose memory and verbal fluency detected. Potential usefulness to support clinical practice.	Robot was not always autonomous. Errors in speech recognition software and lack of naturalness; limited verbal interactions. User needs to wait for 'beep' to reply. Robot's behaviour did not vary between sessions.	✓	✓	✓	✓
Pou-prom et al. (2020)	- Milo, small humanoid - Cognitive stimulation - PLWD	- n=19 PLWD - setting NN (Canada) - 3 ss over 1 year	Picture description task with human, teleoperated and autonomous robot. Kinect sensor for facial features analysis. Linguistic analysis from transcripts. Pre- and post-questionnaires.	Proof of concept of automatic linguistic analysis to monitor dementia. Richest conversations with users with milder cognitive impairment. Autonomous robot unable to recover from conversation breakdowns.	The autonomous robot lacked prosodic, gestural and facial cues. Rule-based dialogue. Limited quality of automatic speech recognition software to target responses of older citizens. Exploratory study with no long-term results.	✓	✓	✓	✓
Salichs et al. (2020)	- Mini, small - Companionship, cognitive stimulation - Older adults, MCI, caregivers	- n=20 HS, carers, relatives - nursing home (Spain) - 2 months	Assess perception of the robot: usability, appearance, satisfactions through a post qualitative questionnaire.	End-users perceived it as a useful tool to motivate towards daily activities but did not recognise the potential to help extend their autonomy.	Robot only follows user requests, does not lead the interaction. No implemented ability to adapt behaviour to user profiles. Results exclusively based on questionnaire data.	✓	✓	✓	✓
Iio et al. (2020)	- CommU twin-robot system, small - Companionship - Older adults, PLWD	- n=30 (13 HS, 4 mild dementia, 13 advanced dementia) - nursing home (Japan) - single ss <30min	Investigate ability to sustain a coherent dialogue despite errors in speech recognition. Word error rate, total dialogue time, user utterance time. Post-questionnaire for user and caregiver impression.	Despite speech recognition errors system maintained dialogue for almost 13min on average. No significant difference in one vs. two-robot scenario on user experience, however twin-robot system may encourage longer dialogue.	Required human supervision to force-quit the program when asked by participants, but speech recognition errors often continued dialogue inappropriately. Lack of continuous sessions, user experience may be biased by novelty effect.	✓	✓	✓	✓
McGinn et al. (2020)	- Stevie, mobile - Companionship, cognitive stimulation - Older adults, caregivers	- n=15 (10 HS, 5 care staff) - care facilities (USA, Ireland) - focus group ss	Applied user-centred design methods to evaluate a SAR prototype. Investigate impression and usefulness through focus groups; determine most suitable voice for robot. Post-questionnaires.	Iterative design process led to successful SAR prototype. Overall positive impression. Concerns raised regarding use by older adults, including fear of stigmatization.	Robot was fully teleoperated in introduction/focus group sessions. Prototype qualitative evaluation, lack of human-robot testing. No objective evidence of effective speech interfaces.	✓	✓	✓	✓

Author (year)	Robot, role*, target users	Participants, setting (country), duration	Methods	Outcomes	Limitations	Q1	Q2	Q3	Q4
Law et al. (2019)	- Silbot, mobile - Companionship, health monitoring, cognitive stimulation - MCI, mild dementia	- 9 HS, 14 experts in aged care - aged-care village and laboratory (New Zealand) - single group ss	Investigate the perceived usefulness through qualitative interviews. Videos of robot prototype and interactions.	Robot-assisted activities perceived as useful, especially reminders and safety checks. Potential to reduce loneliness and caregiver burden. Consider familiarity with technology.	Robot not tested with MCI individuals. Lack of personalization. Issues with speech recognition: little time for user response. WoZ required if user speech was not understood.	✓	X	X	X
Portugal et al. (2019)	- GrowMu, mobile - Companionship - Older adults	- estimated n=100 HS, caregivers, visitors - senior care centre (The Netherlands) - 1 wk, interactions of up to 5min	Usability and acceptability study. Robot would detect presence of people, greet them and propose activities. Questionnaires for qualitative feedback.	Robot perceived as friendly and accessible. Network connection issues reported. Limited set of words recognized (65.93% accuracy). APIs mentioned as potential solution.	Feedback from target users unspecified. Limited speech interfaces. Use of predefined TTS recordings. Lack of integration between robot's expressions and speech input/output.	✓	✓	X	✓
Ritschel et al. (2019)	- Reeti, small - Companionship - Older adults	- n=2 HS - users' homes (Germany) - 1 wk, free interactions	Field usability study. Interaction logs to analyse the robot's adaptation progress. Pre- and post- questionnaires.	Robot perceived as easy to use and attractive. Users mentioned the usefulness of adding a 2-way speech interface, for which privacy concerns were raised.	Limited sample. Lack of speech recognition. Additional hardware with physical buttons needed for human input may turn it less practical for home-based use.	X	✓	X	X
Sawami et al. (2019)	- Sota, small - Cognitive stimulation, clinical therapy - Older adults, PLWD	- n=71 - setting NN (Japan) - 1 therapy ss/wk (90min) for 7 wk	Feasibility study. Robot-assisted memory and dance activities. Comparison of pre- and post-effects using a cognitive evaluation scale. Self-report.	Cognitive dance therapy can improve cognitive abilities. Improvement in self-reported mood, satisfaction, vitality. Correlation between cognitive functioning and mental state.	Lack of validated neuropsychological outcome measures. No follow-up evaluation. No evidence of effectiveness of speech interfaces. Unclear whether the robot operates autonomously.	X	X	NN	X
Yamazaki et al. (2019)	- Telenoid, small, humanoid - Companionship - PLWD	- n=5 PLWD - nursing home (Japan) - 3 ss/wk (20min) for 10 wk	Exploratory field trial. NPI-NH measures to assess dementia symptoms over time.	Verified tendency to improve behavioural and psychological symptoms of dementia, including anxiety and appetite.	Limited sample size. System required teleoperation. Lack of personalized conversation topics. Further control groups needed.	X	✓	✓	X
Dino et al. (2019)	- Ryan, small humanoid (torso) - Companionship, cognitive stimulation, clinical therapy - Older adults, MCI, PLWD	- n=4 older adults with mild-to-moderate depression - senior living facility (USA) - 2 ss/wk (1h) for 4 wk	Pre- and post- mental health evaluation tests. Dialogues were transcribed. Face scale for mood evaluation. Qualitative interview after last session.	Robot able to administer internet-delivered cognitive behaviour therapy. Average sentence length increased over sessions. Overall improvement of patient mood.	Limited sample to generalize results. Dialogue engine based on graph structure may limit naturalness and flexibility of HRI. Use of a finite-state automaton to frame sessions.	X	✓	✓	X
Khosla et al. (2019)	- PaPeRo platform ('Betty', 'Matilda'), small - Companionship - PLWD	- n=5 PLWD - users' homes (Australia) - 3 months	Investigate user engagement and experience. Mixed-method: observations, log data, video recordings, engagement measures, post-trial user experience questionnaires.	Verified usefulness of social robot for home-based engagement with PLWD. Human-robot engagement average of 24-36 min daily.	Limited sample size. Verbal HRI analysis excluded due to privacy concerns. Further research needed to explore long-term adoption and behaviour changes. Autonomous conversational ability unspecified.	X	✓	NN	X
Bajones et al. (2018)	- Hobbit, mobile - Companionship - Older adults	- n=18 HS - users' homes (Austria, Greece, Sweden) - 3 wk	Investigate usefulness in a field trial with autonomous robot. Qualitative interviews and questionnaires in different stages of trials. Log data recorded by the robot.	Perceived usefulness in real homes. Potential to support independent living. Several functions lacked stability over time due to technical issues.	Although conversation was not the main function of robot, speech recognition did not work well for many users.	✓	✓	NN	✓
Caleb-Solly et al. (2018)	- Kompai, mobile - Companionship, cognitive stimulation, health monitoring, clinical therapy - Older adults, PLWD	- n=11 - lab trials (n=8), residential care (n=1), homes (n=2) (UK, The Netherlands) - 2 days	Usability and user experience evaluation. Robot teleoperated. Post-task questionnaires. Behaviour and speech recorded with robot's camera.	A phased introduction of the robot enhances acceptance. User adaptation in face of robot's constraints may lead to more engaging HRI.	Poor speech recognition and issue with the robot's listening mode. Technical instability and lack of autonomy prevented long-term testing.	✓	X	NN	X

Author (year)	Robot, role*, target users	Participants, setting (country), duration	Methods	Outcomes	Limitations	Q1	Q2	Q3	Q4
Di Nuovo et al. (2018)	- Doro Robot-Era, mobile - Companionship - Older adults	- n=72 HS - laboratory (UK) - single individual ss	Usability and acceptability study: free HRI. Experience food delivery task. Post-session questionnaires. Video recordings for behaviour analysis, e.g., user attention.	Overall acceptance and positive impression. Multimodality is key for acceptance and adoption, particularly for users less experienced with technology.	Pre-defined interactive scenarios instead of spontaneous. Issues with speech interface; understands simple commands only; restricted to task-directed dialogue, e.g., food delivery.	✓	X	✓	X
Gerlowska et al. (2018)	- RAMCIP, mobile - Companionship, health monitoring - Older adults, MCI	- n=18 (n=10 HS, n=8 MCI) - laboratory (Poland) - single individual ss	Usability and acceptability study. Participants experience pre-defined use case scenarios, e.g., medication intake, social interaction. Post-task questionnaire.	Overall acceptance and likeness. Neutral feedback on usability: longer HRI needed to verify the robot's real value. No significant differences between healthy and MCI groups.	Results may be biased by the group size effect. Unspecified effectiveness and autonomous ability of the robot's speech interfaces.	✓	X	NN	X
Paletta et al. (2018)	- Pepper, mobile humanoid - Companionship, health monitoring, cognitive stimulation - PLWD	- n=6 PLWD, other caregivers, relatives - setting: NA (Austria) - Focus groups	Investigate attitudes of end-users, caregivers and relatives about robot utility. Qualitative feedback from 6 individual interviews, 3 focus groups.	Overall user interest but scepticism about the use of social robots. Communication and avoiding danger were the main HRI applications identified.	Initial evaluation study in the AMIGO project, not yet tested with end-users in real setting. No details about the system's feature to adapt communication and interaction.	X	X	NN	X
Hebesberger et al. (2017)	- SCITOS, mobile - Companionship - PLWD, aged care staff	- n=70 volunteers, staff - care-hospital (Austria) - 15-day trial, 5-day pilot study	Field acceptance and experience study. Qualitative mixed-method: observation (12h), interviews (n=10), questionnaires (n=70).	Perceived utility of robot depends on proper functioning. User frustration when failing to start a conversation with robot. Reluctance of staff to share workspace with a robot.	Lack of speech recognition engine, which the authors state as a fundamental feature to develop. Robot with limited affective capacity. Lack of validated measure of acceptance.	X	X	X	X
Rudczicz et al. (2015)	- ED, mobile - NA (activities of daily living) - PLWD	- n=10 older adults with Alzheimer's - simulated home environment (Canada) - single ss 2,5h	Examine aspects of PLWD-robot conversations and efficacy of speech recognition. Post-interviews. Explored relation between speech recognition accuracy and mini-mental state exam (MMSE) scores.	Speech recognition remains a challenge with PLWD. Importance of conversation repair features to overcome user confusion or lack of interest in verbal HRI.	Lack of autonomous functionality; robot teleoperated during tasks, including to trigger conversations. Very simple spoken prompts were used. Lack of expressiveness in voice.	✓	X	X	X
Broadbent et al. (2014)	- IroblQ, small - Companionship, health monitoring, cognitive stimulation - Older adults	- n=29 HS - users' homes (New Zealand) - 6 wk	Randomised trial (robot and non-robot groups). Geriatric outcome measures for quality of life and depression. Medication Adherence Report Scale for self-reported adherence. Interviews, questionnaires and log data.	Robot was acceptable by end-users and feasible to deploy in the retirement village; no significant improvement in quality of life or medication adherence verified.	Technical issues reported required a researcher to go to trial location. Unspecified 2-way autonomous conversational capacity. User input mainly via touch screen.	X	X	NN	X
Schroeter et al. (2013)	- CompanionAble, mobile - Companionship, cognitive stimulation, health monitoring - PLWD	- n=11 (5 PLWD-caregiver couples, 1 single PLWD) - simulated home (The Netherlands) - 2 days	User experience studies. Qualitative semi-structured interviews questionnaires, daily observation sheets.	Robot perceived as useful and enjoyable. Benefits of reduced caregiver burden. Robot initiative (navigating to users, giving reminders) considered the most useful feature.	Speech recognition engine had to be deactivated by the second trial day given poor performance. Longer term effects on wellbeing beyond trials were not evaluated. Small sample and short-term study.	✓	X	X	X
McColl et al. (2013)	- Brian 2.1, small humanoid (torso) - Cognitive stimulation - Older people, MCI, PLWD	- n=40 - care facility (Canada) - robot placed in facility for 2 days	Acceptability study. Robot-assisted activities: eating, memory card game. Observed user engagement (frequency, attention), compliance (cooperative behaviours towards the robot). Post-questionnaires.	Average length of interactions of 13 min. Promise for robot-assisted cognitive interventions. Overall engagement. Ability to display emotions through facial expressions and tone of voice was the most liked characteristic.	Lack of long-term testing. Unclear whether speech interfaces worked smoothly during sessions. Lags between user and robot's reply were reported. Lack of integration of facial expressiveness with speech.	NN	✓	NN	X

Author (year)	Robot, role*, target users	Participants, setting (country), duration	Methods	Outcomes	Limitations	Q1	Q2	Q3	Q4
Tanaka et al. (2012)	- Nodding Kabochan, small, toy-like - Companionship; Cognitive stimulation - Older adults	- n=34, all women living alone (robot group n=18, control group with toy n=16) - users' homes (Japan) - 8 wk	Investigate effects on cognitive function. Randomized controlled trial. Evaluation of MMSE, blood and saliva samples, subjective fatigue, motivation. Questionnaires.	For the robot group, cognitive scores of MMSE increased after trials; decreased saliva level, loss of fatigue and enhanced motivation were reported with the robot.	Tested with women only. Robot with static face/emotion. Unclear whether the robot recognizes human speech. Unspecified robot autonomy.	NN	X	X	✓
Sabelli et al. (2011)	- Robovie, mobile humanoid - Companionship - Older adults	- n=55 HS - senior care centre (Japan) - 1 or 2 ss/wk for 3.5 months	Observed interactions. Semi-structured interviews. Collected transcripts of conversations (conversation topics, information shared with robot, emotional involvement).	Overall acceptance. Daily greetings and being called by their name increased user responsiveness.	Robot was teleoperated. A set of conversations and gestures was pre-programmed prior to deployment. Unspecified effectiveness of the robot's speech interface.	X	X	X	X
Kanoh et al. (2011)	- Iibot, small, toy-like - Cognitive stimulation - Older adults	- n=10 older adults with cognitive impairments - health care facility (Japan) - duration NN	Feasibility study of robot-assisted activities. Video recorded sessions for behaviour assessment. Focus group interview.	Overall positive reaction to robot's activities (e.g., songs, quizzes). Little talk directed to the robot was reported.	Participants discontinued the use of this robot. Unspecified effectiveness of speech interface. Robot currently out of production due to poor sales.	✓	X	NN	X
Tapus et al. (2009)	- Bandit, mobile humanoid, (torso) - Cognitive stimulation, clinical therapy - MCI, PLWD	- n=10 MCI/PLWD - care facility (USA) - 1 ss/wk for 6 months	Longitudinal pilot study. Adaptive algorithm to adjust difficulty of game based on reaction time and number of correct answers; performance evaluated over time.	Participants were able to engage and retain attention. Adaptive behaviour to level of user disability maximized performance over time.	Small sample size. Robot uses the same pre-recorded human voice for all interactions with participants.	X	✓	X	X
Heerink et al. (2008)	- iCat, small - Companionship - Older adults	- n=30 HS - senior care facility (The Netherlands) - 3min ss, free public use for 5 days	Investigate acceptance of conversational robot. Video recorded sessions and collected amount of usage. Post-questionnaires.	Overall positive response. Perceived enjoyment should be part of a robot acceptance model for older adults as it has an effect on the intention to use a robotic system.	Robot used simulated conversational ability with pre-recorded speech for output. Unspecified effectiveness of speech interface.	X	X	NN	X
Pineau et al. (2003)	- Pearl, mobile - Companionship, Cognitive stimulation - Older adults, MCI	- n=6 - nursing home, USA - 5 days	Investigate feasibility and user attitudes. Robot leads the subject between locations in the facility and verbally interacts with reminders. Observation.	Overall positive response. Robot ability to remind and autonomously guide users stood out. Issues with speech recognition engine reported. Adaptive behaviour to accommodate each user's needs is imperative.	Study focused on the robot's functionalities, not on people's perceived acceptance/experience. Various adjustments were needed to accommodate user's walking speed, voice and auditory abilities.	✓	X	X	X

A2 In-Home Behavioural Analysis

Table A2.1: Comparison analysis of clustering methods. Silhouette score was used as a measure of coherence to compare different clustering methods with 16 clusters. The total number of utterances clustered accounts for the cohort of participants. Note a pre-trained language model was applied to embed each utterance into a 768-dimensional vector.

Clustering Method	Silhouette
K-Means	0.326
Hierarchical Clustering	0.317
Gaussian Mixture Models	0.308
Spectral Clustering	0.317
Total utterances	6965

Table A2.2: Distribution of topic triggers by time of day and novelty period. Distribution of relevant topics in the morning (before 12:00), afternoon (between 12:00 and 17:00 exclusive) and evening (from 17:00) slots for each participant, represented in percentage. The total counts of topic triggers are also shown for the two periods analysed. Note that numbers in bold are the ones used to describe the results.

PID	Topic	total	Novelty period			Post-novelty period			
			mor	aft	eve	total	mor	aft	eve
P2	Weather	25	68	12	20	190	85.3	10.5	4.2
P2	Timers	1	0	100	0	216	53.7	41.2	5.1
P6	Entertainment	9	55.6	44.4	0	27	51.6	44.4	3.7
P6	Reminders/Time/Date	12	66.7	25	8.3	19	52.6	42.1	5.3
P12	Weather	12	75	16.7	8.3	8	62.5	12.5	25
P14	News	4	50	0	50	9	66.7	33.3	0

Table A2.3: Cluster distributions within the selected time period. Supplementary information for each time period analysed in Figure 3.8, including: start and end dates; total weeks considered; distribution (%) of activity sequences across morning, afternoon, and evening; cluster distribution of sequences within the selected time period. Note that bold numbers highlight key results for each time period analysed.

Time Period			Time Slot (%)				Cluster in Period (%)			
ID	Start date	End date	wk	mor	aft	eve	cl_1	cl_2	cl_3	cl_4
X1	2022-04-03	2022-05-07	6	61.1	33.0	5.9	84.7	9.1	5.3	0.5
X2	2021-10-23	2021-11-13	4	94.7	5.3	0	5.6	16.7	77.8	NA
X3	2021-11-09	2022-01-21	11	100	0	0	1.9	22.2	11.1	64.8
X4	2022-01-25	2022-04-07	11	100	0	0	75	1.8	1.8	21.4

Table A2.4: Cluster distributions outside the selected time period. Supplementary information for each time period, with cluster distribution of activity sequences outside the selected time period. Note that bold numbers highlight key results for each time period analysed.

Time Period			Time Slot (%)				Cluster outside Period (%)			
ID	Start date	End date	wk	mor	aft	eve	cl_1	cl_2	cl_3	cl_4
X1	2022-04-03	2022-05-07	6	61.1	33.0	5.9	11.9	37.8	34.1	16.2
X2	2021-10-23	2021-11-13	4	94.7	5.3	0	34.8	43.9	21.3	NA
X3	2021-11-09	2022-01-21	11	100	0	0	57.4	9.0	4.1	29.5
X4	2022-01-25	2022-04-07	11	100	0	0	24.2	18.3	8.33	49.2

A3 Cultural Feasibility Study in India

Table A3.1: OME metrics used to assess engagement during HRI sessions.

OME metric	ID	Scale
Attention to stimulus during engagement	A1	7 very attentive 6 attentive 5 somewhat attentive 4 not attentive 3 somewhat disruptive 2 disruptive 1 very disruptive
Attitude to stimulus during engagement	A2	7 very positive 6 positive 5 somewhat positive 4 neutral 3 somewhat negative 2 negative 1 very negative
Cognitive difficulty with task	CD	5 very difficult 4 difficult 3 moderately difficult 2 slightly difficult 1 not at all difficult 0 could not determine
Disruptive frequency rate	D1	3 most or all the session 2 some of the session 1 a little (less than 20 sec) 0 none
Distraction frequency rate	D2	3 most or all the session 2 some of the session 1 a little (less than 20 sec) 0 none

Table A3.2: Engagement metrics assessed from video analysis. Engagement assessment from video analysis with the behaviours observed for each category (emotional, visual and verbal).

Engagement Metric	Behaviour Observation
Emotional engagement	
Positive	Smiling, laughing to robot. Greeting the robot with gestures. Overall positive attitude and facial expressions communicating with the robot.
Negative	Shows anger, anxiety or fear, or sadness. Physical aggression, yelling, cursing, drawing eyebrows together, pursing lips, narrowing eyes. Crying, frowning, eyes drooped, moaning, sighing, eyes/head turned down. Voice shaking, shrieking, repetitive calling out, line between eyebrows, tight facial muscles.
Neutral	No sign of discrete facial expression
Visual engagement	
Engaged	Appears alert and maintaining eye contact with robot. Eyes following robot or looking at robot.
With facilitator	Appears alert and maintaining eye contact with facilitator or others. Eyes following facilitator or others. Looking at facilitator or others.
Not engaged	Appears inattentive; blank stare into space. Does not make eye contact with robot.
Verbal engagement	
Positive	Participating and verbally responding to the robot's statements/questions. Willing to maintain a conversation.
Negative	Not participating and maintaining conversation. Uses repetitive generalized answers such as 'fine,' 'yes,' and 'no.' Not responding when prompted. Verbalizes the desire to leave. Refuses to participate. Lack of response.

Table A3.3: Items used in the robot acceptance questionnaire.

Theme	ID	Description	Items
Perceived sociability	SOC	The perceived ability of the robot to perform sociable behaviour	<i>The robot is a pleasant conversational partner.</i> <i>The robot could improve my social interaction skills.</i>
Perceived ease of use	EAS	The degree to which one believes that using the robot would be free of effort	<i>I could interact with the robot without any help.</i> <i>I could use the robot independently if I had a set of instructions.</i>
Perceived usefulness	USE	The degree to which one believes the robot would be assistive in the daily context at home	<i>I would find the robot useful in my life at home.</i> <i>The robot could help in activities of daily living (such as taking vitals).</i> <i>The robot could support wellbeing and independence at home.</i>
Anxiety	ANX	Evoking anxious or emotional reactions when using the robot	<i>If I were to use the robot at home, I would be afraid to make mistakes with it.</i> <i>I find the robot intimidating.</i>
Trust	TRU	The belief that the system performs with personal integrity and reliability	<i>I would trust the robot when it gives me advice.</i> <i>I would follow the advice the robot gives me.</i>

Table A3.4: OME results in robot-assisted cognitive engagement. Results from the OME, which considered the following parameters to assess engagement during robot-assisted clinical sessions: participant's attention on a scale of 1 (very disruptive) to 7 (very attentive); attitude to stimulus rated on a scale of 1 (very negative) to 7 (very positive); duration of engagement (time until not interested); frequency rate, 0 (none) to 3 (most or all the session), of talking to the robot, talking about the robot with the nursing assistant, being disruptive or distracted. The measures were observed by a trained nursing assistant.

Measure		Session 1	Session 2	Session 3
Attention	Average	7	6	5
	Highest	7	7	7
Attitude	Average	4	4	3
	Highest	6	4	4
Duration of engagement		9 min 35 s	10 min 47 s	18 min 1 s
Talking to the robot		3	3	3
Talking about the robot		0	0	0
Disruptive		0	0	0
Distracted		0	1	2

A4 Picture Description Cognitive Task

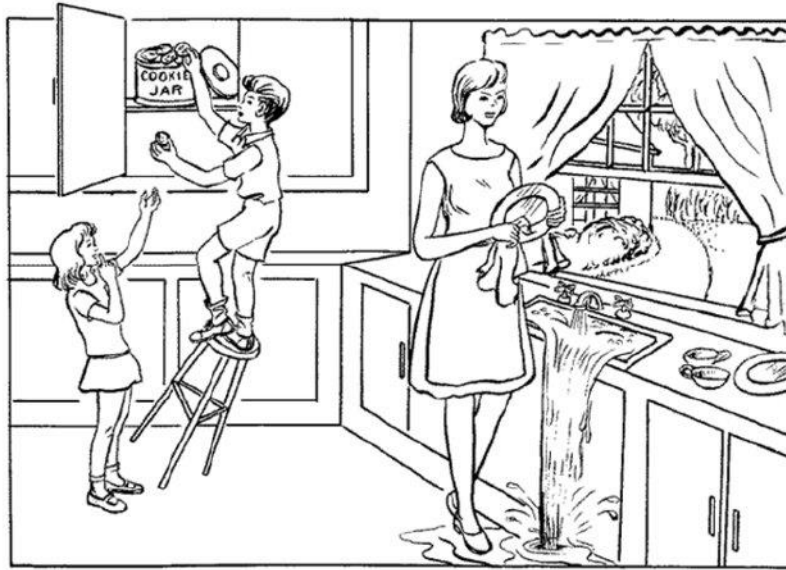


Figure A4.1: Cookie Theft picture. This picture is used in NPT and taken from the Boston Diagnostic Aphasia Examination.



Figure A4.2: Picnic Scene picture. This picture is used in NPT and taken from the Western Aphasia Battery-Revised.

A5 LLM-Powered Robot-Administered Cognitive tasks

Table A5.1: LLM-generated response parameters. Input parameters used in OpenAI's API calls.

Parameter	Value
model	GPT-4o
messages	<transcript>
stop	null
temperature	1
frequency_penalty	0
presence_penalty	0
max_tokens	256
top_p	1

Table A5.2: Groups used in LLM-generated hints during verbal cognitive tasks. Groups used in picture description and semantic fluency tasks with pre-defined areas and actions. The content units for the picture descriptions were selected based on literature about neuropsychological evaluations [106].

Groups	Examples of content units and words
Picture: <i>Cookie Theft</i>	
Boy	brother, stool, falling, reaching, cookies, jar
Girl	sister, laughing, reaching, help
Woman	mother, sink, washing, dishes, water, overflowing
Atmosphere	disaster, mess, chaos, family
Feelings	excited, determined, distracted, anxious, sneaky, curious
Picture: <i>Picnic Scene</i>	
Couple	picnic, reading, book, drink, basket, radio
Boy	kite, dog, water, edge
Child	beach, playing, sandcastle, spade
People	fishing, jetty, pier, sailing, water, boat
Objects	car, garage, flag, flying, tree, house
Semantic Category: <i>Animals</i>	
Pet	Dog, cat, hamster, guinea pig, fish, rabbit
Jungle	Lion, tiger, elephant, wolf, bear, gorilla
Can fly	Bat, eagle, sparrow, owl, falcon, pigeon
Ocean	Whale, shark, dolphin, octopus, sea turtle, squid
Farm	Cow, sheep, goat, horse, pig, chicken
North pole	Polar bear, penguin, seal, walrus, reindeer
Semantic Category: <i>Fruits</i>	
Citrus	Orange, lemon, lime, tangerine, clementine
Yellow	Banana, mango, lemon, pineapple
Tropical	Mango, pineapple, coconut, passionfruit, kiwi
Berries	Strawberry, raspberry, blackberry, blueberry, cranberry
With a pit	Peach, plum, cherry, apricot, mango
Used for juices	Strawberry, blueberry, grape, orange, mango, apricot

Table A5.3: Statistics results for linguistic analysis during the picture description task. Spearman's correlation results between traditional linguistic features and MMSE scores, as well as comparison results between Session 1 and Session 5 across the cohort using the paired Wilcoxon signed-rank test.

Feature	Description	MMSE $r(p, p_{corrected})$	Session 1 vs. Session 5 $W(p, p_{corrected})$	Cohen's d
Type-Token Ratio	The ratio of unique words to total words (tokens) in a transcript corrected for total length, measuring lexical diversity.	0.416 (0.054, 0.220)	33 (0.001*, 0.009*)	0.65
Propositional Idea Density	The number of distinct propositions (i.e. distinct facts or notions contained in a text) divided by total words, assessing vocabulary richness.	-0.227 (0.31, 0.465)	84 (0.273, 0.273)	0.29
Noun Rate	The proportion of nouns relative to total words in the transcript.	0.137 (0.543, 0.652)	68 (0.167, 0.2)	0.4
Pronoun Rate	The proportion of pronouns relative to relative to total words.	0.008 (0.973, 0.973)	67.5 (0.095, 0.142)	0.41
Verb Rate	The proportion of verbs relative to relative to total words.,	-0.245 (0.272, 0.465)	54.5 (0.034*, 0.102)	0.5
Adjective Rate	The proportion of adjectives relative to total words.	0.389 (0.073, 0.220)	44.5 (0.074, 0.142)	0.57

Table A5.4: Acceptance scores for robot-administered sessions. The results are presented as mean (SD).

Session	Sociability	Enjoyment	Usefulness	Anxiety	Trust
S2	4.4(0.7)	4.7(0.5)	4.2(0.8)	4.4(0.6)	3.5(1.2)
S3	4.5(0.6)	4.7(0.6)	4.2(1.0)	4.7(0.4)	3.5(1.2)
S4	4.6(0.6)	4.7(0.6)	4.2(1.0)	4.7(0.5)	3.7(1.2)

Table A5.5: Spearman's correlation between robot acceptance ratings and BFI personality traits. Mean acceptance scores are considered from three robot-administered sessions, presented as r (p , $p_{corrected}$).

	Sociability	Enjoyment	Usefulness	Anxiety	Trust
Extraversion	0.381 (0.080, 0.670)	0.119 (0.598, 0.830)	0.427 (0.047*, 0.590)	0.078 (0.732, 0.858)	0.490 (0.021*, 0.518)
Agreeableness	0.141 (0.532, 0.830)	0.158 (0.482, 0.830)	0.120 (0.594, 0.830)	0.132 (0.558, 0.830)	0.191 (0.394, 0.830)
Conscientiousness	-0.177 (0.431, 0.830)	0.170 (0.448, 0.830)	-0.076 (0.737, 0.858)	0.315 (0.153, 0.763)	-0.031 (0.890, 0.923)
Neuroticism	0.225 (0.313, 0.830)	0.022 (0.923, 0.923)	0.170 (0.449, 0.830)	-0.049 (0.829, 0.901)	0.271 (0.223, 0.797)
Openness	-0.072 (0.750, 0.858)	-0.215 (0.336, 0.830)	0.071 (0.755, 0.858)	0.323 (0.142, 0.763)	-0.279 (0.209, 0.797)

A6 Ethics Approval IRB

This section details the IRB approval by the University of Southern California (UP-24-00154), including the official approval (after amendment), the study protocol, and the updated IRB. All documents in this IRB submission were created by the author (PI: Professor Maja Matarić).

University of Southern California Institutional Review Board
3720 S. Flower Street, Suite 325
Los Angeles, CA 90089
Telephone: (323) 442-0114
Fax: (323) 224-8389
Email: hrpp@usc.edu

Date: Aug 16, 2024, 04:37pm
Action Taken: **Approve**
Principal [Maja Mataric](#)
Investigator: COMPUTER SCIENCE
Faculty
Advisor:
Co- [Jesse Thomason, PhD](#)
Investigator(s): COMPUTER SCIENCE
Project Title: [New questionnaire and design modification](#)
Amendment **UP-24-00154-AM001**
ID:
Funding
Types:

Your submission was reviewed and approved by the University of Southern California Institutional Review Board (IRB) under expedited review 45 CFR 46.110(b)(1)(ii) on 8/16/2024.

You are required to ensure that this research and the actions of all project personnel involved in conducting the submission will conform with the research project and its modifications approved by the IRB, IRB Policies and Procedures, and applicable state laws. University of Southern California is in compliance with requirements of human subjects research, including 45 CFR 46, 21 CFR 50 and 56.

Any unanticipated problems involving risks to participants or others, deviations from the approved research, non-compliance, and complaints must be reported to the IRB in accordance with University of Southern California Human Research Protection Program policies and procedures. If this study includes ongoing oversight by a Data Safety Monitoring Board (DSMB) or other such committee, reports generated by the DSMB or oversight committee must be submitted to the IRB.

The following materials were considered for review and approved:

1. iSTAR Application UP-24-00154-AM001
2. Secondary protocol amendment (0.01)
3. Study protocol amendment (0.01)
4. Mini Mental State Examination(0.01)
5. Spanish Consent Form(0.01)

Attachments:

Social-behavioral health-related interventions or health-outcome studies must register with **clinicaltrials.gov** or other International Community of Medical Journal Editors ([ICMJE](#)) approved registries in order to be published in an ICJME journal. The ICMJE will not accept studies for publication unless the studies are registered prior to enrollment, despite the fact that these studies are not applicable “clinical trials” as defined by the Food and Drug Administration (FDA). For support with registration, go to www.clinicaltrials.gov or contact Jean Chan (jeanbcha@usc.edu, 323-442-2825).

Approved ICs and HIPAA forms: [view](#)

This is an auto-generated email. Please do not respond directly to this message using the "reply" address. A response sent in this manner cannot be answered. If you have further questions, please contact iStar Support at (323) 276-2238 or istar@usc.edu.

The contents of this email are confidential and intended for the specified recipients only. If you have received this email in error, please notify istar@usc.edu and delete this message.

Study Title: Promoting Cognitive Health in Elder Care Using Socially Assistive Robots

Principal Investigator: Prof. Maja Mataric

Study Procedures

1. Background/Rationale

As the world population ages, the number of people living with dementia is increasing, with 10 million new cases each year [1]. An estimated 6.7 million Americans aged 65 and older are living with Alzheimer's in 2023 [2]. There is a strong need for accessible technologies that can promote cognitive health and detect early signs of cognitive deterioration leading to more comprehensive evaluation by medical professionals.

2. Purpose/Objectives/Aims/Research Questions

This study aims to investigate the feasibility of using socially assistive robots (SARs) to promote and assess cognitive health in older adults. We will conduct a within-subjects 5-week study with three consecutive weekly sessions with a SAR and two with a human. Each session will include two interactive cognitive tasks aimed at promoting and assessing cognitive health. In both the human and robot sessions, the administrator will act as a coach, providing hints during a series of interactive tasks based on validated cognitive tools. By using a human baseline in the first session, we aim to assess task performance and absolute learning after three weekly sessions with a SAR. We also aim to evaluate robot acceptance in a cognitive assessment and support context. Additionally, we aim to explore how older adults engage with the robot and how much they enjoy interacting with the robot.

This within-subjects 5-week study will pursue the following research questions to explore how to engage target populations with SARs to promote and assess cognitive health:

Q1 How well do older adults accept a SAR that administers cognitive tasks?

Q2 How does the older adult's task performance change in three weekly sessions with a SAR?

Q3 How much does the older adult's task performance change in the same cognitive tasks administered by a human after three weekly sessions with a SAR?

Q4 What are the attitudes and experiences of older adults who interact with a SAR in this context?

3. Participants (sample)

- a) We will recruit 20 participants aged 65 and older. Participants will be recruited from Front Porch, a partnering collaborator of the USC Interaction Lab. See support letter in Section 6d.1.10.
- b) Describe the inclusion and exclusion criteria for enrollment.
 - i) Inclusion criteria: older adults (aged 65 and older) who are fluent in English or Spanish and with normal or corrected-to-normal hearing. We will recruit participants who are fluent in English or Spanish because the robot speaks these languages and the study personnel are English and Spanish speakers.
 - ii) Exclusion criteria: Inability to consent to the study information sheet. Vision or hearing impairments that preclude hearing and seeing the robot.
- c) We will aim for gender balance in the participant cohort. Ideally, half of the group will self-identify as female, and the other half as male. Other genders will also be welcome though unlikely in the population being recruited. The study gender distribution is constrained by the resident population at Front Porch and their willingness to participate in the study.
- d) Regarding race and ethnicity, we aim for diversity, and ideally, no more than half of the cohort will be White/Caucasian. This is also constrained by the distribution of residents at Front Porch and their willingness to participate in the study.
- e) No children will be included because the research is aimed at older adults and age-related cognitive screening and support.

4. Recruitment/Screening Process (sampling strategy)

- a) Our partnering collaborators at Front Porch will recruit participants by posting flyers in their residential locations. Front Porch will engage with participants interested in the study, organizing and coordinating times for the human-robot interaction sessions.
- b) Interested participants who meet our criteria from Section 3 will complete a consent form.
 - i) Front Porch will share a printed version of the consent form with participants.
 - ii) Front Porch will request a phone number to contact participants via phone call to remind them of their sessions.
 - iii) A day before each appointment, a reminder phone call will be made by study personnel to the participant including the day, location, and time of the session.
- c) Recruitment methods: our collaborators at Front Porch will recruit participants by posting flyers in their residential locations.
- d) Recruitment materials:
 - i) A recruitment flyer will be distributed in Front Porch's residential locations - See Section 24.2: *Flyer*.
 - ii) Before the participants start the study, an information sheet and consent form will be provided and filled out by each participant.

5. Methods

a) Project Description

Recruitment and Consent:

- i) Our recruitment process will be coordinated by our partners at Front Porch. Flyers will be posted in multiple Front Porch residential locations.
- ii) The project description and consent form will be shared in print with participants prior to the first day of the study.
- iii) Front Porch will phone call participants a day before each appointment to remind them of the session, including the day, location, and time of the session.
- iv) Participants will be informed that audio and video will be recorded.
- v) Participants will be informed that the data collected during the study could be shared in anonymized form for research purposes. All the responses from the participants will be anonymous.

Study Procedure:

The study will consist of five sessions conducted over five weeks, as shown in Figure 1. Participants will engage in three consecutive weekly sessions with a SAR, starting with an initial human-administered session to establish a baseline for each participant. The final session will also be human-administered to assess performance after the three weekly robot sessions and evaluate absolute learning. During each session (human and robot-administered), the task administrator will provide instructions and offer encouragement and hints when necessary. We will use two versions of each task, alternating between them in the sequence V1, V2, V1, V2, V1 such that the first and last human-administered sessions have the same task version. These versions are based on validated cognitive tools and include slight variations in content (for instance, in the picture description task, a different picture will be used) to prevent participants from learning the tasks by memory over time.

Week	W1	W2	W3	W4	W5
Task version	V1	V2	V1	V2	V1
Administrator	HUMAN	ROBOT	ROBOT	ROBOT	HUMAN

Task performance & engagement in 3 weekly robot sessions

Baseline: learning after 3 weekly robot sessions

Fig. 1. Within-subjects 5-week study design. Two task versions will be used (V1, V2) in three consecutive sessions with a SAR in addition to two human-administered sessions in the first and last weeks.

Each session is expected to last 30-45 minutes. Participants will engage in interactive cognitive tasks based on validated cognitive assessment tools. On the first day of the study, participants will bring the consent form, which they will have received previously and had time to read and voluntarily agreed to participate by signing the form.

Each session will start with the researcher explaining the overall motivation and objectives of the study, and answering questions or concerns from the participant (five minutes). The researcher will follow a script to maintain consistency across sessions and participants.

In robot-administered sessions, the researcher will introduce the robot and briefly explain how to interact with the robot before proceeding with cognitive tasks, each of which will take up to seven minutes to complete and be followed by a short pause of less than five minutes. Participants will engaged in these two cognitive tasks:

- Task 1: Picture description, adapted from the Boston Diagnostic Aphasia Examination [3]. The administrator (human or robot, as shown in Fig. 1) will direct participants to a printed picture and prompt them to describe everything they see in the picture. The administrator will encourage participants to maintain engagement, offering hints when there are long pauses, guiding their attention to specific regions in the picture.
- Task 2: Semantic task, adapted from the Addenbrooke's Cognitive Examination [4]. The researcher will ask participants to name as many words as possible within a given category, such as animals or fruits. The administrator (human or robot, as shown in Fig. 1) will encourage participants to keep engaging, providing hints about additional words not yet mentioned by the participant.

In all sessions, we will collect data through audio and video recordings, observations, and questionnaires. In the first session only, participants will complete demographic and personality questionnaires (see details in Instrumentation below). In addition, participants will complete the following questionnaires after the interactive cognitive tasks in robot sessions:

- The robot acceptance and experience questionnaire based on the Almere model of technology acceptance, tailored for SARs in geriatric care [7].
- The system usability scale questionnaire [8].

In the last human-administered session, participants will complete the Mini-Mental State Examination (MMSE), a standard tool that assesses five areas of cognitive function: orientation, registration, attention and calculation, recall, and language. It takes only 5-10 minutes to administer. The maximum score is 30, which will be calculated following previous research and available guidelines [9].

Data management and privacy:

Video and audio recordings will be made using a camera outside of the robot. The data will be uploaded to HIPPA-compliant locked cloud storage tools that are accessible only to authorized researchers. Consent forms and other confidential data will be placed in a locked location and not disclosed to third parties. Identifiable study data will only be available to researchers on this IRB.

Dissemination:

All publishable results of this work will be shared in conference and/or journal publications, using only de-identified aggregate statistics. One possible outcome of this data collection is the creation of a dataset that could be used for training machine learning models targeting dementia screening from cognitive tests. Therefore, we may make available by request anonymized features derived from recorded patient speech and video; such data sharing is standard because those features cannot reconstruct the participant identities, they are purely numerical and do not include original audio or video, and are numerically processed through machine learning-based feature extraction methods.

b) Instrumentation

- i) We will use Blossom, a safe, affordable, and versatile stationary, table-top, soft, affective robot platform with interactive capabilities that is 3D-printed, low cost, and open source (see Fig. 2). Blossom has been used in multiple other USC IRB-approved studies in our lab.
- ii) Participants will interact verbally with the robot to complete a series of tasks adapted from cognitive tests supported by the literature. For each task, the robot will act as a coach that provides task instructions and offers hints and encouragement.
- iii) Questionnaires/Surveys
 - 1) Demographic Questionnaire - See Section 21.2: *Demographic Questionnaire*.
 - 2) Big Five Inventory-10 [10] - See Section 21.2: *Big Five Inventory Shortened*
 - 3) System Usability Questionnaire - See Section 21.2: *Robot Acceptance Questionnaire*
 - 4) Robot Acceptance and Experience Questionnaire - See Section 21.2: *System Usability Questionnaire*
 - 5) MMSE Questionnaire - See Section 21.2: *Mini Mental State Examination*
- iv) Qualitative Instruments
 - 1) We will also assess user experience, enjoyment, comfort, and subjective perceptions of the human- and robot-administered interactions in a semi-structured interview after the tasks in each session - See Section 19.1: *Semi Structured Interview*.



Fig. 2. Blossom robot platform used in the study.

c) Data Analysis

- i) Data analysis techniques may include standard statistical tools including parametric and non-parametric tests.
- ii) We will assess human-robot engagement through task completion rate, time-on-task (before disengagement), video and audio data obtained from recordings to get objective measures of attention and behavior.
- iii) We will compare task performance over the sessions: for the picture description task, we consider the number of correct information content units and overall level of detail, for the semantic task we consider the number of correct words for the category, for the story recall task we consider the level of detail recalled.

References

- [1] Dementia key facts (2021), [online] Available: <https://www.who.int/news-room/fact-sheets/detail/dementia>.
- [2] Alzheimer's Disease Facts and Figures (2023), Alzheimer's Association, [online]. Available: <https://www.alz.org/alzheimers-dementia/facts-figures>
- [3] Giles, E., Patterson, K. and Hodges, J.R., 1996. Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer's type: missing information. *Aphasiology*, 10(4), pp.395-408.
- [4] Bruno, D. and Schurmann Vignaga, S., 2019. Addenbrooke's cognitive examination III in the diagnosis of dementia: a critical review. *Neuropsychiatric disease and treatment*, pp.441-447.

- [5] Wechsler, D., 1945. A standardized memory scale for clinical use. *The Journal of Psychology*, 19(1), pp.87-95.
- [6] Gavett, B.E., Gurnani, A.S., Saurman, J.L., Chapman, K.R., Steinberg, E.G., Martin, B., Chaisson, C.E., Mez, J., Tripodis, Y. and Stern, R.A., 2016. Practice effects on story memory and list learning tests in the neuropsychological assessment of older adults. *PloS one*, 11(10), p.e0164492.
- [7] Heerink, M., Kröse, B., Evers, V. and Wielinga, B., 2010. Assessing acceptance of assistive social agent technology by older adults: the almere model.
- [8] Bangor, A., Kortum, P.T. and Miller, J.T., 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human-Computer Interaction*, 24(6), pp.574-594.
- [9] Folstein, M.F., Folstein, S.E. and McHugh, P.R., 1975. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3), pp.189-198.
- [10] Rammstedt, B. and John, O.P., 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality*, 41(1), pp.203-212.

Amendment: New questionnaire and design modification (UP-24-00154-AM001)

Amendment Information

AM1.1. * Identifier for this Amendment:
New questionnaire and design modification

AM1.2. Type of Amendment (check all that apply):
Type

☐ Editorial

☐ Drug or Device

☐ Enrollment Status Change

☐ Funding

☐ Informed Consent and/or Addenda

☐ Investigator's Brochure / Package Insert

☐ Locations

☐ Number of Subjects

☒ Procedures and/or Protocol

☐ Recruitment Materials

☐ Risks/Harms

☐ Study Personnel

☒ Subject Population

☐ Subject Reimbursement / Compensation

☐ Other

If this is a change initiated to eliminate an **Immediate Hazard**, please DO NOT WAIT for IRB approval before making changes to protect the health and welfare of study participants. In such a circumstance, please also (1) continue with the amendment process and (2) submit a reportable event of the type "Protocol Change Initiated to Eliminate Immediate Hazard".

AM1.3. Will the changes you are making affect the accuracy of the study abstract?
☐ Yes ☒ No

AM1.6. Has the risk/benefit ratio changed?
☐ Yes ☒ No

AM1.7. Will you be submitting a Simple amendment request?
☒ Yes ☐ No



If you intend to submit this application as a Simple Amendment, please make sure to read the [Simple Amendment Guidelines](#) before you indicate ☒ Yes ☐ No above. This type of amendment will get review prioritization provided it meets the qualifications stipulated by the IRB. The IRB can remove this prioritization at any time if it does not meet these requirements and it will be placed into the normal IRB queue in the order it was received. If you have questions regarding this application, please contact the analyst assigned to the study.

Amendment: New questionnaire and design modification (UP-24-00154-AM001)

AM8. Procedures and/or Protocol

This screen is required if you indicated this amendment involves changes to Procedures and/or the Protocol (Question AM1.2.)

AM8. Please describe the nature and rationale for changes to the protocol, procedures or methods:

We are adding the Mini Mental State Examination (MMSE) to measure participants' cognitive function across five standardized domains: orientation, registration, attention and calculation, recall, and language. The MMSE is a brief questionnaire that takes only 5-10 minutes to administer. This change will not affect the agreed-upon duration of the experiment because the questionnaire will be administered in the last session of the study. We will also inform participants that the MMSE is a standard assessment of cognition and will be available to answer any questions they may have. Including the MMSE will enhance our post-experiment analysis by providing valuable insights into cognitive abilities and potential mild cognitive impairment within our cohort.

We also updated our study design to include one phase instead of two and use the first human-administered session to establish a baseline for each participant. This minor change does not impact the number of sessions for participants nor the data collected. We still have our proposed four research

questions which now happen within a single phase.

You must make your changes in the body of the main study application. In addition to the changes in the Methods and Procedures sections (items 9, and 12 through 21 inclusive), please remember other likely places of change include the informed consent (item 24) and protocol (5.2). If you have a revised protocol, it must be uploaded in the main study application, including the sponsor's Summary of Changes (if available).

Amendment: New questionnaire and design modification (UP-24-00154-AM001)

AM13. Subject Population

This screen is required if you indicated this amendment involves Subject Population changes (Question AM1.2.)

AM13. Please describe each of the changes to the Subject Population and provide a rationale for the change:

We are expanding our inclusion criteria to include fluency in English or Spanish because our robot can now communicate effectively in both languages. This will enhance participant recruitment while maintaining our maximum sample size of 30. This change will also enable us to better represent the elderly population in California, thereby increasing the relevance of our findings. A native Spanish speaker from our team reviewed the Spanish document versions used for recruitment and informed consent and conducted the sessions with participants.

You must make your changes in the body of the main study application. Inclusion and exclusion criteria must be updated at item 10, and special subject populations are in item 22. If these revisions require changing any other information, including the Protocol or Informed Consent, you must revise that information as well.

Amendment: New questionnaire and design modification (UP-24-00154-AM001)

AM99. Ready to make Changes

Following this page is a copy of the study application in which all changes specified must be made. When you have completed your changes do not forget to submit the amendment to the IRB.



1. Project Identification and Abstract

1.1. * Type of Submission:

- ☒ Research Protocol or Study on Human Subjects
- ☐ Use of Humanitarian Use Device (Not Research)
- ☐ Grant/Contract Only
- ☐ Rely on another IRB (Ceded)
- ☐ Right To Try

1.2. * Full Title of Research Protocol

Promoting Cognitive Health in Elder Care Using Socially Assistive Robots

1.3. * Short Title

Promoting Cognitive Health in Elder Care Using Socially Assistive Robots

1.4. * Abstract: Provide a simple explanation of the study and briefly address (in 1 to 2 sentences) each of the following points: rationale; intervention; objectives or purpose; study population or sample characteristics; study methodology; description of study arms (if appropriate); study endpoints or outcomes; follow-up; statistics and plans for analysis.

As the world population ages, the number of people living with dementia is increasing, with 10 million new cases each year. An estimated 6.7 million Americans aged 65 and older are living with Alzheimer's in 2023. There is a strong need for accessible technologies that can promote cognitive health and detect early signs of cognitive deterioration leading to more comprehensive evaluation by medical professionals. This study aims to explore novel ways to promote and assess cognitive health maintenance and monitoring in seniors by using robots in an interactive way. Study participants (N=20, aged 65 and older) will engage with a socially assistive robot during a series of interactive cognitive tasks based on validated dementia screening tools. We will conduct a within-subjects 5-week study with three consecutive weekly sessions with a SAR and two with a human. Each session will include two interactive cognitive tasks aimed at promoting and assessing cognitive health. In both the human and robot sessions, the administrator will act as a coach, providing hints during a series of interactive tasks based on validated cognitive tools. We aim to evaluate robot acceptance in a cognitive support context and compare task performance in the two conditions. The second phase involves three weekly robot-administered sessions to assess performance and learning on similar cognitive tasks. Additionally, we aim to explore how older adults engage with the robot and how much they enjoy interacting with the robot. Data will be collected through video recordings, observations, and questionnaires, which will be analyzed to produce novel results for peer-reviewed publications. Additionally, this data collection from seniors would allow us to generate a dataset that could be used for training machine learning models targeting dementia screening from cognitive tests.

1.5. Is this a clinical trial? [The NIH defines a clinical trial as a prospective research study to evaluate the effects of one or more interventions on health-related biomedical or behavioral outcomes.]

- ☐ Yes ☒ No

1.6. * Select which IRB you are requesting review from:



☐ USC - Biomedical IRB

☒ **USC - Social Behavioral IRB**

☐ CHLA - Institutional Review Board (CHLA)

Select **"USC - Biomedical"** if you are conducting biomedical research activities which fall under FDA jurisdiction (drugs, devices, etc.).

Select **"USC - Social Behavioral"** if you are conducting social behavioral research activities (community/field-based research such as survey and interview procedures, etc. that are not FDA regulated.).

NOTE: IRBAs and designated reviewers (Chairs, Vice Chairs, etc.) are currently assigned to review applications based on the type of research conducted. Selecting the incorrect board may result in a delay of the review process; and the application may be returned for the correct review board to be selected.

1.7. *** To the investigator's knowledge, does the Institution have financial and/or intellectual property interests in the sponsor or the products used in this project? An institutional conflict may occur when a financial interest of the institution has the potential to bias the outcome of research conducted by its employees or students or to create an unacceptable risk to human subjects.**

☐ Yes ☒ **No**

2. Study Personnel

2.1. Study Personnel and their roles:

	Last Name	First Name	Organization	Study Role	Certifications	Obtain Consent	Interact with Participants	Access Identifiable Data	Manage Audit Access to PHI/ePHI (CHLA Only)
View	Mataric	Maja	COMPUTER SCIENCE	Principal Investigator	☒ HS ☒ GCP ☒ HIPAA	no	no	yes	
View	Thomason	Jesse	COMPUTER SCIENCE	Co-Investigator	☒ HS ☒ HIPAA	no	no	yes	
View	Hulyalkar	Avni	VITERBI SCHOOL OF ENGINEERING	Research Assistant or Associate	☒ HS ☒ HIPAA	yes	yes	yes	
View	Lima	Maria	University of Southern California (USC)	Research Assistant or Associate	☒ HS ☒ HIPAA	yes	yes	yes	
View	Nagahara	Alethea	VITERBI SCHOOL OF ENGINEERING	Research Assistant or Associate	☒ HS ☒ HIPAA	no	yes	yes	
View	O'Connell	Amy	VITERBI SCHOOL OF ENGINEERING	Research Assistant or Associate	☒ HS ☒ HIPAA	yes	yes	yes	
View	Pinto	Alva	VITERBI SCHOOL OF ENGINEERING	Research Assistant or Associate	☒ HS ☒ HIPAA	yes	yes	yes	
View	Zhou	Feiyang	any	Research Assistant or Associate	☒ HS ☒ HIPAA	no	yes	yes	

Who may be included as "key personnel" on an IRB submission?

Key Personnel are individuals who contribute to the scientific development or execution of a project in a substantive, measurable way, whether or not they receive salaries or compensation under the protocol. Individuals who should be named on an IRB application are those who engage in the following:

- a. conducting research through an interaction or intervention with human subjects for research purposes
- b. participating in the consent process by leading it or contributing to it
- c. directly recording or processing identifiable private information, including protected health information, related to those subjects for the purpose of conducting the research study

Who should NOT be listed as key personnel on an IRB submission:

Individuals paid by the institution to perform a service not part of or paid by the research project performing services that are typically performed for non-research purposes or fee for service:

- honest broker
- pharmacy employees dispensing investigations drugs
- hospital employees obtaining blood through a blood draw or collect urine and provide such specimens to investigators as a service
- radiology clinic employees performing chest x-rays and sending results to investigators as a service
- routine laboratory analyses of blood samples for investigators as a commercial service
- transcription of research study interviews as a commercial service
- not administering any study intervention being tested or evaluated under the protocol

2.2. **Is the Principal Investigator a staff member, student, resident, fellow, postdoctoral scholar, other trainee, visiting scholar, or visiting faculty member at USC/CHLA?**

☐ Yes ☒ **No**



2.4. Does your study include specific aims or endpoints that relate to cancer risk factors, cancer screening or diagnosis, cancer prevention, cancer treatment, or supportive care?

Note: For studies of diet, tobacco, sun exposure, or other cancer risk factors, or any study that includes cancer patients, please answer yes.

☐ Yes ☐ No

2.5 Specify the group/organization who has reviewed this study for scientific merit:

☒ Federal Agency (e.g. FDA, NIH, CDC, DOE, NSF, DOJ, etc.)

☐ USC Norris Clinical Investigations Committee

☐ Doctoral Dissertation Committee

☐ Other

☐ None

4. Funding Information

4.1. * What existing or pending support will be used for this study? (check all that apply)

☐ Cooperative Group (SWOG, COG, RTOG, etc.)

☐ CTSI

☐ Departmental/Institutional Funds

☒ Federal Grant/Contract

☐ Non-Profit (including Foundations and Universities)

☐ Industry

☐ Intramural/Internal Grant

☐ State or Local Grant/Contract

☐ No Funding

☐ Other

Change of existing/addition of new funding sources and agencies will require an amendment.

4.1.1. Will you be submitting an Urgent Review request?

☐ Yes ☒ No

4.1.2. Will any funds come from the National Institutes of Health (NIH)?

☐ Yes ☒ No

4.1.3. Will any of the funds come (either directly or indirectly) from the Department of Defense (DoD)?

☐ Yes ☐ No

4.2. If the funding source has undergone separate review by the IRB (i.e., cooperative group grants, umbrella grants, multi-project/program grants, center grants), try to select it from the list using the "Add" button. If the funding source is not displayed in the list, enter the information in question 4.4.

Grant #	Principal Investigator	Grant Title
There are no items to display		

4.2.1. If the grants selected in question 4.2 fund multiple studies, please attach the specific pages of the grant that are relevant to THIS study.

Name Version Modified

There are no items to display

4.4. Add the details of each source of funding for this study.

Sponsor	Principal Investigator	Type of Funding
View National Science Foundation	Maja Mataric	Federal: Grant *
View NSF (National Science Foundation)	Maja Mataric	Federal: Grant *

4.5. If there are related awards or proposals, please use the respective button below to link it to this study.

Related Proposals from Cayuse (uncheck to remove):

Proposal Number	Project Title	Principal Investigator	Sponsor	Proposal Start Date	Proposal End Date	Status Name
There are no items to display						

Related Awards from Cayuse (uncheck to remove):

Award Number	Project Title	Principal Investigator	Sponsor	Award Start Date	Award End Date	Status Name
23-0846-A0001	011576-00001	Maja Mataric	National Science Foundation	9/1/2019	9/30/2024	Active

5. Type of Study Review

5.1. Select the type of review that you are requesting for this study:

- ☐ Full Committee Review
- ☒ Expedited Review
- ☐ Exempt Review

5.2. All studies must have a fully developed protocol. For investigator-initiated studies, use the appropriate protocol template. Social Behavioral, Biomedical, or Secondary Analysis protocol templates are located at the [HRPP website](#). Cancer Center Protocol templates can be found on the [CISO website](#). If you are uncertain which protocol template to use, consult with the IRB before developing your protocol. Using an incorrect protocol will result in your application being returned as incomplete.

****Please note that a grant application cannot serve as the protocol.****

Attach the protocol below.

Name	Version	Modified
 Secondary protocol(0.02)	0.02	8/9/2024 9:12 AM
 Study protocol(0.02)	0.02	8/9/2024 9:11 AM

5.3. Attach the sponsor's template informed consent here.

Name	Version	Modified
There are no items to display		

5.4. If any study documents are password protected, enter the passwords here.

5.5. If there is a sponsor protocol number associated with this file, specify it here:

9. Methods and Procedures - Selected Descriptors/Community Engaged Research

Note: The list of items below IS NOT an all-inclusive list of methods and procedures available to investigators. The list only includes items that will trigger additional questions specific to areas of research or are necessary for the review process.

9.1. **NOTE:** Review the information under the blue question mark before answering this question.

- ☒ The study procedures involve interactions or interventions with participants (e.g., medical procedures, online surveys, in person interviews).
- ☐ Secondary Analysis: This study involves use of data/specimens **that have been or will be collected/generated for non-research purposes** (e.g., medical records, student records).
- ☐ Secondary Analysis: This study involves use of data/specimens **that have already been collected/generated for another research purpose** (e.g., existing research datasets, existing specimens collected for research).

9.2. Study Procedures: (check all that apply)

- ☒ Audio/Video Recordings or Photographs
- ☒ Behavioral Observations and/or Behavioral Experimentation
- ☐ Behavioral Interventions
- ☐ Deception
- ☒ Interview/Focus Groups
- ☒ Population-based Field Study
- ☐ Psychophysiological Testing



- ☒ Surveys/Questionnaires/Psychometric Testing
- ☐ Approved/Investigational Devices
- ☐ Biohazardous Substances (e.g. fresh tissue or tissue fluids, infectious agents, microorganisms, recombinant DNA, or shipment of biological material)
- ☐ Blood Collection
- ☐ Creation of a Data or Tissue Repository
- ☐ Magnetic Resonance Imaging (MRI)
- ☐ Stem Cells or Cell-Based Therapy

- 9.5. Will data from this study be subject to the NIH Genomic Data Sharing (GDS) policy?
☐ Yes ☒ No
- 9.6. Does your study involve community-engaged research (community-engaged research addresses community needs and involves the community in research plan, conduct of study, etc)?
☐ Yes ☒ No
- 9.7. Will data from this research be submitted to the US Food and Drug Administration or will data be held for inspection by the FDA?
☐ Yes ☒ No

10. Characteristics of the Study Subject Population

- 10.1. What is the maximum number of subjects you plan to recruit for this site? (Integer values only)
30
- 10.1.1. If this is a multi-site study, indicate the projected total subject accrual. (Integer values only)
30
- 10.1.2. If necessary, provide further explanation of accrual goals for all subject populations.
- 10.3.1. If there are any age, ethnic, language, or gender-based exclusion criteria, please provide justification.
We will recruit participants aged 65 and older.

Inclusion criteria: older adults (aged 65 and older) who are fluent in English or Spanish and with normal or corrected-to-normal hearing. We will recruit participants who are fluent in English or Spanish because the robot speaks these languages only and the study personnel are English and Spanish speakers.

Exclusion criteria: Inability to consent to the study information sheet. Vision or hearing impairments that preclude hearing and seeing the robot.

We will aim for gender balance in the participant cohort. Ideally, half of the group will self-identify as female, and the other half as male. Other genders will also be welcome though unlikely in the population being recruited. The study gender distribution is constrained by the resident population at Front Porch and their willingness to participate in the study.

Regarding race and ethnicity, we aim for diversity, and ideally, no more than half of the cohort will be White/Caucasian. This is also constrained by the distribution of residents at Front Porch and their willingness to participate in the study.

19. Methods and Procedures - Interview/Focus Groups

- This screen is required if you indicated the use of Interview or Focus Groups as a procedure (Question 9.2.)
- 19.1. Attach copies of any scripts and/or questions that will be used to guide the interviews/groups.

Name	Version	Modified
 Semi Structured Interview.pdf(0.01)	0.01	2/7/2024 11:59 AM

21. Methods and Procedures - Surveys/Questionnaires/Psychometric Testing

- This screen is required if you indicated the use of Surveys, Questionnaires, or Psychometric Testing (Question 9.2.)
- 21.2. Attach copies of all measures/instruments that will be used for this study.

Name	Version	Modified
 Big Five Inventory Shortened.pdf(0.01)	0.01	2/8/2024 1:48 AM
 Demographic Questionnaire.pdf(0.01)	0.01	2/8/2024 1:48 AM



Name	Version Modified	
 Mini Mental State Examination(0.01)	0.01	7/4/2024 12:32 PM
 Robot Acceptance Questionnaire.pdf(0.01)	0.01	2/8/2024 1:48 AM
 System Usability Questionnaire.pdf(0.01)	0.01	2/8/2024 1:48 AM

22. Special Subject Populations

22.1. Indicate any special subject populations you intend or expect to enroll in the research: (check all that apply)

- ☒ Healthy Volunteers Undergoing Clinical/Medical Procedures for Research Purposes
- ☐ Employees or Students
- ☐ Adults not Competent to Consent (or likely to lose the capacity to consent during the study)
- ☐ Non-English Speaking Populations
- ☐ Minors (subjects under 18 years of age)
- ☐ Pregnant Women / Human Fetuses
- ☐ Neonates (infants under 30 days old)
- ☐ Prisoners/Detainees
- ☐ Wards
- ☐ Military Personnel
- ☐ American Indian and/or Alaskan Native (AI/AN), Native Hawaiian and other indigenous communities
- ☐ None of the above

23. Study Resources

23.1. Describe the time the investigators have available to conduct and complete the research and justify that it is sufficient. Please check-off the items that apply to this study.

- ☐ Employed interns, residents, fellows, or postdocs with dedicated time to conduct this research.
- ☒ Employed faculty and or staff with dedicated time to conduct this research.
- ☒ Students with dedicated time as part of their training to conduct this research.
- ☐ Volunteers
- ☐ Other

23.2. Describe the staff and justify their qualifications. Please check-off the items that apply to this study.

- ☐ All biomedical investigators are privileged and credentialed to perform the study activities in the study locations.
- ☒ All study staff are trained and credentialed to perform the duties assigned to them.
- ☒ All study staff have fulfilled the training mandated by their respective departments or institutions.
- ☐ Other

24. Subject Recruitment and Informed Consent

24.1. Recruitment Tools that will be used by the local site (check a box only if your site will control the use or distribution of the recruitment tool): (check ALL that apply)

- ☐ Brochures
- ☐ Clinical Data Warehouse
- ☐ Email/Electronic Mailing Lists



☒ Flyers

☐ Letters

☐ Newspaper/Magazine Advertisements

☐ Radio/Television Announcements

☐ Subject or Participant Pools

☐ Telephone Scripts

☒ Verbal (Personal Solicitation)

☐ Websites / Social Media Outlets

☐ Other

☐ None of the above

24.1.2. Describe how you will be obtaining contact information:
Our collaborators at Front Porch will select appropriate study candidates based on their age (aged 65 and older), hearing and language ability (fluency in English or Spanish). Front Porch will engage with participants interested in the study, organizing and coordinating times for the human-robot interaction sessions.

24.2. Attach copies of all recruitment tools that will be used by the local site.(Do not attach any advertising or recruitment materials that will not be used at the local site or under control of the local site.)

Name	Version Modified	
 Flyer Spanish version(0.01)	0.01	8/9/2024 9:06 AM
 Flyer.pdf(0.01)	0.01	2/7/2024 5:58 AM

24.3. Informed Consent and Waivers:

**** Please note that child assent and parental permission will be addressed on subsequent pages. Do not complete the following consent questions if adults will not be participating in the study. ****

Check the type(s) of consent or waiver of consent planned for this study: (check ALL that apply)

☒ Written/signed consent (participants will sign an informed consent document)

☐ Alteration of the elements of consent (participants will sign a consent document, but one or more of the basic required elements of consent will be altered or waived)

☐ Waiver of signed informed consent (informed consent document without the signature sections will be provided)

☐ Waiver of consent (participants will not be asked to sign or be given a consent document)

24.7. Attach copies of the informed consent documents(s), information sheets, and any statements of new information/findings or consent addenda (as applicable) that will be used in this study. This set should also contain any assent or parental permission documents that will be used.

Name	Version Modified	
 Consent form Spanish version(0.01)	0.01	8/9/2024 8:53 AM
 Consent form.pdf(0.01)	0.01	2/7/2024 6:00 AM

[Click here to obtain an IRB Informed Consent Template and instructions for preparing the consent form.](#)
Consent forms submitted to the IRB should comply with these instructions.

**A qualified interpreter is an individual who is fluent (can speak, read and write) in English and the language of the subject, and (preferably) understands human research informed consent requirements. The interpreter should not be a member of the potential subject's family. Although it may be necessary in some rare cases to have a bilingual family member or staff person serve as a medical interpreter, keep in mind the following issues.*

1. The routine use of ad hoc interpreters should be avoided.
2. Children should not be asked to serve as an interpreter.
3. Complex ideas and treatment regimens may demand that a trained professional be employed.
4. Issues of privacy must be considered if family members are asked to translate.

Personnel from section 2.1 obtaining consent/permission/assent:				
Last Name	First Name	Organization	Study Role	Certifications Obtain Consent
Hulyalkar	Avni	VITERBI SCHOOL OF ENGINEERING	Research Assistant or Associate	yes
Lima	Maria	University of Southern California (USC)	Research Assistant or Associate	yes
O'Connell	Amy	VITERBI SCHOOL OF ENGINEERING	Research Assistant or Associate	yes



Last Name	First Name	Organization	Study Role	Certifications Obtain	Consent
Pinto	Alva	Leticia	VITERBI SCHOOL OF ENGINEERING	Research Assistant or Associate	yes

If the above list is incomplete or incorrect, please navigate to item 2.1 and make your changes there.

24.8. Describe the circumstances and location of the process of informed consent: (check ALL that apply)

- ☒ In a private area
- ☐ In a waiting room, open ward, group, or public setting
- ☐ Remotely (online/electronic, telephone, mail, or fax)
- ☐ Other

24.9. Describe how you will assess the individual's comprehension of the research and what it means to participate, including understanding of the voluntary nature of participating. (check ALL that apply)

- ☐ An assessment tool will be used. (attach a copy of the tool below)
- ☒ This will be verbally assessed. Individuals will be asked to answer the following questions (as applicable): (1) What are you being asked to do? (2) What question is this study trying to answer? (3) What are the potential risks of participating in this study? (4) How often will you need to come in for study visits? (5) What is the difference between participating in this study and your standard medical care? (6) What should you do if you decide to withdraw from the study?
- ☐ Other (specify below)

24.10. Describe all measures that will be taken during the recruitment and consent process to ensure that individuals have adequate time to consider participation and to safeguard against potential coercion and undue influence: (check ALL that apply)

- ☒ The informed consent process will begin with a concise and focused presentation of the key information that is most likely to assist a prospective participant in understanding the reasons why one might or might not want to participate in the research
- ☒ They will not be forced, threatened or coerced in any way to participate in this research, and no undue influence or other form of constraint will be used to recruit individuals to participate in this research or to retain currently enrolled subjects. (Note: "Coercion" is the use or threat of the use of force to gain compliance. "Undue influence" is when an individual who is in a position of authority (e.g., physician, teacher, employer) exerts inappropriate or excessive manipulation to gain power and compliance over a vulnerable individual (potential research subject). "Constraint" means force, obligation, or pressure.)
- ☒ They will not be punished or denied something which they would normally receive (e.g., threatening to withdraw health services to which an individual would otherwise be entitled) if they choose not to participate in this research or choose to withdraw early from participation.
- ☒ They will be given an adequate amount of time to consider participation in the study relative to the initiation of study procedures.
- ☒ The information presented to individuals during recruitment and consent will reflect that provided in the informed consent document/informed consent script.
- ☒ The recruitment and consent process will not promise them a certainty of cure or benefit beyond what is outlined in the informed consent document/informed consent script.
- ☒ The recruitment and consent process will take place in an area in which it is possible to maintain privacy and confidentiality.
- ☒ They will be given the opportunity to take the informed consent document home in order to discuss participation with their family, friends and/or others before making a definitive decision.
- ☒ They will receive payment for their participation, but the amount of payment will be commensurate with their participation (not an inducement for participation), and receipt of the payment will not be contingent upon the individual's completion of the study. (Note: The specific method, schedule, and amount of payment must be outlined in the payment section of the application.)
- ☐ Other (explain below)

25. Financial Obligation and Compensation

25.1. Financial Obligation: Choose the response that best describes the cost to participants.

- ☐ All costs are covered by the sponsor or funder.
- ☐ Research costs are paid by the sponsor or funding agency; routine health care costs are the responsibility of the participants and/or their healthcare plans.
- ☐ All costs are the responsibility of the participants and/or their healthcare plans.
- ☐ Drug trials sponsored by the National Cancer Institute or other national institutes.
- ☒ There are no costs related to participation.
- ☐ Other

25.1.A. Consent Text: The following financial obligation statement must be contained in the informed consents for this study: (edit only as necessary. If your study has a contract, this language must be consistent with the contract language)
There is no cost to you for taking part in this study.

25.2. Payment for Participation: Describe how much, if any, financial or other form of compensation will be provided to the subject/family. Describe the requisite conditions that must be fulfilled to receive full or partial compensation. Describe the proposed method of timing and disbursement. If children are involved, please specifically address how the compensation will be distributed to children.
Participants will be compensated with a \$50 gift card (in the form of a VISA to facilitate their use among older people) for their participation over the five sessions. This will be given out after a participant has completed all the sessions.

In the case that the participant requests to stop the study due to discomfort during the human-robot interaction sessions, they will still receive compensation (they will be informed of this during consent). This is to avoid any pressure to continue the experiment if the participant feels uncomfortable.

26. Participant Privacy and Data Confidentiality

- 26.1. Privacy Protections:** Privacy is a participant's ability to control how other people see, touch, or obtain information about his/her self. Violations of privacy can involve circumstances such as being photographed or videotaped without consent, being asked personal questions in a public setting, being seen without clothing, being observed while conducting personal behavior, or disclosing information about abortions, HIV status, or illegal drug use.

Select the provisions to protect the privacy of the individual during screening, consenting, and conduct of the research: (check ALL that apply)

- ☒ Research procedures will be conducted in person in a private setting.
- ☒ Data will be captured and reviewed in a private setting.
- ☒ Only authorized research study personnel will be present during research related activities.
- ☒ The collection of information about participants is limited to the amount necessary to achieve aims of the research.
- ☒ Participants will not be approached in a setting or location that may constitute an invasion of privacy or could potentially stigmatize them.
- ☐ Other (specify below)

- 26.2. Confidentiality Precautions:** Confidentiality is an extension of the concept of privacy; it refers to the participant's understanding of, and agreement to, the ways identifiable information will be collected, stored, and shared. Identifiable information can be printed information, electronic information, or visual information such as photographs.

How will the research data/specimens be labeled? (check ALL that apply)

- ☐ Data and/or specimens will be directly labeled with personal identifying information. (Identifiable)
- ☒ Data and/or specimens will be labeled with a code that the research team can link to personal identifying information. (Coded)
- ☐ Data and/or specimens will not be labeled with any personal identifying information, nor with a code that the research team can link to personal identifying information. (Anonymous)
- ☐ Other (explain below)

- 26.3. Study data/specimen will be stored:**

- ☐ Physically
- ☒ Electronically

Which devices will have study data:

- ☒ Local computers/laptops
- ☒ Removable drives (USB, external drives)
- ☒ Local Server(s)
- ☒ External Servers (including cloud based services)

Which company/organization will be providing the service?
Qualtrics, Google, Amazon

Note: Please ensure that the third party's technical infrastructure meets or exceeds the minimum standards set out by ITS.

Please confirm that, at a minimum, the following measures will be taken and enforced:

- ☒ Electronic data will be stored with appropriate electronic safeguards, such as unique usernames/passwords, and limited to authorized study personnel. Dual factor authentication will be used, if feasible.
- ☒ Security software (firewall, antivirus, anti-intrusion) will be installed and regularly updated in all servers, workstations, laptops, and other devices used in the study
- ☒ All computers with access to study data will be scanned regularly (for viruses and spyware, etc.) and problems will be resolved
- ☒ Data stored on a removable drive will be encrypted and have proper access controls
- ☒ Data transfer will be encrypted

Other Measures:

- ☐ Audio and/or video recordings will be transcribed and then will be destroyed



☐ Audio and/or video recordings will be modified to eliminate the possibility that study participants could be identified

☐ Photos or images will be modified to eliminate the possibility that study participants could be identified

26.4. Will identified data and/or specimens be reviewed and/or accessed by a third party institution (such as a study sponsor, federal agency, or another institution)?

☐ Yes ☒ No

26.5. What will happen to the research data and/or specimens at the conclusion of the study? (check ALL that apply)

☒ Direct identifiers and/or the key to the codes will be destroyed upon completion of the research (all data/specimens will be stripped of identifying information and/or the key to codes destroyed, paper documents shredded, electronic files purged, electronic media securely erased)

☐ Retained for study record keeping purposes per institutional policy

☒ Retained by the investigator for future research use

☐ Retained for future research use (submit data or tissue to an existing repository/bank)

☐ Restricted use data will be destroyed or returned to the source

☐ No direct or indirect identifiers are being collected. The anonymous data and/or specimens will be retained at the discretion of the investigator

☐ This research is a clinical trial conducted under FDA regulations. Direct identifiers and/or the key to the codes will be destroyed as directed by the sponsor (IND/IDE holder) in accordance with FDA regulations

☐ The NIH requires that the records be retained for three years following the completion of the study

☒ Other (specify below)

26.5.1 Please specify:

1. Anonymized data (after coding keys and other identifying links are discarded) will be kept indefinitely on a computer in password protected files or in locked cabinets for future research purposes.

2. Consent forms collected for this study will be kept for three years after the study is completed.

26.6. Do you have, or plan to apply for, a Certificate of Confidentiality for this study?

☐ Yes ☒ No

NOTE: NIH-funded research in which identifiable, sensitive information is used, including research that:

- Meets the definition of human subjects' research, including exempt research in which subjects can be identified
- Is collecting or using human bio specimens that are identifiable or that have a risk of being identifiable
- Involves the generation of individual level human genomic data
- Involves any other information that might identify a person

is automatically protected by a certificate of confidentiality from the NIH.

26.7. If audio/video recordings or photographs will be used, will you be anonymizing or de-identifying materials? If so, how will this be done and when.

To get objective measures of attention and behavior our study will use facial affect and gaze from video recordings, precluding the possibility of de-identifying the materials. All identifiable data will be securely stored to prevent non-study personnel from viewing or accessing it.

For audio recordings, we will anonymize the recordings with standard audio privacy-preserving techniques that allow machine learning models to learn effectively but preserve the participants' privacy.

27. Risk/Benefit Assessment - Risks

27.1. Risks, Discomforts and Potential Harms: Describe the risks associated with each research intervention. Include consideration of physical, psychological, social, and other factors. (check all that apply)

☐ Discrimination based on genetic findings.

☐ Some people may find it upsetting to learn that they have certain mutations or errors in genes that could lead to future health problems for themselves or their children.

☒ Some of the questions may make the participant feel uneasy or embarrassed.

☒ There is a small risk that people who are not connected with this study will learn a participant's identity or their personal information.

☐ The participants are providing highly sensitive, personal information in this study. If people not connected with the study learn this information, they could have problems getting a new job, keeping their current job, finding housing, or getting insurance (health, disability, or life insurance). In highly unlikely situations, they could be charged with a crime.

☐ Biomedical risks, including drug, device, biologics, radiation, surgery or other research procedures (please specify).

☐ The research includes the risk or disclosure that a participant may engage in self-harm or attempt suicide.

☐ Venipuncture risks including: mild discomfort (or pain), bruising and swelling around the puncture site, dizziness or fainting, or infection (rare).

☐ Other (specify below)

27.2. Describe the precautions that will be taken to minimize risks/harms. (check all that apply)



- ☒ We will use our best efforts to keep the findings in this study as confidential as possible.
- ☒ Subjects can choose to skip or stop answering any questions that make them uncomfortable.
- ☒ Data will be coded and identity stored separate from data.
- ☐ Data will be collected anonymously.
- ☐ Biomedical precautions, including precautions relating to drugs, devices, biologics, radiation, surgery or other research procedures (please specify).
- ☐ Venipuncture by individuals certified and privileged to perform the procedure.
- ☐ Other (specify below)

28. Risk/Benefit Analysis - Potential Benefits and Alternatives

28.1. Describe any potential for direct benefits to participants in the study: (check all that apply)

- ☒ There are no direct benefits to research participants
- ☐ Improvement in some or all of participants' symptoms
- ☐ Improvement in some or all of participants' survival or longevity
- ☐ Information gained from testing or monitoring procedures
- ☐ Provision of drug or device
- ☐ Reduced side effects
- ☐ Other (explain below)

28.2. Describe potential benefits to society, if any. (check all that apply)

- ☒ The advancement of knowledge
- ☐ A new treatment or therapy for the condition under study
- ☐ None
- ☐ Other (explain below)

28.3. What are the alternatives to participation? (check all that apply)

- ☒ Not participating
- ☐ Continue current medical care for their condition
- ☐ Participation in other research studies
- ☐ Palliative care
- ☐ No treatment or therapy
- ☐ Participate in other subject pool activities
- ☐ Other (specify below)

28.4. Risks in relation to benefits:

- ☐ The potential benefits to the research participants justify exposure of the participants to the risks.
- ☒ The potential benefits to humanity justify exposure of the participants to the risks.
- ☐ Other (specify below)

35. Is the HIPAA Privacy Rule Applicable?

35.1. Do you intend to access, review, collect, use, or disclose Protected Health Information (PHI/ePHI) which includes either patient and/or participant data, in your research? Answer yes if you intend to do any of the following:

- Look at medical records (paper or electronic) to identify potential research participants
- Look at clinic logs to identify potential research participants
- Record demographic information obtained from medical records (paper or electronic)
- Record health information obtained from medical records (paper or electronic)
- Obtain information from laboratory reports, pathology reports, radiology reports or images, or other reports from medical or mental health testing and treatment



- Obtain information from medical billing records
- Record or use medical record numbers or other information that could be used to identify an individual (review the list of HIPAA identifiers below)
 - Name/Initials
 - Street address, city*, county*, precinct*, zip code*, or equivalent geocodes*
 - All elements of dates (except year) directly related to an individual (date of birth, admission date, discharge date, date of death)*
 - Elements of date, including year, for persons 90 or older
 - Telephone number
 - Fax number
 - Electronic mail address
 - Social Security Number
 - Medical record number
 - Health plan identification number
 - Account number
 - Certificate/license number
 - Vehicle identifiers and serial numbers, including license plate number
 - Device identifiers and serial number
 - Web addresses (URLs); Internet IP addresses
 - Biometric identifiers, including finger and voice print
 - Full face photographic images and any comparable images
 - Any other unique identifying number, characteristic, or code*

☐ Yes ☒ No

39. Conflict of Interest Information

39.1. Instructions: **Click on the name of each study team member** and indicate whether they have a potential financial conflict of interest related to this study. If they have a conflict of interest, click on the conflict that has been reported in diSClose. If the conflict is not listed, ask the study team member to update their disclosure information in diSClose before completing this section.

Study Staff	Role	COI Annual Disclosure Status	Conflicts
Maja Mataric	Principal Investigator	Current: due on 8/31/2025	No conflicts identified
Jesse Thomason	Co-Investigator	Current: due on 8/31/2025	No conflicts identified
Avni Hulyalkar	Research Assistant or Associate	No disclosure on file	No conflicts identified
Maria Lima	Research Assistant or Associate	No disclosure on file	No conflicts identified
Alethea Nagahara	Research Assistant or Associate	No disclosure on file	No conflicts identified
Amy O'Connell	Research Assistant or Associate	No disclosure on file	No conflicts identified
Leticia Pinto Alva	Research Assistant or Associate	No disclosure on file	No conflicts identified
Feiyang Zhou	Research Assistant or Associate	No disclosure on file	No conflicts identified

CHLA policy requires all study team members to have a current COI disclosure in DiSClose. If a study team member has not completed a disclosure form or the disclosure form has expired, you must submit a new disclosure in DiSClose. Hover over the annual disclosure status for more details.

Please note that first time users of diSClose will need to create a diSClose account using the "Login with CHLA Okta" option to ensure disclosure data from diSClose appears in this Section.

40. Additional Supporting Documents



IMPORTANT!
Care must be taken to ensure that any information containing PHI is not entered into the system. Do not enter or upload documents with personal identifiers (i.e., patient names and MRNs) into iStar. Any direct identifiers should be redacted from documents prior to submission to the IRB.

40.1. Attach any other documents that have not been specifically requested in previous questions, but are needed for IRB review.

Name Version Modified
There are no items to display

40.2. If there is any additional information that you wish to communicate about the application include it below. Please note, this section should not be used instead of the standard application items.

45. Study Locations



45.1. Identify the locations where the research activities described in this application will be performed (check all that apply):

- ☐ USC HSC - Health Sciences Associated Locations
- ☒ USC UPC - University Park Associated Locations
- ☐ Los Angeles General Medical Center Locations
- ☐ Other

45a. USC Location(s)

This screen is required if you indicated a USC Location (Question 45.1.)

▼ USC UPC Locations

45a.2. UPC Locations (check all that apply and provide detail where indicated):
Location

- ☒ Campus location (includes ISI and ICT)
- ☒ Off-campus location

45a.2.1. Specify Campus Location:
Ronald Tutor Hall (RTH) 423, 422 and 406.

45a.2.2. Specify Off-Campus Location:

50. Required Approvals (Post-Submission)

This screen indicates the approvals received once the proposal has been submitted.

50.1. Pending Division/Department Approvals:

Name Division/Department Parent Campus
There are no items to display

50.2. Received Division/Department Approvals:

Name Division/Department Parent Campus
There are no items to display

99. Instructions for Submission

You have reached the end of the application. When you are sure of the content, the following steps may be taken to submit your application for review.

1. Click the "Finish" button on the top or bottom application navigator bar to return to the workspace.
2. Use the Hide/Show Errors above to determine that all sections of the application are filled out correctly.
3. Use the "Send Study Ready Notification" activity to send an email to the Principal Investigator and Co-Investigators with instructions for reviewing and submitting the application.
4. All listed Co-Investigators (indicated in item 2.1.) must use the "Agree to Participate" activity and answer yes.
5. Once all the Co-Investigators have agreed to participate, the Principal Investigator (indicated in item 2.1.) can submit the application by using the "Submit Application to _____", where _____ indicates the IRB you are submitting to.
6. The PI will have to check the PI endorsement box. The PI will also have to check the student endorsement box if it is applicable.
7. The application is submitted. The state indicator in the top left of the workspace will no longer display Pre Submission.
8. The PI and Study Contact Person will receive an email confirming the application has been submitted.

4.4. Funding Source

Please enter the fields below and click 'OK' when done.

4.4.1. * Sponsor Name:

Use "%" to enhance filter criteria. For example, to narrow your query and find the sponsor "USC (University of Southern California)", you can type "%usc" or "%university of southern".
National Science Foundation

If you cannot find the entity in the list above, click here: ☐

4.4.2. * Named Principal Investigator:

Maja Mataric

4.4.3. Institution awarded the grant-award:

USC



- 4.4.4.

Grant-award number provided by the Sponsor:
NSF IIS-1925083
- 4.4.5.

Title of the Funding Project, if applicable:
Communicate, Share, Adapt: A Mixed Reality Framework for Facilitating Robot Integration and Customization
- 4.4.6.

* Type of Funding:
Federal: Grant *
- 4.4.7.

Attach a copy of the proposal/contract/grant with the project budget. Copies of the documents submitted to the funding agency, including the budget detail, should be submitted. (salary information need not be displayed or included.)

Name	Version	Modified
There are no items to display		

4.4. Funding Source

Please enter the fields below and click 'OK' when done.

- 4.4.1.

* Sponsor Name:
Use "%" to enhance filter criteria. For example, to narrow your query and find the sponsor "USC (University of Southern California)", you can type "%usc" or "%university of southern".
NSF (National Science Foundation)

If you cannot find the entity in the list above, click here: ☐
- 4.4.2.

* Named Principal Investigator:
Maja Mataric
- 4.4.3.

Institution awarded the grant-award:
5445095084
- 4.4.4.

Grant-award number provided by the Sponsor:
IIS-1925083
- 4.4.5.

Title of the Funding Project, if applicable:
NRI: FND: Communicate, Share, Adapt: A Mixed Reality Framework for Facilitating Robot Integration and Customization
- 4.4.6.

* Type of Funding:
Federal: Grant *
- 4.4.7.

Attach a copy of the proposal/contract/grant with the project budget. Copies of the documents submitted to the funding agency, including the budget detail, should be submitted. (salary information need not be displayed or included.)

Name	Version	Modified
There are no items to display		

