

Towards Natural Human-Robot Dialogue Through a Reflective LLM

Gaëlle Snoeckx

Student number: 02107325

Supervisor: Prof. dr. Tony Belpaeme

Counsellors: Maria Jose Pinto Bernal, Ir. Eva Verhelst

Master's dissertation submitted in order to obtain the degree of Master of Science in Biomedical Engineering

Academic year 2025-2026

TOWARDS NATURAL HUMAN-ROBOT DIALOGUE THROUGH A REFLECTIVE LLM

GAËLLE SNOECKX
STUDENT NUMBER: 02107325

Supervisor: Prof. dr. Tony Belpaeme
Counsellor: Maria Jose Pinto Bernal, Ir. Eva Verhelst

Master's dissertation submitted in order to obtain the academic degree of Master of Science in Biomedical Engineering

Acknowledgements

First of all, I would like to express my gratitude to my daily supervisors, Maria José Pinto Bernal and Eva Verhelst, for their guidance and dedication throughout this research project. While our bi-weekly meetings provided a consistent framework and kept the project on track, they were always approachable and ready to answer any questions that arose in between. I would also like to thank them for carefully proofreading this thesis. Their thoughtful feedback and attention to detail greatly improved its quality.

My thanks also go to my supervisor, Prof. Dr. Tony Belpaeme, for his academic oversight and for giving me the opportunity to undertake this research.

I would like to thank all the participants who took part in this study. Their willingness to contribute their time and insights was essential to the research and is greatly appreciated.

Finally, I would like to thank my family and friends for their constant support and encouragement throughout my studies. Their patience and belief in me during the more challenging moments made all the difference.

With gratitude for my time here in Ghent, I look forward to what lies ahead.

Gaëlle Snoeckx, May 2026

Admission to loan

“The author(s) gives (give) permission to make this master’s dissertation available for consultation and to copy parts of this master’s dissertation for personal use. In the case of any other use, the copyright terms have to be respected, in particular with regard to the obligation to state expressly the source when quoting results from this master’s dissertation.”

Toegang tot bruikleen

“De auteur(s) geeft (geven) de toelating dit afstudeerwerk voor consultatie beschikbaar te stellen en delen van het afstudeerwerk te kopiëren voor persoonlijk gebruik. Elk ander gebruik valt onder de beperkingen van het auteursrecht, in het bijzonder met betrekking tot de verplichting de bron uitdrukkelijk te vermelden bij het aanhalen van resultaten uit dit afstudeerwerk.”

Thesis as part of an exam

“This master’s dissertation is part of an exam. Any comments formulated by the assessment committee during the oral presentation of the master’s dissertation are not included in this text.”

Gaëlle Snoeckx, May 2026

Towards Natural Human-Robot Dialogue Through a Reflective LLM

by
GAËLLE SNOECKX

Master's dissertation submitted in order to obtain the academic degree of Master of
Science in Biomedical Engineering

Supervisor: Prof. dr. Tony Belpaeme
Counsellor: Maria Jose Pinto Bernal, Ir. Eva Verhelst

Faculty of Engineering and Architecture
Ghent University

Department of AIRO
Chair: prof. dr. Tony Belpaeme, Maria Jose Pinto Bernal, ir. Eva Verhelst, Giulio
Antonio Abbo

Abstract - Loneliness among older adults in care settings is a growing public health problem linked to cognitive decline, depression and premature death. Social robots offer a promising way to provide consistent social engagement, but current systems struggle to maintain long-term engagement. Large language models (LLMs) improve fluency but suffer from structural shortcomings, including no persistent memory, no proactive initiative and no stable persona. This thesis therefore proposes the Dialogue Analyser, a supervisory LLM layer that runs in parallel with a primary conversational agent, providing turn-level guidance to address these shortcomings in real time. The system was evaluated in a controlled within-subjects experiment involving 16 participants and the humanoid robot Pepper. Outcomes were assessed through subjective self-report scales, behavioural latency metrics and an LLM-as-a-Judge transcript analysis. The results showed that the Analyser significantly improved perceived conversational proactivity and suggested improved emotional responsiveness during stressful discussions. The LLM-as-a-Judge analysis revealed that the Analyser entirely eliminated question stacking on emotionally charged turns in the stress condition, while latency metrics indicated a trend toward faster human response times. However, the Analyser did not affect the robot's perceived warmth, comfort or enjoyment. These findings demonstrate that a supervisory LLM layer can improve specific aspects of social robot behaviour, though replication with the intended elderly population and longer interaction protocols remains necessary.

Keywords - Social Robots, Long-Term Engagement, Reflective LLM Layer

Towards Natural Human-Robot Dialogue Through a Reflective LLM

Gaëlle Snoeckx

Prof. Dr. Tony Belpaeme

Maria Jose Pinto Bernal, Ir. Eva Verhelst

Abstract—Although current LLM-powered social robots can hold fluent conversations, they struggle to maintain meaningful long-term engagement. They forget previous interactions, never take the initiative and don’t have a stable personality. This thesis presents the Dialogue Analyser, a real-time supervisory LLM layer that runs in parallel with a primary conversational agent. It steers responses through pre- and post-analysis and proactive intervention, without replacing the underlying generative model. The system was evaluated in a controlled 2x2 within-subjects experiment involving 16 participants and the humanoid robot Pepper [9]. The experiment compared a baseline GPT-4o system with the same system with the Dialogue Analyser active, across both positive and stressful conversation topics. The results showed that the Analyser significantly improved perceived conversational proactivity ($p_{FDR} = .009$, $d = +0.95$) and produced a significant crossover interaction on perceived emotional responsiveness, with the Analyser outperforming the Baseline during stressful conversations. In that condition, the Analyser eliminated question stacking entirely on emotionally charged turns (0% versus 8.5% for the Baseline). These findings suggest that a reflective supervisory layer can enhance specific aspects of social robot behaviour, such as proactivity, emotional attunement in stressful situations and perceived understanding.

Index Terms—Social Robots, Long-term Engagement, Reflective LLM Layer

I. INTRODUCTION

Loneliness among older adults in care settings is a growing public health crisis linked to cognitive decline, dementia and premature death [1, 2]. Social robots offer a promising solution, as they are available on demand and do not require an external human partner. However, traditional systems rely on pre-scripted dialogue, which limits conversational depth [4, 6]. While large language models (LLMs) improve fluency, they introduce three structural shortcomings: a lack of persistent memory across sessions [3], a lack of proactive initiative [5] and an inability to maintain a stable persona [3]. When a robot speaks naturally, users instinctively apply the same social expectations they have for human companions, expecting the robot to remember them, recognise their emotional state and engage with them proactively [6, 7]. Current systems cannot meet these expectations, producing interactions that feel fluent in the moment but reset-like across sessions [6].

This thesis proposes the Dialogue Analyser as a solution to this issue. Rather than replacing the primary language model, the module runs in parallel as a real-time supervisory layer inspired by the dual-process evaluation framework of Constitutional AI [8], monitoring the conversation and steering responses through pre- and post-analysis and proactive intervention. The aim is to

create a robot that feels like a consistent, emotionally present companion that enhances long-term engagement.

Two research questions guide the evaluation: Firstly, does the Dialogue Analyser improve the conversational experience compared to a baseline system in terms of perceived empathy, emotional appropriateness, conversational naturalness, felt comfort and interaction enjoyment? Secondly, does the effect of the Analyser depend on the emotional load of the conversation? In other words, is its benefit greater when participants are discussing something stressful rather than something positive?

II. SYSTEM ARCHITECTURE

The Dialogue Analyser is built on the open-source TEN Framework, which is a real-time voice pipeline that processes audio, text and contextual data as a directed graph of asynchronous, independent nodes [10]. This architecture enables speech transcription, memory retrieval and social error analysis to run in parallel, without any task blocking another [11]. Once a user’s turn has been confirmed, the primary language model (GPT-4o) waits for the analyser to return a guidance signal before generating its response. If the signal arrives within the timeout window, it is included in the prompt. Rather than rewriting the next turn, the Analyser steers responses through directional instructions, such as ‘*Ask if they are still looking forward to the hiking trip they mentioned yesterday.*’ preserving the naturalness of the primary model’s output. An overview is given in Figure 1. The Analyser operates across three parallel phases.

A. Pre-Analysis (Prevention)

The pre-analysis phase examines the user’s input before the robot responds and has two primary functions. First, it checks for safety signals, such as crisis language or self-harm cues. This ensures that the robot never probes further when a user expresses acute distress. It also applies a deterministic question-stacking rule. If the robot has asked questions in three or more consecutive turns and the user’s replies were brief, a correction is enforced regardless of other signals. This redirects the robot towards making a warm statement or reflection rather than asking another question.

B. Post-Analysis (Correction)

The post-analysis phase evaluates the robot’s previous response in light of what the user says next, which is something that the pre-analysis cannot do. For example, if a user says, ‘My

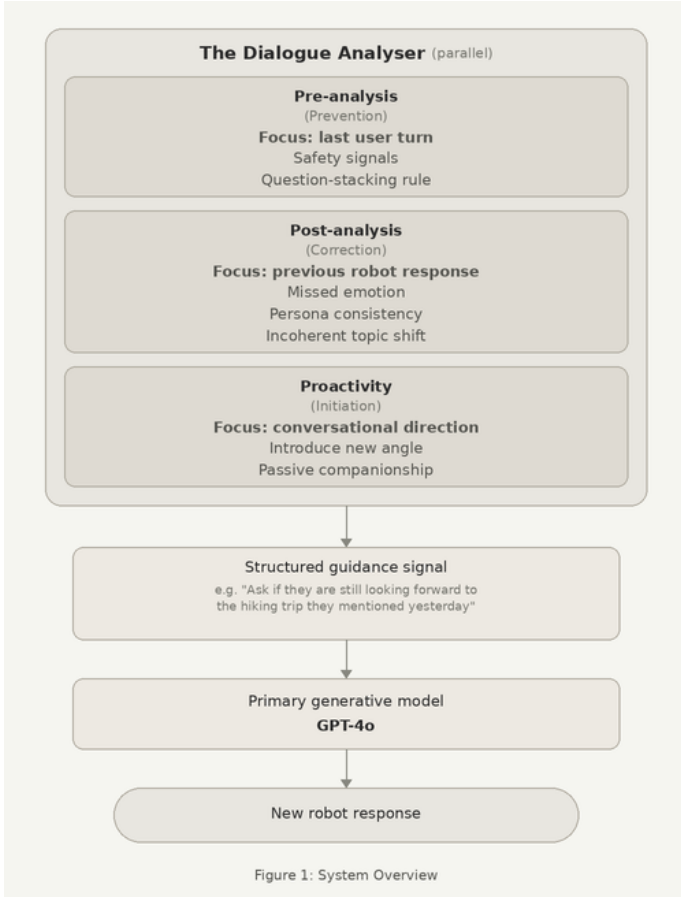


Figure 1: System Overview

sister called yesterday. We hadn't spoken in years,' and the robot replies, 'That's wonderful! What else did you do yesterday?', the pre-analysis would see nothing unusual. The user has answered and the conversation has moved on. Post-analysis, however, recognises that a personal moment has been overlooked with a single word and an immediate generic follow-up. It would flag this as a missed emotional moment and instruct the robot to return to it in the next turn. Post-analysis monitors for three types of failure:

- A missed emotion flag is raised when the user expresses something emotionally meaningful, but the robot moves past it without acknowledging it.
- A persona consistency flag is raised when the robot's tone conflicts with its defined persona, such as detecting the user's emotional distress but reacting too cheerfully to it.
- An incoherent topic shift flag is raised when the robot changes the subject abruptly or revisits something that the user has already moved on from.

The intervention threshold is intentionally high, firing only when the analyser is confident that a response has caused concrete social failure.

C. Proactivity (Initiation)

The proactivity module operates alongside the other two phases. Rather than waiting for the user to drive the conver-

sation forward, it determines whether the robot should raise an unacknowledged detail, introduce a new angle or withdraw entirely and provide quiet companionship when the user's social energy is low. A cooldown system prevents the same topic from being raised more than once per session, unless the user actively engages with it.

III. EXPERIMENTAL METHODOLOGY

The system was evaluated in a controlled 2x2 within-subjects experiment that crossed two independent variables: system condition (Baseline versus Dialogue Analyser active) and conversation topic (positive versus stressful). A total of 16 participants (22.2 ± 1.2 years) were recruited. Each participant completed two eight-minute conversations with the humanoid robot Pepper, which embodied the companion persona Noa. The order of topics and conditions was fully counterbalanced across four groups to control for carry-over effects and participants were not informed which condition was active during which session. Outcomes were measured using three complementary methods. Subjective experience was captured using the Perceived Empathy of Technology Scale (PETS) [13] and the Robotic Social Attributes Scale (RoSAS) [14], as well as single-item measures of naturalness, felt comfort and enjoyment and five custom conversational dynamics items targeting proactivity and timing. Behavioural data were extracted from session silence logs to calculate human turn duration and robot-to-human and human-to-robot response latencies. Objective, turn-level quality was evaluated using an LLM-as-a-Judge that rated each robot response using the Empathic Communication Coding System (ECCS) [12] and extracted metrics such as the multi-question and grounded reference rates while remaining blind to the experimental condition. All hypothesis tests were corrected for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) procedure and all findings are treated as exploratory given the sample size.

IV. RESULTS

Three findings stand out. Conversational proactivity was the most consistent and statistically robust finding. The Custom Proactivity Index was the only outcome to survive FDR correction under RQ1 ($p_{FDR} = .009$, $d = +0.95$), with participants rating the Analyser significantly higher for active conversational guidance ($M = 3.75$ versus $M = 3.40$). This effect was consistent across both topic conditions and at the individual level: 11 out of 16 participants showed a positive difference between the Analyser and the Baseline, with only one showing a negative difference. Item-level analysis identified 'kept conversation flowing' as the dominant driver ($RBC = +0.71$). The only exception to this trend was the timing item 'didn't rush topic changes' ($d = -0.43$), suggesting that, while the Analyser made the robot more directive, it occasionally changed topics before participants were ready to move on.

The second key finding was context-dependent emotional responsiveness. Under RQ2, the PETS Emotional Response produced a significant crossover interaction ($F(1,14)=12.68$, $p_{FDR}=.025$, $\eta_p^2=.475$), as can be seen in Figure 2. The analyser outperformed the baseline on perceived emotional responsiveness during stressful conversations ($M=4.00$ versus $M=3.88$), while the reverse was

true for positive topics ($M=3.60$ versus $M=3.77$). The objective LLM-as-a-Judge analysis provides a behavioural explanation for this pattern. During emotionally loaded turns in the stress condition, the Analyser never asked more than one question per response (0 out of 82 emotional interactions, 0%), whereas the Baseline did so in 8.5% of cases (6 out of 71). This difference was significant and survived FDR correction (Fisher’s exact test, $p_{FDR} = .036$). The forced-choice preference data support this conclusion: 77% of participants who expressed a preference felt more understood in the Analyser session ($p_{raw} = .046$).

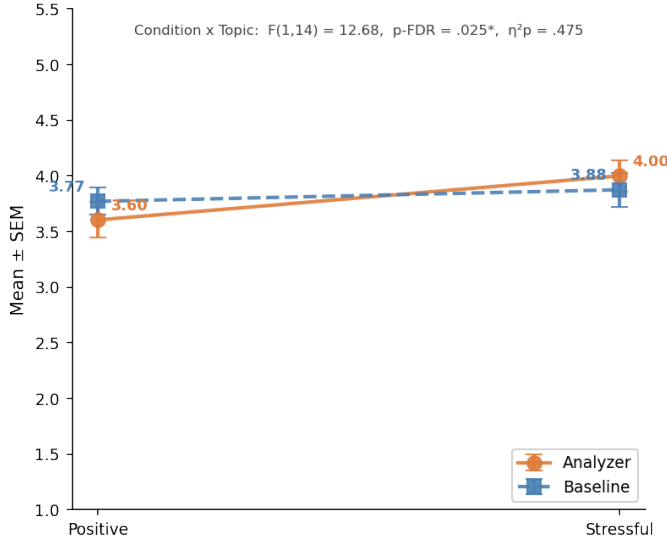


Figure 2: Subjective: PETS Emotional Response

The parallel architecture was associated with significantly faster robot response times ($H \rightarrow R$ latency: $M = 7,045$ ms versus $M = 8,843$ ms, $p_{FDR} = .009$; $RBC = -0.809$), though this may partly reflect API variability across sessions rather than an architectural property. There was also a marginally significant trend, suggesting that participants responded more quickly after the robot had finished speaking in the Analyser condition ($R \rightarrow H$ latency: $M = 1,644$ ms versus $M = 2,033$ ms; $p_{FDR} = .071$; $d = -0.625$), which is a medium to large effect.

V. DISCUSSION AND CONCLUSION

The Dialogue Analyser did not make the robot warmer, more enjoyable or more comfortable overall. However, it did significantly improve conversational proactivity and enhanced perceived emotional responsiveness during stressful exchanges. During these emotionally loaded turns, the Analyser prevented question stacking entirely.

Two limitations require acknowledgement. Firstly, the proactivity module occasionally redirected conversations before participants were ready to move on, resulting in negative timing and suggesting the need for an adaptive pacing mechanism in the future. Secondly, the sample of engineering students aged 20-24 differs from the intended deployment population of elderly care home residents, so replication with a properly powered target sample is needed.

A functioning prototype and a responsibly deployed system are

not the same thing. The risks of substituting human contact, reinforcing false memories due to hallucinations in cognitively vulnerable users and collecting intimate personal data without meaningful consent must be resolved before these systems can be used outside of the research lab. Social robots can complement human care meaningfully, but only if the line between technology that supports people and technology that exploits them is taken into account.

VI. ACKNOWLEDGEMENTS

I would like to thank my daily supervisors, Maria José Pinto Bernal and Ir. Eva Verhelst, for their guidance and Prof. Dr. Tony Belpaeme for giving me the opportunity to undertake this research. I am also grateful to all participants for their time and willingness to participate and for my family and friends for their constant support.

REFERENCES

- [1] G. Coppens and M. Petrovic, “Maatschappelijke impact van depressie bij ouderen met dementie,” *Tijdschrift voor Geneeskunde*, 2011.
- [2] A. R. Sutin, Y. Stephan, M. Luchetti, and A. Terracciano, “Loneliness and risk of dementia,” *The Journals of Gerontology: Series B*, 2020.
- [3] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, “Personalizing Dialogue Agents: I have a dog, do you have pets too?,” *arXiv*, 2018.
- [4] Ruben Janssens, Andre Pereira, Gabriel Skantzé, Bahar Irfan, and Tony Belpaeme, “Online Prediction of User Enjoyment in Human-Robot Dialogue with LLMs,” 2024.
- [5] Chapa Sirithunge, A. G. Buddhika P. Jayasekara, and D. P. Chandima, “An Evaluation of Human Conversational Preferences in Social Human-Robot Interaction,” *Applied Bionics and Biomechanics*, 2021.
- [6] Jesse Fox and Andrew Gambino, “Relationship Development with Humanoid Social Robots: Applying Interpersonal Theories to Human-Robot Interaction,” *Cyberpsychology, Behavior, and Social Networking*, 2021.
- [7] Aurora An-Lin Hu, Dimity Crisp, Sharni Konrad, Damith Herath, and Janie Busby Grant, “From Failing to Exceeding User Expectations: A Large-Scale Study Exploring the Role of Expectation Gap Magnitude and Direction in Robot Adoption Intention,” *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion '26)*, ACM, 2026.
- [8] Yuntao Bai et al., “Constitutional AI: Harmlessness from AI Feedback,” *arXiv*, 2022.
- [9] A. K. Pandey and R. Gelin, “A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of Its Kind,” *IEEE Robotics & Automation Magazine*, vol. 25, no. 3, pp. 40–48, 2018, doi: 10.1109/MRA.2018.2833157.
- [10] TEN Ecosystem, Agora. *TEN Framework Documentation: Vision*. 2024. URL: <https://theten.ai/docs#vision> (visited on 05/13/2026).
- [11] M. J. Pinto and T. Belpaeme, “Predictive Turn-Taking: Leveraging Language Models to Anticipate Turn Transitions in Human-Robot Dialogue,” in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, Aug. 2024. URL: <https://ieeexplore.ieee.org/document/10731379/>.
- [12] M. J. Pinto-Bernal, K. Boels, and T. Belpaeme, “Fostering Social Connection and Emotional Engagement to Mitigate Social Isolation in Care Homes Using Empathetic Human-Robot Conversation,” Preprint via Research Square, 2024. URL: <https://doi.org/10.21203/rs.3.rs-6811088/v1>.
- [13] M. Schmidmaier et al., “Perceived Empathy of Technology Scale (PETS): Measuring Empathy of Systems Toward the User,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, Honolulu, HI, USA: ACM, May 2024, pp. 1–18. DOI: 10.1145/3613904.3642035. URL: <https://dl.acm.org/doi/10.1145/3613904.3642035>.
- [14] C. M. Carpinella et al., “The Robotic Social Attributes Scale (RoSAS): Development and Validation,” in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*, Vienna, Austria: ACM, Mar. 2017, pp. 254–262. DOI: 10.1145/2909824.3020208. URL: <https://dl.acm.org/doi/10.1145/2909824.3020208>.

Table of Contents

Acknowledgements	3
Abstract	6
Extended Abstract	10
List of Abbreviations	1
1 Introduction	2
2 Literature Overview	4
2.1 The Psychology and Evolution of Robotic Companionship	4
2.1.1 The Expectation Gap and the CASA Paradigm	4
2.1.2 The Fragmentation Gap: from Task Execution to Open-Domain Dialogue	5
2.2 Approaches to Relational Coherence	6
2.2.1 Episodic and Semantic Memory Architectures	6
2.2.2 Contextual Proactivity and Anticipatory Frameworks	6
2.2.3 Persona Modeling and Identity Evolution	6
2.3 Supervisory Frameworks and Model Self Evaluation	7
2.4 Contribution	8
3 Architecture	9
3.1 Bridging the Gap: The Dialogue Analyser	9
3.2 System Overview	9
3.3 The Dialogue Analyser	11
3.3.1 Architecture and Delivery	11
3.3.2 Pre-Analysis: Prevention Before the Response	12
3.3.3 Post-Analysis: Correction After Robot Response	13
3.3.4 Proactivity: Initiating rather than Reacting	13
3.3.5 Persona and Prompt Architecture	14
4 Experimental design	15
4.1 Experimental Conditions	15
4.1.1 System Condition	15
4.1.2 Conversation Topic	15
4.1.3 Counterbalancing	16
4.1.4 Participants	16
4.2 Settings	16
4.3 Procedure	17
4.4 Metrics	18
4.4.1 Perceived Empathy	18
4.4.2 Emotional Appropriateness	18
4.4.3 Conversational Naturalness	19
4.4.4 Interaction Enjoyment	19
4.4.5 Felt Comfort	19
4.4.6 Robot Social Attributes (RoSAS warmth and discomfort) [13]	19
4.4.7 Conversational Dynamics	20
4.4.8 LLM-as-a-Judge	20
4.5 Data Analysis Plan	20
4.5.1 Scale Reliability and Scoring	20
4.5.2 Normality Assessment and Multiple Comparisons Correction	21

4.5.3	Statistical Tests	21
5	Results	23
5.1	Data Preprocessing and Diagnostics	23
5.1.1	Scale Reliability Diagnostics	23
5.1.2	Manipulation Check	24
5.2	RQ1: Main Effect of the Dialogue Analyser	25
5.3	RQ2: Condition x Topic Interaction	26
5.4	Secondary Analyses and Methodological Checks	27
5.4.1	Forced-Choice Preference	27
5.4.2	Within-Person Consistency	27
5.4.3	Session Order Effects	28
5.4.4	Demographic Covariates	28
5.5	Conversational Dynamics and Latency	28
6	Discussion	30
6.1	Effect on Timing and Proactivity	30
6.2	Context-Dependent Emotional Attunement	31
6.3	Latency Dynamics and Architectural Efficiency in Parallel Processing	32
6.4	Conclusion	33
7	Conclusion and future work	35
7.1	Limitations	35
7.1.1	Methodological and Experimental Limitations	35
7.1.2	Architectural Shortcomings	36
7.2	Future Work	36
7.3	Conclusion	37
8	Societal Reflection	38
A	Questionnaires	46
A.1	Per- and Post-Session Questionnaire	46
A.2	Demographics Questionnaire	50
B	Use of Generative AI	52

List of Abbreviations

AI	Artificial Intelligence
ANOVA	Analysis of Variance
API	Application Programming Interface
CASA	Computers Are Social Actors
ECCS	Empathic Communication Coding System
FDR	False Discovery Rate
GDPR	General Data Protection Regulation
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
HRI	Human-Robot Interaction
LLM	Large Language Model
MOS	Mean Opinion Score
OCEAN	Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
PENJ	Perceived Enjoyment
PETS	Perceived Empathy of Technology Scale
PETS-ER	Perceived Empathy of Technology Scale - Emotional Response
RBC	Rank-Biserial Correlation
RoSAS	Robotic Social Attributes Scale
RQ	Research Question
SD	Standard Deviation
SEM	Standard Error of the Mean

Introduction

Loneliness and social isolation among older adults are a global public health crisis. Prolonged isolation has been consistently linked to depression, cognitive decline, dementia, cardiovascular disease and premature death. [74]. Between a third and half of all older adults experience loneliness or social isolation [41]. In Belgium, this affects around 500,000 adults aged 65 and older, representing 22% of this age group [22]. This widespread feeling of loneliness has serious clinical implications, increasing the risk of developing dementia by approximately 50% [30] and stroke by 25% [71]. This issue is difficult to address because the harm is not driven by the number of social contacts a person has. It is the subjective perception of being connected that most strongly predicts health outcomes [19]. Someone who sees family regularly but feels lonely is at the same risk as someone who sees nobody at all. Subjective loneliness and objective social isolation increase the likelihood of early mortality by 26% and 29% respectively [33].

This issue is most severe for older adults in care homes. With declining fertility rates and rising life expectancy, today's older adults have smaller family networks and fewer relatives available for regular visits [77]. In care homes, this issue is made worse by the physical separation from the wider community and the cognitive or health-related decline that limits meaningful relationships with others. As a result, the responsibility for social interaction falls largely on care staff, who are already overwhelmed by their primary caregiving duties. This is a burden that the system does not have the resources to sustain, especially when considering that Belgium alone is projected to have a deficit of around 120,000 healthcare and elderly care workers by 2040 [2]. Video calls offer a partial solution and are generally well received, but they depend on a willing family member being available, as well as the resident having the cognitive capacity to manage the interaction [81]. For many residents, neither condition is reliably met. On top of that, organising these calls places an extra scheduling burden on the understaffed facilities [52]. Although passive alternatives such as television and radio provide stimulation, they lack the active, cooperative elements necessary for building social connections [62, 76]. In contrast, interactive experiences have been shown to significantly reduce loneliness and increase feelings of connectedness [38]. Therefore, an interactive solution that does not depend on external availability and does not impose practical barriers is needed, one that can provide consistent, on-demand social engagement. Socially interactive robots are a promising solution.

Social robots are physical agents designed to interact with humans in a natural and socially aware way. Unlike passive technologies, they are available on demand, do not require an external human partner and can sustain a continuous conversational presence. Paro, Pepper, and Nao are examples of systems that have shown real success in this area. Studies report that they help lower anxiety and increase feelings of companionship among older adults [17, 48]. These results are encouraging, but these systems are also still limited. Traditional social robots rely heavily on pre-scripted dialogue and rule-based interaction systems, which limits the depth and naturalness of the conversations they can sustain [37].

They can recognise broad emotional states, such as sadness or agitation, but they cannot interpret the deeper psychological context behind them. This means that they cannot adapt in a meaningful way when a resident shares something unexpected, painful or personally significant [25, 50]. This is important because physical presence alone is not the key factor in effective social companionship. It is emotional companionship that reduces loneliness [23, 31]. A robot that responds from a fixed script cannot reliably produce this perception, regardless of how lifelike its appearance or how expressive its movements are. The limiting factor is not the hardware, but the rigidity of the language used.

Large Language Models (LLMs) address this limitation (partially). As they are trained on extensive datasets, these models possess an ability to understand context, make logical deductions and integrate diverse information from multiple sources [84]. They demonstrate remarkable adaptability, successfully executing complex instructions without the need for task-specific programming [10, 44]. This flexibility enables artificial agents to handle unpredictable open-domain dialogue and liberates them from predetermined conversation pathways [37, 70]. As a result, the nature of interaction shifts from mechanical exchanges to continuous, dynamic collaboration [40, 72].

LLMs make robotic conversation more fluent, but this creates a new problem. When a robot speaks naturally and has a human-like appearance, users instinctively apply the same social expectations they have for human companions. They expect to be recognised and remembered and to be proactively engaged with [25, 34]. Yet the language model alone has no mechanism to meet these expectations. Three concrete shortcomings explain why.

First, language models process each session within a bounded context. Without a structured mechanism to preserve what was discussed in prior interactions, earlier information is forgotten, topics are repeated and the sense of having an accumulated shared history that underlies trust is never established [82]. Secondly, current systems remain entirely passive, requiring the user to initiate every interaction. For socially withdrawn, physically fatigued, or uncertain older adults who do not know how to begin, a system that waits passively for initiation forces the most vulnerable people to carry the full burden of sustaining the relationship [37, 68]. Thirdly, without explicit mechanisms to stabilise the robot’s relational identity across sessions, the tone, register and apparent personality of the system can shift in subtle yet cumulative ways, which undermines the consistency required for affective trust to develop [25, 82].

These three deficiencies are not separate engineering problems. A system with perfect recall but no capacity for initiative remains a passive tool. A robot that engages proactively but forgets previous conversations produces interactions that feel superficial rather than warm. Together, these issues point to a single underlying absence. The current architecture has no dedicated module to manage how a relationship evolves continuously across multiple sessions. Solving this specific structural problem is the focus of this thesis.

This thesis proposes the Dialogue Analyser as an architectural response to the structural gap identified above. The Analyser operates in parallel with the primary language model, rather than replacing it. Inspired by the dual-process evaluation framework of Constitutional AI [5], it operates as a reflective supervisory layer positioned above the base model. This separation is necessary because primary language models are known to struggle with self-critiquing complex social rules, often rephrasing their own errors with equal confidence rather than correcting them [35]. Therefore, an independent layer is required to evaluate the interaction from an external perspective.

The Dialogue Analyser runs alongside the primary language model, monitoring the conversation in real time. When it detects a social error or missed opportunity, it sends a brief corrective instruction to the main agent before the next response is generated. It does not rewrite the robot’s replies, but rather steers them.

The Analyser aims to address the three shortcomings identified above in three phases. Before responding, the Analyser reviews what the user has just said and sends a short guideline to the main agent, ensuring that the upcoming output is emotionally appropriate and consistent with the robot’s established persona. After responding, it evaluates the reply for relational missteps, such as missed emotional cues or inconsistencies in its personality. Throughout the session, the Analyser also monitors the conversation for opportunities to steer it proactively. To do so, it draws on both the current exchange and stored memory from previous sessions.

The result should be a robot that feels like a consistent and emotionally present companion across weeks of interaction, rather than a fluent but forgetful conversational tool.

The remainder of this thesis is structured as follows: Chapter 2 provides a review of the relevant literature. Chapter 3 provides a detailed overview of the system architecture. Chapter 4 outlines the experimental methodology. Chapter 5 presents the results, which are discussed in Chapter 6. Chapter 7 summarises the conclusions, limitations and directions for future work. Finally, chapter 8 presents the social reflection.

Literature Overview

This chapter establishes the theoretical basis for the system proposed in this thesis. It starts from the psychology of why social robots so often fall short. It then reviews the state of the art in memory, proactivity and persona modeling before establishing why an independent supervisory layer is required to bring these components together in real time.

2.1 The Psychology and Evolution of Robotic Companionship

To understand why current social robots fall short, their technical limitations and the psychological expectations they trigger must be examined. This section traces the gap from its cognitive origins to its architectural consequences, exploring how these elements interconnect to define the precise relational criteria that social robots must fulfil in order to succeed.

2.1.1 The Expectation Gap and the CASA Paradigm

When people first encounter a social robot, they tend to respond positively. The novelty of interacting with a machine that can speak, move and respond generates curiosity and engagement. However, once this initial curiosity dissipates, users begin to judge the system against strict social standards, that most current systems are not able to meet [3]. The reason for this lies in a cognitive tendency known as the Computers Are Social Actors (CASA) paradigm. When a machine takes on a humanoid form and communicates in natural language, users automatically apply human social expectations without consciously choosing to do so. The robot is no longer evaluated as a tool, but it is instead evaluated as a social partner. However, a social partner that fails to follow basic conversational norms, such as taking turns appropriately, responding at the right time, or acknowledging what has just been said, is experienced as unsettling or untrustworthy [3, 7, 56]. This sense of unnaturalness does not arise from physical awkwardness or mechanical noise, but from the absence of social cognition [7]. For example, a robot lacks social cognition when it cheerfully asks a follow-up question immediately after a user shares distressing news, failing to recognise that the social context requires a shift to a more supportive tone.

To understand what users actually require from a social robot, it is necessary to delve into four psychological layers that make a robot a companion.

1. **Coherence:** A robot must remain logically consistent within and across conversations, connecting new exchanges to previous discussions to sustain a continuous dialogue rather than a sequence of isolated responses [15, 29]. Without coherence, even a technically fluent robot will seem inconsistent, because the user cannot develop an understanding of what the robot knows, remembers or understands. Coherence is what maintains the impression of a thinking presence.
2. **Empathy:** Users need to feel that the robot understands what they are saying, rather than just producing an appropriate-sounding response. This distinction matters because empathy operates at the level of perceived understanding, not linguistic accuracy [4, 42, 59]. For example, a user who shares that they are grieving does not need a technically correct statement about loss. They need a response that acknowledges the specific weight of what they have just said. A reply that is too flat suggests that the robot has not properly registered the emotional content. Similarly, a reply that is exaggerated suggests that the robot is acting rather than feeling empathy. Effective empathic

responses find a balance between these extremes: they reflect back to what the user shared in a way that feels appropriate, specific and sincere.

3. **Personalisation:** A companion is not just someone who responds well to anyone. They respond specifically to you. In order to produce this experience, a robot must draw on its accumulated knowledge of the individual user, adapting its tone and choice of topics to their particular history, preferences and current emotional state [11]. Even if a robot can respond in a natural way, this does not mean that it knows the person it is talking to. Over repeated sessions, this absence of personalisation progressively undermines the sense that the relationship is developing.
4. **Behavioural dynamics:** What is said is very important, but also how or when it is said matters. The mechanics of spoken interaction, including response latency, backchannelling and conversational pacing, are as important as the content of the words themselves [4, 7]. For older adults in particular, these dynamics require careful calibration. This group may speak more slowly, pause more frequently or require more time to process a response [8]. A system that interrupts too early, responds too quickly or fails to signal active listening will be perceived as socially intrusive, regardless of the appropriateness of its verbal output.

When these four layers are functioning together, the result is what the literature calls social comfort, i.e. a state in which the user experiences the interaction as natural and safe [68]. Social comfort is not a fixed threshold, but rather a dynamic equilibrium. Failure in any one of the four layers can destabilise this equilibrium. A robot that is empathetic but incoherent feels unpredictable. One that is coherent but impersonal feels indifferent. One that says the right things at the wrong pace feels intrusive. Social comfort is only achieved when all four conditions are met simultaneously, and it is this requirement that defines the engineering challenge in human-robot interactions (HRI).

2.1.2 The Fragmentation Gap: from Task Execution to Open-Domain Dialogue

Historically, robotic architectures were unable to fulfil these four psychological conditions. Early conversational robotics focused on task execution. Systems were designed to receive commands, retrieve information and perform specific actions [29], with social engagement being neither a design goal nor a success criterion. However, as robots began entering environments where sustained human interaction was important, a different need emerged. A robot that could only respond to commands seemed cold and impersonal in settings that required warmth and continuity. This drove the development of systems capable of open-ended, low-stakes conversation intended to build relationships and maintain engagement over time [18].

LLMs appeared to resolve what scripted dialogue or fixed response systems could not. Trained on large amounts of human language data, these models can generate contextually relevant, grammatically fluent responses on any topic without the need for task-specific programming [37, 82]. Yet the advance was incomplete. Although individual responses are fluent, extended interactions with LLM-powered robots still feel mechanical and reset-like [3, 56]. While each response may be locally appropriate, the conversation as a whole lacks continuity, initiative and relational depth. While the model produces high-quality output at the level of the single turn, it remains compartmentalised across sessions [18]. It does not carry forward information from previous sessions. It does not recognise when a topic important to the user has been overlooked and it does not initiate.

In addition to these structural shortcomings, LLM-powered systems tend to ask several questions in a single response. In emotionally charged conversations, this can create an excessive cognitive burden, as discussing sensitive topics requires some mental and emotional processing. Research evaluating empathetic social robots designed to discuss personal memories with older adults found that asking several questions simultaneously disrupts the natural rhythm of the conversation and overwhelms the listener [61].

These examples illustrate the fragmentation gap: the robot is fluent enough to pass as a conversational partner in the moment, but lacks the long-term continuity to feel like one over time [12, 37, 68]. The following sections examine the specific architectural components required to close this gap.

2.2 Approaches to Relational Coherence

Bridging the fragmentation gap requires more than fluent language generation. A robot that feels like a companion must maintain a coherent memory of who it is talking to, take initiative at the right moments and present a stable and recognisable personality across weeks of interaction. This section examines each of these three requirements.

2.2.1 Episodic and Semantic Memory Architectures

In friendships, people remember what you told them last week, notice when something has changed and bring up details you shared without being asked. This accumulated shared history is what transforms a series of pleasant exchanges into an actual relationship [25]. To produce the same effect, a social robot needs two distinct types of memory working together: episodic memory, which stores specific events and conversations, and semantic memory, which retains stable, long-term facts about the user [43].

Early attempts to implement this in social robotics relied on structured, rule-based approaches. The CARESSES project, for example, used ontology-based memory to store user-specific preferences, such as whether they preferred tea or coffee and used this information to personalise future interactions [29]. The Oro server took a more architectural approach, dividing memory into short-term, episodic and long-term categories with different lifespans and retrieval rules [43]. These approaches represented meaningful progress. They could remember that a user had mentioned their daughter, for example, but they could not recognise that the user has been anxious about a family visit in the past three days [25, 43]. Temporal reasoning, the ability to detect patterns that unfold across days or weeks, remains out of reach.

More recent frameworks address this limitation by moving away from static fact retrieval towards continuously updated knowledge structures. Rather than operating as a list of stored facts, these systems build an interconnected and evolving map of everything the system knows about the user, tracking their situation, concerns and conversational history as a structure that evolves with each session. MemU is one such example [53]. This underlying intention tracking enables proactivity, as the system always has a comprehensive and up-to-date understanding of who it is talking to and what matters to them.

2.2.2 Contextual Proactivity and Anticipatory Frameworks

A robot that only responds when spoken to is still a passive tool. It is proactivity that distinguishes a reactive system from one that feels present. In the literature, proactivity is defined as the ability to anticipate a situation and act on it without waiting for an explicit prompt [12]. For a socially withdrawn older adult who is unsure how to begin a conversation or simply does not feel like making the effort on a particular day, a robot that waits silently for initiation places the entire relational burden on the person. Specifically in elderly care, proactive behaviour takes the form of low-pressure, naturally timed interventions, such as suggesting a light activity, asking whether a family member has visited recently or simply opening the conversation with a reference to something the user mentioned last time [14, 46, 54]. These small gestures signal awareness and continuity, giving the feeling that the robot has been paying attention.

However, proactivity is also one of the easiest things to get wrong. An intervention that arrives at the wrong moment feels intrusive rather than warm [12, 14, 46]. This is why the literature emphasises the need for robotiquette, which is a set of social timing rules that govern when and how a robot can take the initiative appropriately. Equally important is avoiding repetition. A proactive callback that remains unchanged across multiple sessions quickly begins to feel scripted and a generic prompt that could have been directed at anyone feels impersonal [14, 37].

2.2.3 Persona Modeling and Identity Evolution

Although memory and proactivity solve the problems of continuity and initiative, they do not address the third requirement, that the robot must appear to be the same entity throughout every session. Users build trust with a robot's specific personality, tone and way of engaging. A robot whose character shifts unpredictably between interactions does not feel like an evolving companion. It feels like a different system each time. Research consistently shows that a clearly defined personality is one of the strongest predictors of user trust in social robots [51, 73].

The most widely used framework for defining this personality is the Big Five OCEAN model, which characterises personality along five dimensions: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism [1]. By anchoring a robot's personality to a specific OCEAN profile, designers can

produce a consistent and predictable relational identity.

Early implementations of this relied on static, preset prompts. At the start of each session, systems were given a fixed persona description, such as that of a 32-year-old counsellor who is warm, patient and supportive, and were instructed to maintain this persona throughout [16, 51, 73, 83]. While this approach works reasonably well in short interactions, it breaks down over time. Furthermore, static persona descriptions also influence how users relate to the robot. For example, framing the robot as a counsellor or care worker can affect user perception. Research shows that older adults respond better to peer-like social relationships than to care-oriented ones, and that overly therapeutic behaviour is frequently perceived as condescending [20, 64].

More recent work has moved towards dynamic persona frameworks, such as Self-Evolving Personalised Dialogue Agents (SPDA), which allow the robot’s identity to adapt gradually in response to the user’s evolving expectations and preferences [16]. This approach addresses the cold start problem of attempting to define an optimal identity before any interaction occurs, as it can now develop through the relationship itself.

However, there are still two problems. The first one is persona drift. This refers to a model’s tone, register or apparent personality gradually and unsupervisedly shifting across successive turns [83]. Unlike the cold start problem, which occurs at the beginning of a deployment, persona drift accumulates silently over time. The robot does not suddenly become a different entity. Instead, it changes in small, subtle ways that the user may not consciously notice, yet will still perceive. Over weeks of interaction, these shifts quietly erode the sense of a stable social presence [25]. The user can no longer predict how the robot will respond and this unpredictability undermines the psychological safety that affective trust requires.

The second problem is that frameworks such as SPDA manage gradual, long-term persona evolution reasonably well, but they don’t operate at turn level. They are not designed to detect real-time discrepancies between the robot’s output and the user’s immediate emotional state. A robot that reacts in a positive way to a user who has just expressed sadness (turn-level), or that stays cheerful when the conversation has turned to grief or loss (multiple turns), is an example of a different kind of persona failure. This is not about the robot drifting across time, but about it not being emotionally in line with the user at that moment.

This distinction is important because both types of failure damage trust, but through different mechanisms and at different timescales. Preventing drift requires tracking consistency across sessions. Preventing emotional misalignment, on the other hand, requires evaluating each response against the user’s current emotional state. Addressing both forms of failure cannot be solved at the level of prompt design alone. It requires a real-time supervisory architecture that monitors the robot’s output across turns and intervenes when the persona begins to deviate, or when a response risks being emotionally inappropriate.

2.3 Supervisory Frameworks and Model Self Evaluation

The previously discussed requirements cannot be met because the current LLM evaluation frameworks do not operate in real time and because language models cannot reliably evaluate their own outputs. When prompted to self-critique, they do not identify and correct errors. Instead, they rephrase the original response with equal confidence [35]. They cannot recognise when a response is contextually inappropriate, temporally inconsistent or emotionally misaligned because they have no mechanism for stepping outside the generation process to evaluate it [63].

Constitutional AI [5] demonstrates the importance of separating evaluation and generation. The system generating the dialogue should not be the same system evaluating it. By separating these roles, a secondary model acts as an independent supervisor, checking the proposed response against a strict set of core principles to catch errors. This two-stage structure reliably identifies errors that the generating model would not detect itself. However, Constitutional AI was designed for offline training where the evaluation criteria are explicit and well-defined, primarily for detecting harmful or toxic content. The rules regarding social naturalness, emotional appropriateness and relational continuity in open-domain social dialogue are highly ambiguous and context-dependent. They cannot be reduced to a fixed checklist, meaning the Constitutional AI framework cannot be transferred directly to live conversational settings without adaptation [5].

Another limitation in the current literature is its temporal nature. The majority of LLM evaluation frameworks operate post hoc, meaning that once the conversation has finished, the interaction is analysed offline to identify where things went wrong [45]. While this is useful for improving future systems, it offers no solution to errors that have already damaged the user experience. In social interactions, missed emotional cues or inconsistent persona responses cannot be corrected afterwards. By the time the analysis

is complete, the conversation will have moved on and relational damage will have been done. What is needed is an evaluation layer operating within the latency budget of a live voice conversation to detect and correct errors before the next response is generated.

2.4 Contribution

As demonstrated by the literature reviewed in this chapter, researchers have developed partial solutions to the discussed challenges. More specifically, individual aspects of memory infrastructure, dynamic persona frameworks and proactivity mechanisms have been addressed, though none of these issues have been fully resolved.

The literature does not yet provide a system that brings these components together and keeps them working in coordination during a live interaction. Memory, proactivity and persona consistency have been treated as three separate research problems, each with its own literature, benchmarks and proposed solutions. No existing framework treats these three factors as interconnected constraints that a supervisory layer must balance in real time.

This is the gap that this thesis addresses. The proposed solution is a secondary reflective layer referred in this thesis as the Dialogue Analyser that sits above the primary language model and acts as a social supervisor during live conversations. Rather than replacing the generative capacity of the main model, the Dialogue Analyser monitors the interaction as it unfolds, detects social errors that the primary model cannot identify itself and provides corrective guidelines before the next response is produced. The specific architecture of this module and its mechanisms for managing memory, proactivity and persona consistency in real time are detailed in the following chapters.

Architecture

3.1 Bridging the Gap: The Dialogue Analyser

This chapter explains how the Dialogue Analyser is constructed. The focus shifts from explaining why the Analyser is needed to showing how it works in practice.

The Analyser operates alongside the primary language model as an independent supervisory layer. The primary model retains full responsibility for generating responses, as the Dialogue Analyser does not generate them. Instead, it monitors the conversation as it unfolds. When it detects a social error or a missed opportunity, it sends a short corrective guideline to the primary model before the next response is generated.

These guidelines are not scripts. The module does not dictate what the robot should say word for word but it provides directional instructions, such as *‘acknowledge the user’s sadness about their deceased dog’* or *‘take the initiative to bring up the hiking trip they mentioned yesterday’*. The primary model then decides how to incorporate the instruction naturally in its own words and style, ensuring that the correction never feels like an interruption. When the conversation is going well, the Analyser stays silent. The robot’s persona, its knowledge of the user and any guidance produced by the Analyser are brought together into a structured prompt that is injected before each robot response is generated, as described in Section 3.3.5. The Dialogue Analyser operates across three functional phases, each targeting a different point in the conversational turn. The specifics of each phase are detailed in Section 3.3.

3.2 System Overview

The system is built on the TEN Framework, which is an open-source, real-time voice pipeline designed for agentic applications requiring the simultaneous processing of audio, text and contextual data with sub-second latency. TEN operates as a directed graph of isolated, asynchronous processing nodes, each running independently [75]. Tasks such as speech transcription, memory retrieval and social error analysis run in parallel, without any task blocking another. This parallel architecture makes it feasible to run a reflective supervisory layer alongside a live conversation without introducing noticeable delay.

Figure 3.1 provides a visual overview of the full pipeline. When the user speaks, their voice is captured and distributed simultaneously to two components: a speech-to-text module (Deepgram), which transcribes the words and a voice activity detection module, which tracks when speech starts and stops. Both processes draw on the same audio input and run in parallel, so neither delays the other.

Once a transcription is available, the system must determine whether the user has actually finished their turn. Using a fixed silence timer alone is unreliable, as it may cut in during a natural pause in the middle of a sentence or wait too long after a short reply [69]. A more effective approach is to use a lightweight model that reads each transcript and judges whether a thought has been completed. A sentence that ends with *‘and’* or *‘because’*, for example, does not necessarily mark the end of a turn, regardless of how long the silence that follows it lasts. This semantic judgement is handled by a dedicated turn detection model, which provides a verdict on whether the user’s thought is complete before the pipeline proceeds [60].

In this deployment context, a particular challenge is that elderly users speak more slowly and pause more frequently during their turns than younger adults do. If the silence thresholds are set too low, the system misinterprets natural pauses as the end of a turn and cuts in before the user has finished speaking. To address this issue, the voice activity detection silence threshold is set higher for elderly users. However,

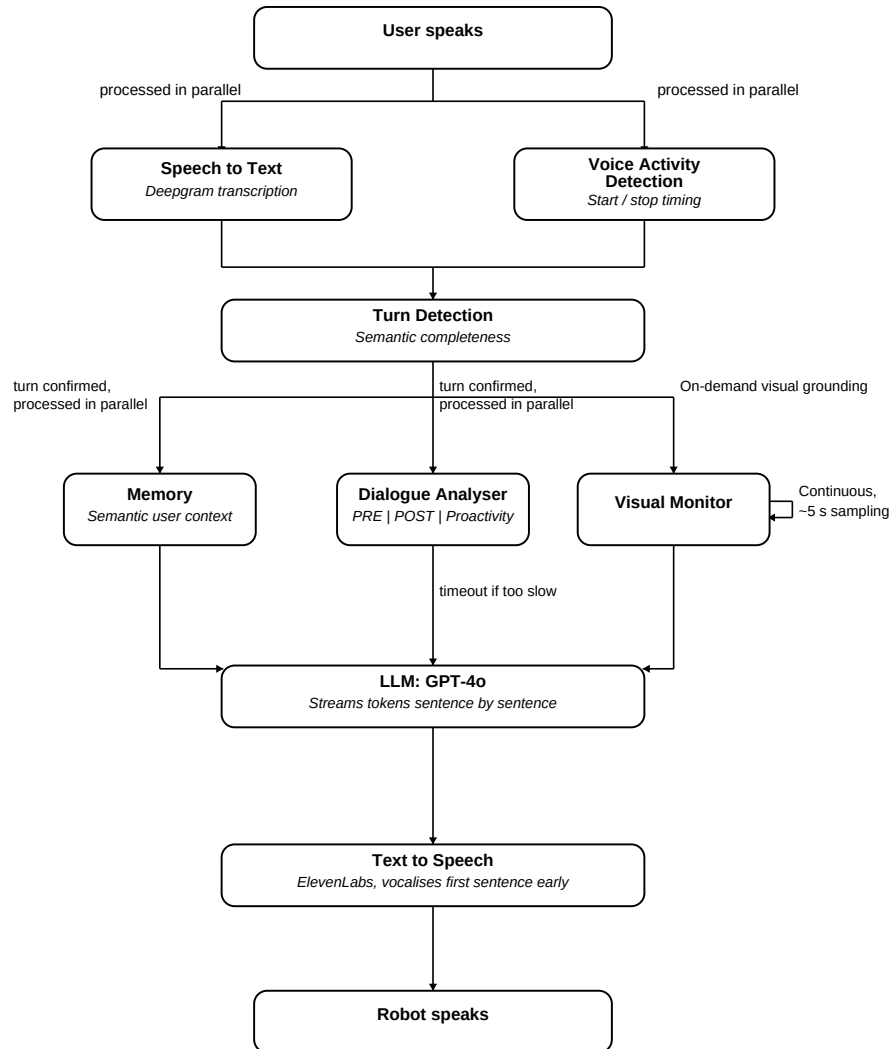


Figure 3.1: Overview of the conversational pipeline built on the TEN Framework. User input is processed in parallel for transcription and voice activity detection prior to turn confirmation. Asynchronous nodes that follow, including Memory, the Dialogue Analyser and the Visual Monitor, feed into the main LLM to ensure a real-time, low-latency response.

this is currently a static configuration rather than a dynamic adaptation, meaning it cannot adjust to variation between individuals within the elderly population. A user who speaks particularly slowly will still encounter premature cut-ins that a longer threshold would prevent. This is a known structural limitation of relying on static silence thresholds for turn-taking [70].

When the turn is confirmed, several things happen simultaneously. The system queries MemU for semantically relevant memories about the user. To capture the physical environment, an ambient monitoring process runs independently on a separate timer, sampling a visual frame approximately every five seconds and scoring its relevance to the ongoing conversation. When an environmental cue exceeds a relevance threshold, it is stored and incorporated into the context of the next turn. Additionally, a reactive visual routing check runs when the user’s words suggest a visual question, such as ‘look at this’, triggering an immediate analysis of the current camera frame.

The user’s input is also sent to the Dialogue Analyser. The primary language model waits until either a timeout is reached or the analyser returns a corrective guideline. If the guideline arrives in time, it is merged into the prompt before the model generates its response. Otherwise, the primary model proceeds without it and the guideline is applied to the next turn.

The primary language model (GPT-4o) generates its response incrementally, streaming tokens as they are produced rather than waiting for the full reply to be completed. As soon as the first complete sentence is formed, it is immediately sent to the text-to-speech module (ElevenLabs), which begins vocalising before the rest of the response has been written. This keeps the total delay from when the user finishes speaking to when the robot’s first audible reply is heard within a range that human listeners experience as natural.

3.3 The Dialogue Analyser

Although the primary language model generates fluent and contextually relevant responses, it has three structural blind spots that make it unsuitable for long-term companion interaction on its own. It does not critique its own previous responses, so if it produces an inadequate reply, it has no mechanism to adjust without external feedback [35]. Secondly, it does not reliably track patterns across turns. For example, it may continue to ask follow-up questions without registering that the user’s replies are becoming shorter, which is a well-documented indicator of disengagement. Furthermore, its default tendencies, such as ending each turn with a question and maintaining forward momentum in the conversation, actively conflict with good companion behaviour for elderly users. It has been shown that an unbroken stream of questions disrupts conversational flow and causes disengagement in this group [80]. The Dialogue Analyser aims to address these errors by monitoring the interaction at every turn and providing corrections without interrupting the main conversation.

3.3.1 Architecture and Delivery

The Dialogue Analyser operates as an independent node within the TEN Framework, receiving the current user turn and the most recent robot response as input. It also uses its own rolling conversation history to interpret the current user turn and robot response in relation to the overall conversation. The Dialogue Analyser triggers three parallel analyses:

- **Pre analysis:** which occurs before the robot responds, examining what the user has just said in the context of the recent conversation history.
- **Post analysis:** which evaluates the previous robot response against both the user’s current reply and the conversational context.
- **Proactivity module:** which determines whether the robot should initiate something.

These results are packaged as a structured guidance block and injected into the primary model’s prompt. To meet the latency constraints of live conversations, the system uses a three-payload delivery approach. The pre-analysis payload is sent as soon as the pre-analysis call completes, allowing the primary model to act on it immediately. Post-analysis follows as the second payload once its analysis is complete. The proactivity decision arrives as the third payload and is merged into the current turn if it arrives within the timeout window. This design ensures social correction is delivered in the same turn wherever possible while guaranteeing the pipeline never stalls waiting for guidance.

The Dialogue Analyser’s internal calls use GPT-4o-mini rather than the full GPT-4o model. Since all three phases perform classification and detection rather than open-ended generation, the reduced model

capacity is sufficient and keeps the pipeline within the required timeout window [28].

Figure 3.2 provides an overview of the three parallel analyses, their respective checks, and how their outputs are merged into a single guidance block before each response is generated.

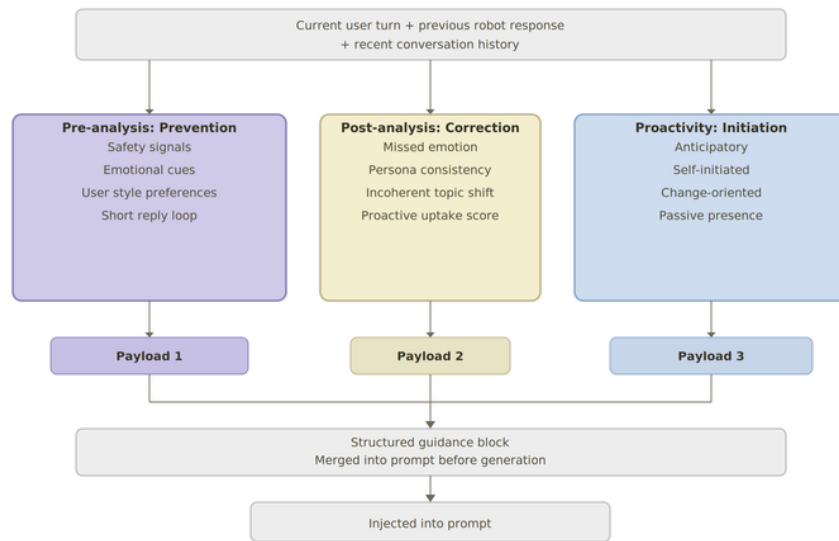


Figure 3.2: The internal structure of the Dialogue Analyser. Each phase produces a payload that is merged into a structured guidance block. This block is then injected into the GPT-4o prompt before the next response is generated.

3.3.2 Pre-Analysis: Prevention Before the Response

Pre-analysis examines the user’s input before the robot responds. Its purpose is to prevent the robot from producing a harmful or socially inappropriate reply by applying a behavioural constraint before generation begins. It applies four checks:

- **Safety signals:** Crisis language and self-harm cues are flagged. In an elderly care context, the robot must never probe further when a user expresses distress. The correct response in this situation is to offer a warm acknowledgement and, if appropriate, gently suggest that they speak with a family member or caregiver [4]. Social robots lack the clinical empathy and judgement required for managing severe psychological crises. An automated system that attempts to intervene in cases of acute distress risks creating a therapeutic misconception whereby the user overestimates the machine’s capacity to provide therapeutic support [39]. Relying on a robot for complex emotional support can worsen a user’s mental state rather than providing comfort [4].
- **Emotional cue detection:** The system flags signs of personal distress or emotional load, ensuring the primary model explicitly acknowledges and validates the user’s feelings rather than moving past them.
- **User style preferences:** Explicit preferences expressed during the session, such as ‘*be less chirpy*’, are captured and stored for use in future prompts of that conversation.
- **Engagement check:** This check operates deterministically without calling a language model. If the robot has asked questions in three or more consecutive turns and at least two of the user’s replies contain fewer than ten words, a correction is forced because short replies are a reliable behavioural indicator of disengagement [58] and an unbroken sequence of questions met with increasingly brief answers is a recognised failure pattern in companion interaction. In such cases, the robot is instructed to respond with a warm statement, reflection or observation, rather than asking another question. This constraint only applies to the current turn, after which normal behaviour resumes.

3.3.3 Post-Analysis: Correction After Robot Response

The pre-analysis focuses on what the user said. Post-analysis, on the other hand, focuses on what the robot said, specifically whether its previous response was socially appropriate in light of what the user said next. While pre-analysis can detect when a user expresses distress, it cannot assess whether the robot’s response to that distress was adequate. Post-analysis considers the entire conversation and asks whether it was handled well.

To illustrate this, suppose a user says, *‘My sister called yesterday, we hadn’t spoken in years.’* The robot might reply, *‘That’s wonderful! What else did you do yesterday?’* If the next user’s response is a normal reply to the question asked, pre-analysis sees nothing unusual and has no reason to flag anything. The user has answered, the conversation is moving and everything looks fine. However, post-analysis sees the full picture, a user who has shared something emotionally significant, a robot that has acknowledged it with a single word and immediately moved on to a generic follow-up question and a user who has followed along without the moment ever being properly addressed. Post-analysis would flag this as a missed emotional moment and instruct the robot to return to what the user said about their sister in their next response, giving the robot a second chance to handle the moment it failed to acknowledge.

A limitation of this retroactive approach is that the robot has already made a conversational misstep and addressing this a turn later might feel unnatural. However, this limitation is mitigated by the fact that the post-analysis output is provided as a guideline and the main generative model (e.g. GPT-4o) uses its own judgement to decide how to incorporate this information naturally. For example, if the user responded to the generic question by saying that they had gone for a walk in the garden, the robot could connect the two topics by asking, *‘A walk in the garden sounds lovely. Did you find yourself thinking a lot about that phone call with your sister while you were walking?’* This ensures that the system validates both the user’s latest input and the previously overlooked emotion without disrupting the conversational flow.

Post-analysis monitors for three issue types.

- A missed emotion flag is raised when the user expresses personal emotion explicitly and the robot fails to acknowledge it.
- A personality consistency flag is raised when the robot violates a protected behavioural value, such as dismissing real distress, repeating itself on a topic that the user has already moved on from, or maintaining a forced cheerfulness that conflicts with the user’s emotional state.
- An incoherent topic shift flag is raised when the robot changes the subject without providing a connection to the previous topic or returns to a topic that the user has indicated they want to leave behind.

The intervention threshold is intentionally high so that post-analysis only fires a corrective guideline when it is confident that the delivered response would cause concrete social failure. False positives that overwrite a good response are more damaging to the interaction than false negatives.

The post-analysis also scores how the user responded to the previous proactive move, determining whether they engaged with it, ignored it or displayed disengagement. This score is fed directly into the proactivity cooldown system. A single low or medium uptake score means that the same topic cannot be used proactively again for the rest of the session. Only a high uptake score allows the same topic to be reintroduced after a fifteen turn cooldown period, a value defined as an adjustable system hyperparameter.

3.3.4 Proactivity: Initiating rather than Reacting

The proactivity module determines whether the robot should take the initiative by raising an unacknowledged detail, introducing a new topic or providing quieter companionship. It runs as a separate process, triggered at a heuristically chosen baseline of six turns, or sooner if pre-analysis detects a drop in engagement or if the user responds with short replies on consecutive turns. The module operates in four modes:

- **Anticipatory:** The system acts before a need arises. It predicts where the conversation is heading and surfaces something relevant (often drawn from memory) before the user explicitly raises it. For example, if a resident remarks, *‘I can’t believe it’s already the end of October’* and the system’s memory notes that it is their deceased spouse’s birthday this week, the robot can recognise the emotional weight and say something like : *‘Time is passing quickly. I know this week is usually tough for you with John’s birthday approaching. How are you feeling about it?’* This validates the emotion without the user having to explain why they are feeling down.

- **Self-initiated:** The robot brings up something that the user has mentioned in passing earlier in the session, or recalls a specific detail from a previous day, which has not yet been explored.
- **Change-oriented:** When the current conversational trajectory is no longer serving the user, the system introduces a fresh angle or topic aligned with the user’s known historical interests to initiate a shift.
- **Passive Presence:** This is the most important for elderly care. When the user’s social energy is consistently low, the system backs off entirely, offering one calm, validating statement with no question or new topic attached.

A cooldown system prevents the same topic from being raised proactively more than once per session unless the user engages with it the first time. Topics are assigned a stable label when introduced and their cooldown period is updated based on uptake scores. This prevents the proactivity module from repeatedly cycling through the same material, which the literature identifies as a primary cause of disengagement once the initial novelty has worn off [80].

3.3.5 Persona and Prompt Architecture

Before generating a response, all relevant information (the robot’s persona, its knowledge of the user, the conversation history and the guidance signal from the Dialogue Analyser) is combined into a single, structured prompt. This prompt can be divided into three layers, each of which contributes to the robot’s next response.

- **The First Layer (Fixed Identity):** The first layer is a fixed system prompt that is assembled at the start of each session. Its core is a single identity statement: ‘You are Noa, a lifelong friend of [user name], not a therapist or care worker.’ The system prompt implements this through concrete behavioural rules: responding to what the user has actually said rather than offering generic comfort, never inferring or naming the user’s emotional state unless they have named it themselves, asking specific follow-up questions rather than open-ended therapeutic prompts and never filling conversational space with empty affirmations.
- **The Second Layer (Background Memory):** At this stage, Noa knows how to behave, but has no idea who they are talking to. The second layer addresses this. At the start of the session, a background task queries MemU for a summary of the user’s profile, interests, current plans and recent life events. If anything is returned, it is added to the per-turn context.
- **The Third Layer (Per-Turn Context):** This layer is rebuilt with every turn. Before each call to the primary language model, a structured context message containing relevant memory (both the general background and specific details retrieved for that turn), the conversation history, and the guidance signal from the Dialogue Analyser is assembled.

The first layer tells the robot how to behave. The second tells it who it is talking to. The third builds the actual per-turn prompt. Without all three working together, the robot might know how to behave, but would lack the context to do so.

Experimental design

The previous chapters established the need for socially coherent robots and the architectural response proposed by this thesis. This chapter describes how this approach is evaluated. The study is guided by two primary research questions:

- **Research Question 1:** Does the Dialogue Analyser enhance the user’s conversational experience compared to a baseline system without it?
- **Research Question 2:** Does the effect of the Analyser depend on the emotional load of the conversation? In other words, is its benefit greater when participants are discussing something stressful rather than something positive?

To address these questions, a controlled within-subjects experiment was conducted, involving the social robot Pepper (see Figure 4.1). Pepper is a widely used, semi-humanoid robot designed specifically for HRI [57]. Each participant engaged in two eight-minute conversations with the robot, where in one conversation the Dialogue Analyser was active and one where it was not.

4.1 Experimental Conditions

The study uses a 2 x 2 within-subjects design that crosses two independent variables: system condition (baseline versus Dialogue Analyser active) and conversation topic (positive versus stressful). Each participant experiences both conditions, one per session. The specific pairing and order are determined by random assignment to one of four counterbalancing groups.

4.1.1 System Condition

In the baseline condition, the social robot Noa, which is embodied by the humanoid robot Pepper, is driven only by a GPT-4o language model and a standard companion role system prompt. This model generates responses one at a time without any retrospective evaluation. It does not critique its own preceding output or track behavioural patterns across turns. Nor does it have a mechanism for detecting tone mismatches between its response and the participant’s emotional state.

In the active Dialogue Analyser condition, the same underlying GPT-4o model and system prompt are used. However, a supervisory layer runs in parallel with every conversational turn, performing three concurrent analyses as described in Chapter 3. All other aspects of the experiment are identical for each participant, including the robot’s voice and physical appearance, the room setup, the experimenter’s behaviour and the length of the conversation. Participants are not informed which condition is active during which session. This information is only revealed during the debriefing session once both sessions have been completed.

4.1.2 Conversation Topic

The second within-subjects factor is the emotional content of the conversation. At the start of each session, the participant is given a prompt card with a conversation topic on it. In the positive topic condition, they are asked to discuss a recent positive experience of their choice, such as a pleasant moment with friends, a personal achievement or an upcoming event. In the stressful topic condition, they are asked to discuss something that is currently worrying or stressful, such as an approaching deadline, a

difficult conversation or an ongoing concern. Participants choose the specific content within each category themselves, ensuring that the emotional material is personally relevant rather than externally imposed. The topic factor tests whether the benefit of the Dialogue Analyser depends on the emotional load of the conversation. It is hypothesised that the module adds most value when empathic attunement matters most, i.e. in the stressful condition. If participants' scores improve when discussing the stressful topic but not the positive one, this suggests that the benefit is driven specifically by emotional sensitivity rather than by a general improvement in conversational fluency. Without the positive topic as a comparison point, this conclusion cannot be drawn.

The idea of a design in which participants would experience only one topic was considered, but rejected because it would not allow the second research question to be tested. In order to determine whether the module's effect depends on emotional load, it is necessary for each participant to experience both a stressful and a positive conversation. A between-subjects comparison of topics would confound the effect of the topic with individual differences in expressiveness, stress sensitivity and baseline comfort with the robot. Therefore, the within-subjects design is the only approach that cleanly separates the module's general effect from its emotionally specific one.

4.1.3 Counterbalancing

As shown in Table 4.1, four counterbalancing groups handle the crossing of condition order (baseline first versus Dialogue Analyser active first) and topic order (positive first versus stressful first). The 16 participants were randomly and equally distributed across four groups ($n = 4$ per group). This counterbalancing controlled for carry-over effects and session order effects. However, residual carry-over effects such as mood, fatigue and adaptation to the robot's voice cannot be fully eliminated given that both sessions occur on the same day.

Group	Session 1	Session 2	n
A	Baseline + Stressful Topic	Analyser + Positive Topic	$n/4$
B	Baseline + Positive Topic	Analyser + Stressful Topic	$n/4$
C	Analyser + Stressful Topic	Baseline + Positive Topic	$n/4$
D	Analyser + Positive Topic	Baseline + Stressful Topic	$n/4$

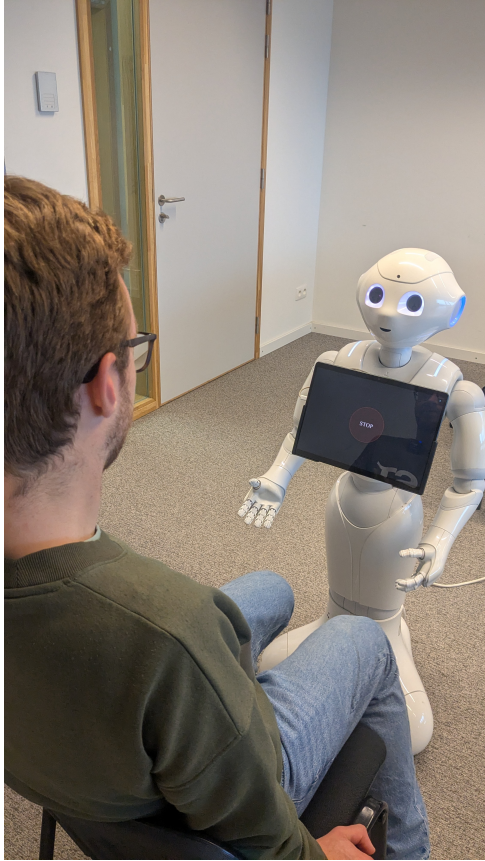
Table 4.1: Experimental Design Conditions

4.1.4 Participants

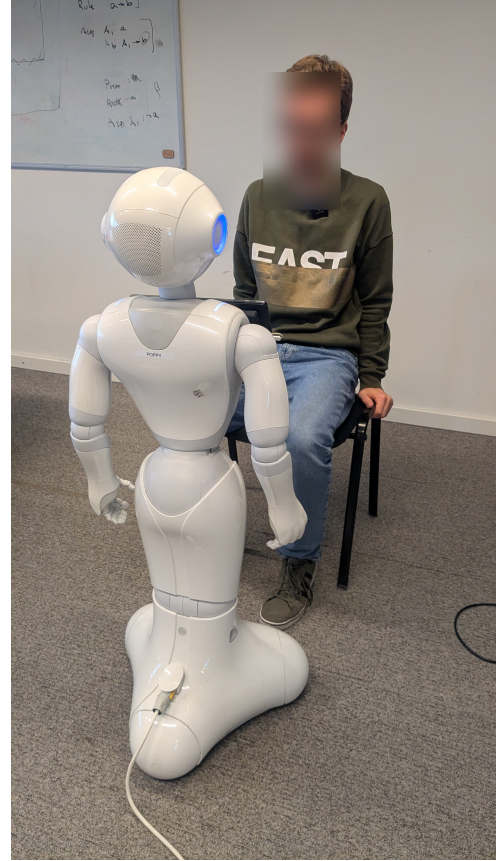
A total of 16 participants (7 male and 9 female, 22.2 ± 1.2 years) were recruited. The group was mainly made up of engineering students, as well as a student of psychology and one of social work. All participants were native Dutch (Flemish) speakers. At the start of the experiment, they were asked if they were comfortable conducting the sessions in English and all participants agreed. Twelve of them reported using AI chatbots on a weekly basis, indicating a high level of general familiarity with conversational technology. However, most of them had no prior experience with physical social robots. Their talkativeness was rated as $M = 6.6$ ($SD = 2.0$) on a scale of 1-10. Participants with severe hearing or speech impairments that would affect their ability to interact naturally with the robot were excluded, as were those with prior exposure to the Dialogue Analyser in an experimental context.

4.2 Settings

All sessions took place in a meeting room at the I Gent building on the campus of Ghent University. Participants sat facing the Pepper robot, which was positioned facing a white wall to optimise its face-tracking performance, as is shown in Figure 4.1b. To overcome the limitations of Pepper's built-in microphones, each participant wore an external lapel microphone to ensure accurate voice detection and clear speech transcription. Audio was played back through a tablet attached to the robot's chest, which also served as the primary interface, as can be seen in Figure 4.1a. The TEN Framework and all processing ran on a high-performance external workstation equipped with a dedicated GPU. By decoupling the computation from the robot in this way, response latency was reduced and it was ensured that the parallel supervisory analyses of the Dialogue Analyser did not compete with the main conversational pipeline for processing resources.



(a) The Pepper robot, featuring the tablet-based primary interaction interface.



(b) Participant orientation and robot positioning to optimise face-tracking performance.

Figure 4.1: Experimental setting

4.3 Procedure

Each session lasted approximately forty to forty-five minutes and the overview can be found in Table 4.2.

Table 4.2: Overview of the experimental procedure and approximate timings per participant.

Phase	Approximate Duration
Participant intake, consent and demographic form	8 minutes
Introduction to Noa	2 minutes
Session 1 (First conversation topic)	8 minutes
Post session 1 questionnaire	5 minutes
System changeover	Background task
Session 2 (Second conversation topic)	8 minutes
Post session 2 questionnaire	5 minutes
Post experiment questionnaire, audio recorded interview and debrief	5 minutes
Total estimated time	40 to 45 minutes

Participants began by completing an informed consent form and a short questionnaire covering their demographics (Appendix A.2), including their age, gender, field and year of study, baseline stress level, prior experience with social robots and AI chatbots, general comfort with technology and any hearing or speech-related issues. They then spent approximately two minutes in an unrecorded conversation to familiarise themselves with Noa, to reduce the novelty effect, before any data collection began. For this short introductory conversation, the opening was ‘*Hi {user name}! I’m Noa, nice to meet you. Tell me a bit about yourself or what you have been up to lately?*’ This was followed by the first recorded eight minute conversation, which began with a topic-specific greeting prompt. For the positive topic,

Noa opened with *‘Hi {user name}! Good to see you. I heard something good happened recently. I’d love to hear about it. What’s been a nice moment for you lately?’* For the negative topic, the opening was *‘Hi {user name}, good to see you. I heard you might have something on your mind today, feel free to share whatever it is, I’m listening.’* Once the conversation had concluded, the participant completed the post-session questionnaire while the experimenter silently switched the system to the alternate condition in the background. A second eight-minute conversation then took place under the alternate condition, after which the participant completed the questionnaire a second time. The experiment concluded with a comparative survey, a brief semi-structured audio-recorded interview and a full debrief, during which the experimenter explained the two-condition structure and described the Dialogue Analyser and its purpose. They also answered any remaining questions. The full questionnaire can be found in Appendix A.1.

4.4 Metrics

Before deciding what to measure, it is important to clarify what a better interaction means in this context. A conversation does not simply feel good because the robot produces fluent sentences. It feels good when the robot notices your emotional state, responds with an appropriate tone and takes the initiative to maintain the conversation rather than passively waiting for you to speak [59, 68]. The Dialogue Analyser is designed to address these issues and each measure in this study targets something that the module is designed to improve or prevent. All scale choices follow the guidelines established by Saad et al. [65] and the GMU HRI Scale Database, which rates scales against a thirteen-item psychometric quality checklist [27]. Only scales that have been rated above 77% on this checklist or have comparable independent validation are included. One scale that did not survive this process is Godspeed, which Saad et al. identify as requiring further investigation before use in HRI contexts, as its sub-scales do not clearly distinguish the constructs they claim to measure.

4.4.1 Perceived Empathy

Perceived empathy captures whether the participant felt that the robot recognised their emotional state and responded in a way that was appropriate to how they were feeling, rather than just saying something kind.

This measure is assessed using the Perceived Empathy of Technology Scale (PETS), which was developed by Schmidmaier et al. through a rigorous, multi-phase process [66]. This makes PETS the most thoroughly validated empathy scale for interactive systems currently available. The scale has two sub-scales. The Emotional Response sub-scale (PETS-ER) captures the emotional side of empathy across six items: Did the robot seem to notice and respond to the participant’s feelings? Did it show interest? Did it sympathise or support the participant through an emotional situation? The Understanding and Trust subscale captures the cognitive side of empathy across four items: did the robot understand the participant’s goals, needs and intentions and did the participant trust it? Items are rated on a five-point Likert scale, with each subscale scored as the mean of its items.

PETS was primarily validated with conversational agents, such as chatbots and voice assistants, rather than with embodied robots. While the constructs it measures do translate directly to embodied interaction, absolute baseline scores may differ and this is acknowledged as a limitation. It is expected that PETS scores will be higher in the active condition of the Dialogue Analyser, with the largest effect occurring during the stressful topic, where empathic attunement matters most. This prediction follows from the architecture of the module. The pre- and post-analysis components detect and respond to negative emotional signals, such as worry and sadness, rather than positive ones, such as excitement or joy. Therefore, the module is most likely to activate when the topic is stressful, which is why the largest empathic benefit is expected there.

4.4.2 Emotional Appropriateness

Emotional appropriateness is related to, but distinct from, empathy. For example, a robot could score highly for overall empathy yet still misjudge the tone of a specific moment, responding cheerfully when someone shares a worry or remaining serious when the conversation turns to lighter topics. The pre-analysis phase of the Dialogue Analyser is designed to identify this type of failure, making this construct useful for determining whether the module is effective.

As there is no validated standalone scale for emotional appropriateness in HRI, it is captured through triangulation across three sources. The PETS item *‘The robot supported me in coping with an emotional*

situation’ assesses whether the response felt appropriate to the situation emotionally. The custom item ‘*The robot seemed aware of changes in my mood or tone*’ directly measures tone tracking. The open-ended reflection question, which asks whether there was a moment when the robot’s response surprised the participant either positively or negatively, often reveals incidents of tone mismatch because participants naturally describe the exact moments when something felt inappropriate.

4.4.3 Conversational Naturalness

Naturalness captures the overall feel of the conversation in terms of pacing, timing and flow. It is measured using a single Mean Opinion Score (MOS)-style item, which is rated on a five-point scale ranging from ‘Very unnatural’ to ‘Very natural’, following the ITU-T Recommendation P.800 standard for subjective assessment of conversational quality [36]. Only one item is used because there is no validated multi-item naturalness scale for HRI in the GMU database. Constructing an unvalidated composite would create the impression of a richer measure without the necessary psychometric foundation. This item is supplemented by five custom conversational dynamics items, which capture aspects of conversational flow quality (Appendix A.1 Block D).

4.4.4 Interaction Enjoyment

Enjoyment integrates all of the participant’s experiences, both positive and negative, into a single overall judgement about whether talking to the robot was worthwhile. In the context of the eldercare deployment targeted by this thesis, it is also the most practically meaningful outcome as it predicts whether a person will engage with the robot again. A robot that scores highly for empathy but feels unnatural or forced may still fail here.

Enjoyment is captured through two complementary instruments. The primary measure is a single Likert item per session, adapted from the ‘Perceived Enjoyment’ (PENJ) construct of the Almere Model [32], which is a framework that has been specifically validated for evaluating social robot acceptance among older adults. The post-experiment questionnaire contains four forced-choice comparative judgement items, asking participants to select which session felt more natural, empathic, understanding and enjoyable. As participants cannot rate both sessions equally, this format eliminates the bias associated with using a scale and is expected to be more sensitive than the single Likert item (participants can’t give both sessions a 4 for example). Given the low reliability of single items, self-report enjoyment ratings are expected to show small or null differences between conditions, while the forced-choice measures are expected to be more discriminating.

4.4.5 Felt Comfort

‘Felt comfort’ captures how relaxed or tense the participant felt during the conversation. It is distinct from both perceived empathy, which evaluates the robot’s behaviour, and enjoyment, which is a retrospective, global judgement. Felt comfort focuses on the participant’s own emotional state during the interaction. This distinction is important because the Dialogue Analyser is designed to make the robot more emotionally attuned, which should, in turn, make participants feel more at ease, especially when discussing stressful topics where someone sharing a personal worry needs the conversation to feel safe.

Felt comfort is measured using a single MOS style item rated on a five-point scale from ‘Very uncomfortable’ to ‘Very comfortable’. A single item is used for the same reason as for naturalness as there is no validated, short, multi-item felt-comfort scale for HRI. To build a more complete picture of what this score captures, it is triangulated with the post-session stress item, the enjoyment item and the PETS Understanding & Trust sub-scale. It is expected that felt comfort will be higher in the active Dialogue Analyser condition, with the greatest effect on the stressful topic and smaller or null effects on the positive topic.

4.4.6 Robot Social Attributes (RoSAS warmth and discomfort) [13]

RoSAS captures the participant’s overall perception of the robot as a social entity, including how warm or cold and comfortable or unsettling they found it throughout the conversation. If the Dialogue Analyser makes the robot feel warmer overall, rather than just being more empathic at certain points, RoSAS Warmth will detect this. Similarly, if the module’s interventions make the robot seem unnatural or overly attentive in a way that makes the participant feel uncomfortable, RoSAS Discomfort will flag this. Only the Warmth and Discomfort subscales are used. The Competence sub-scale is excluded because the

Dialogue Analyser targets emotional behaviour rather than task performance and including it would add six items per session for limited theoretical gain.

The warmth sub-scale has only five questions because the 'organic' question has been omitted, as it is vague and would not provide any useful information to participants. This sub-scale is a positive outcome, scored as the mean of its items. Higher scores indicate a greater perception of warmth. Discomfort (six items) is a negative outcome, meaning that a higher score indicates greater discomfort. Given that RoSAS is a holistic measure and the module targets specific moments, modest effects on Warmth are expected. Discomfort should remain near the minimum in both conditions.

4.4.7 Conversational Dynamics

Five custom items target specific functions that the Dialogue Analyser is designed to perform, such as taking the initiative in guiding the conversation, responding at the right moment, maintaining the flow of the conversation and noticing shifts in mood or tone. As no published HRI instrument covers this combination of constructs, these items were written specifically for this study. Prior to analysis, the topic-pacing item is reverse scored, and items can be averaged into a single Dialogue Analyser responsiveness composite or examined individually, depending on the research question.

4.4.8 LLM-as-a-Judge

The LLM-as-a-Judge approach complements self-report measures. Using an LLM as an evaluation tool is a validated methodological approach in HRI research. It has been shown to produce ratings that correlate well with human judgement when applied to well-defined coding schemes [85]. An LLM, in this case Claude Opus 4.7, receives each conversation transcript without any information about the experimental condition. Empathic response quality is rated using the Empathic Communication Coding System (ECCS) as applied by Pinto et al. [61], which evaluates the appropriateness of empathic responses. The robot response is rated on a scale from 0 to 6 based on how effectively it acknowledges and addresses the emotional content of the preceding user turn, as can be seen in Table 4.3. In addition to the ECCS score, three structural metrics are extracted per turn: whether the robot grounded its response in specific content from the user's turn (named place, person, etc.), whether it acknowledged the user's emotion and whether it asked more than one question in a single turn. Combining self-report and automated scoring means that conclusions about the effect of the Dialogue Analyser rest on two independent sources of evidence rather than one.

Table 4.3: Coding levels for the Empathic Communication Coding System (ECCS), with illustrative robot examples.

Level	ECCS (0–6): Empathic appropriateness
0	Ignores, denies, or contradicts emotion. <i>User: "I'm sad." Robot: "That's not true."</i>
1	Minimal acknowledgment; flat response. <i>"Okay." / "I see."</i>
2	Recognition: names or echoes emotion without depth. <i>"You're happy about that."</i>
3	Reflection: mirrors emotion with warmth. <i>"That sounds really special for you."</i>
4	Pursuit: encourages elaboration. <i>"That seems important—tell me more about what it meant."</i>
5	Confirmation: deeper, accurate understanding. <i>"I can tell it gave you comfort just being together."</i>
6	Additive empathy: provides new meaning or reframe. <i>"It sounds bittersweet—you miss those days, but they also gave you strength you carry now."</i>

4.5 Data Analysis Plan

The analysis is structured to address the two research questions sequentially. Before any hypothesis tests are performed, a series of preparatory steps ensure that the data is ready for analysis.

4.5.1 Scale Reliability and Scoring

Before computing composite scores, the internal consistency of each multi-item scale is examined using Cronbach's α . This step is necessary because averaging items into a single score only makes sense if those

items measure the same underlying construct. If internal consistency falls below acceptable thresholds, items will be inspected individually and removed or reorganised where appropriate. If the items within a scale measure different concepts, those items will be grouped into separate indexes rather than being forced into a single composite score. Items showing severe floor and ceiling effects (zero variance) across all participants cannot contribute to any statistical analysis and will be excluded from hypothesis testing, being reported descriptively only.

4.5.2 Normality Assessment and Multiple Comparisons Correction

For paired comparisons (RQ1), normality is assessed using the Shapiro-Wilk test on the difference score for each participant (Analyser - Baseline). This is the appropriate approach for within-subject designs, since it is the distribution of the difference that matters, rather than that of each condition separately. For the interaction effect (RQ2), the normality of the residuals is assessed before running the mixed ANOVA. Eight outcomes (overview in Table 5.2) are tested simultaneously for each research question. Testing eight hypotheses at an α level of .05 creates a roughly 34% chance of at least one false positive (Type 1 error) occurring by chance alone. To address this issue, the Benjamini-Hochberg false discovery rate (FDR) procedure is applied to all eight raw p-values for both RQ1 and RQ2. This approach is preferable to the Bonferroni correction, as the latter would be too strict for correlated outcomes in a small sample.

4.5.3 Statistical Tests

For RQ1, if the difference score is approximately normally distributed, a paired t-test is used. Where normality is violated, the Wilcoxon signed-rank test is used instead. This adaptive strategy is applied to each outcome rather than selecting one test for all measures. For RQ2, a 2 x 2 mixed ANOVA is used, with condition (Analyser versus Baseline) as the within-subjects factor and topic (stressful versus positive) as the between-subjects factor. The central focus is on the interaction between condition and topic: does the benefit of the Analyser differ depending on the topic assigned to the participant? An overview of the test selection procedure for RQ1 is given in Figure 4.2.

Due to the small sample size of the main analysis cohort, all findings are treated as exploratory rather than confirmatory. Statistical power refers to the probability of detecting a true effect if it exists. A post-hoc power analysis using a non-central t-distribution revealed that with $N = 16$, the study achieves adequate power ($\geq 80\%$) only when the true effect is large ($d \geq 0.75$). For a medium effect ($d \approx 0.5$), a sample of approximately $N = 33$ would be required to reach the same level of power. In practical terms, this means the study is well equipped to detect strong, robust effects but may miss smaller yet meaningful ones. Non-significant results should therefore be interpreted as inconclusive rather than as evidence of no effect.

For the LLM-as-a-Judge analysis, the empathic quality of each robot response to emotionally marked turns is compared between the Analyser and Baseline conditions using Mann-Whitney U tests, based on the ECCS scale output. Fisher's exact tests are used to compare the multi-question rate and grounded reference rate between conditions, as these are appropriate for binary outcomes with small cell counts. FDR correction using the Benjamini-Hochberg procedure is applied to all tests.

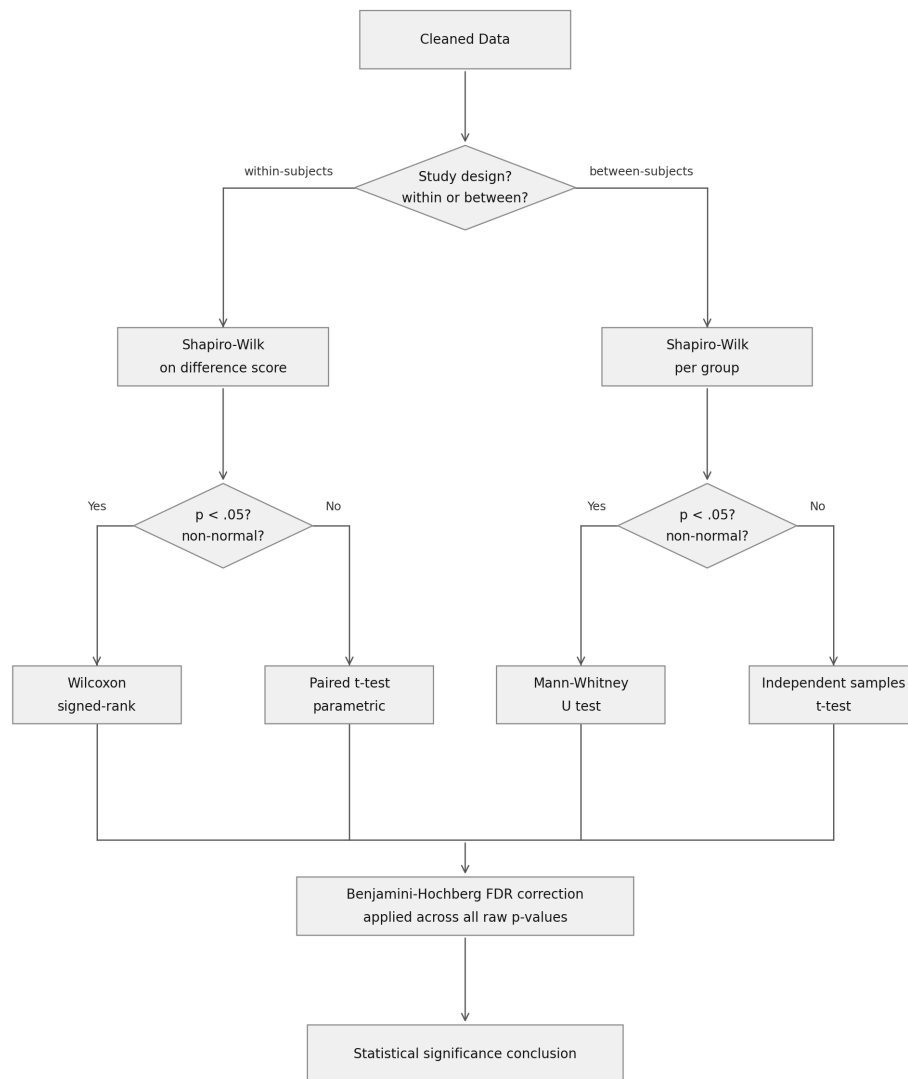


Figure 4.2: Test selection procedure for RQ1 based on study design and normality. For RQ2, a 2x2 mixed ANOVA was used to test the condition x topic interaction.

The chapter is organised as follows: Section 5.1 covers preliminary analyses, including scale reliability diagnostics and a manipulation check. Sections 5.2 and 5.3 address the two primary research questions. Section 5.4 presents secondary analyses, including forced-choice preference data, within-person consistency profiles, session order effects and demographic covariates. Section 5.5 reports conversational dynamics and latency.

5.1 Data Preprocessing and Diagnostics

5.1.1 Scale Reliability Diagnostics

Before computing composite scores and conducting any statistical tests, the internal consistency of each scale was examined using Cronbach’s α . According to the established benchmarks of George and Mallery [26], $\alpha \geq .80$ was considered good, $\alpha \geq .70$ acceptable, $\alpha \geq .60$ questionable but usable and $\alpha < .60$ poor, indicating that a scale may not measure a single coherent construct. The results, along with the decision taken for each scale, are presented in Table 5.1.

Table 5.1: Reliability Analysis for Study Scales

Scale	Items	α	Verdict	Decision
RoSAS Warmth (5-item)	5	.585	Questionable	4-item composite, ‘Seemed happy’ removed
RoSAS Discomfort	6	.544	Poor	Excluded from inferential analysis
PETS Emotional Response	6	.596	Questionable	Retained as composite
PETS Understanding and Trust	4	.693	Acceptable	Retained as composite
Custom Dialogue (5-item)	5	.386	Poor	Split into Proactivity and Timing indexes

Two scales required structural modification prior to analysis. For the RoSAS Warmth scale, the item ‘*Seemed happy*’ showed near-zero inter-item correlations with all other items ($r \approx .00$ to $.15$), suggesting that robots in laboratory settings tend to be rated as uniformly happy, regardless of experimental condition. Removing this item increased the Cronbach’s alpha coefficient from $.585$ to $.629$, so all subsequent analyses use a 4-item Warmth composite. For the RoSAS Discomfort scale, the item ‘*Seemed dangerous*’ exhibited zero variance across all observations ($M = 1.00$, $SD = 0.00$), indicating a floor effect. Items with zero variance contribute no statistical information, as no correlation can be computed and no test retains power. The Discomfort composite was therefore excluded from all inferential analyses. The floor effect is itself a finding, as no participant perceived the robot as threatening or dangerous in either condition. The PETS Emotional Response Scale showed an α of $.596$, with mild ceiling effects on two items (‘*Sympathised with me*’, $M \approx 4.3$; ‘*Showed interest in me*’, $M \approx 4.5$).

The PETS Understanding & Trust scale performed best in the battery ($\alpha = .693$), with all four items (‘*Understood my goals*’, ‘*Trusted the robot*’, ‘*Understood my needs*’, ‘*Understood my intentions*’) cohering well. It was retained as a composite scale without modification.

The overall α of $.386$ for the Custom Dialogue Scale was poor. Inspection of the inter-item correlation matrix revealed that the five items did not form a single construct but fell into two clusters with negative

cross-cluster correlations. Items 1, 3 and 4 (*‘Took initiative in guiding’, ‘Kept the conversation flowing’, ‘Aware of mood/tone changes’*) form a proactivity subscale, capturing the robot’s active conversational guidance. Items 2 and 5 (*‘Responded at the right moment’* and *‘Did not rush topic changes (Reversed)’*) form a Timing subscale, which captures the appropriateness of the robot’s conversational pacing. The low subscale α values (.138 and .545, respectively) suggest that these subscales function as indexes (the sums of related items) rather than unidimensional scales [9]. Therefore, statistical findings for these composites are interpreted item-by-item rather than assuming the composite captures an underlying trait. Following the scale modifications described above, eight outcome measures were taken forward for hypothesis testing. These are presented in Table 5.2.

Table 5.2: The 8 Outcome Measures Used in Hypothesis Testing

Outcome	Scale	Type
RoSAS Warmth (4-item)	RoSAS	Composite
PETS Emotional Response	PETS	Composite
PETS Understanding and Trust	PETS	Composite
Custom Proactivity (items 1, 3, 4)	Custom Dialogue	Index
Custom Timing (items 2, 5rev)	Custom Dialogue	Index
Naturalness	Single item	Item
Felt Comfort	Single item	Item
Enjoyment	Single item	Item

5.1.2 Manipulation Check

To verify that the topic assignment produced a meaningful difference in subjective stress levels, the post-session stress ratings were compared by contrasting each participant’s rating after their stressful topic session with their rating after their positive topic session. This kind of manipulation check is needed to verify that the experimental manipulation, in this case the topic assignment, has produced the desired difference between conditions [7]. If the stressful topic did not produce higher stress levels than the positive topic, then any differences between the two topic conditions could not be linked to emotional load.

The difference scores were not normally distributed (Shapiro-Wilk $W = 0.880$, $p = 0.039$), so a Wilcoxon signed-rank test was employed and the manipulation check reached significance. As can be seen in Figure 5.1a, the mean stress score after the stressful topic session was 4.00 ($SD = 1.51$), compared to 3.19 ($SD = 1.26$) after the positive topic session ($W = 10$, $p = .021$, $RBC = 0.744$). Participants reported significantly higher stress after the stressful topic session than after the positive topic session, confirming that the topic assignment produced a meaningful difference in subjective stress.

Examining the stress trajectory across the experiment provides additional context (Figure 5.1b). Stress was measured at three timepoints: before the experiment began and after each session. The highest post-session stress was recorded after the Baseline condition with a stressful topic ($M = 4.38$, $SD = 1.51$) and the lowest after the Baseline condition with a positive topic ($M = 2.50$, $SD = 1.07$). The Analyser condition with a stressful topic produced lower post-session stress ($M = 3.62$, $SD = 1.41$).

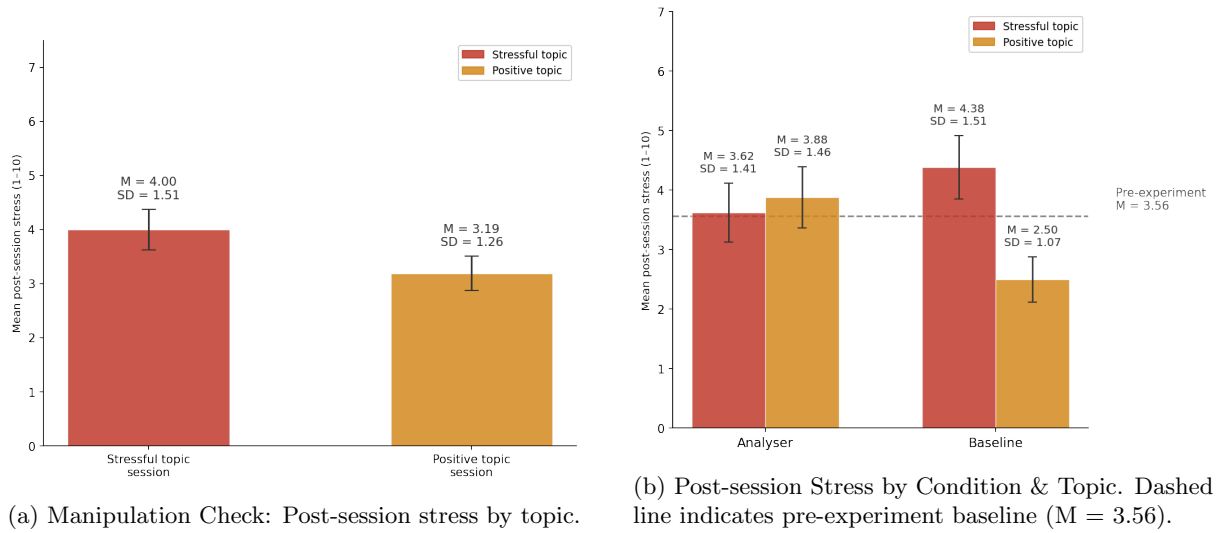


Figure 5.1: Analysis of user stress levels following the interaction. Panel (a) confirms the successful manipulation, showing higher post-session stress in the stressful topic condition than in the positive topic. Panel (b) shows the absolute stress levels in each of the four experimental conditions relative to the pre-experiment baseline.

5.2 RQ1: Main Effect of the Dialogue Analyser

The first research question was whether activating the Dialogue Analyser produced higher scores than the Baseline condition across the eight outcome measures. Those outcome measures and their results are presented in Table 5.3.

From the eight outcomes tested, only the Custom Proactivity index survived FDR correction ($p_{\text{FDR}} = .009$, $d = +0.95$). Participants rated the robot significantly higher for proactive conversational guidance when the Analyser was active, with a mean score of 3.75 compared to 3.40 under the Baseline condition. This was the most consistent finding in the dataset, with 11 out of 16 participants showing a positive Analyser-Baseline difference on this index (and only one person showing a negative difference). The remaining seven outcomes did not reach significance after correction. The Custom Timing index trended in the opposite direction ($d = -0.32$, $p_{\text{FDR}} = .699$), indicating that participants rated timing-related behaviour lower with the Analyser active than with the Baseline condition.

Table 5.3: RQ1 Results: Paired Comparisons Between Analyser and Baseline (N = 16, FDR-corrected)

Outcome	Test	Analyser M	Baseline M	Δ	Effect	p_{FDR}
RoSAS Warmth (4-item)	t-test	3.59	3.45	+0.14	$d = +0.27$	.699
PETS Emotional Response	t-test	3.80	3.82	-0.02	$d = -0.05$	.827
PETS Understanding and Trust	t-test	3.58	3.62	-0.05	$d = -0.08$	.827
Custom Proactivity (1,3,4)	t-test	3.75	3.40	+0.35	$d = +0.95$	.009**
Custom Timing (2,5rev)	t-test	2.75	3.03	-0.28	$d = -0.32$	.699
Naturalness	Wilcoxon	3.50	3.38	+0.12	RBC = +0.20	.827
Felt Comfort	Wilcoxon	3.88	4.00	-0.12	RBC = -0.25	.827
Enjoyment	Wilcoxon	4.00	3.94	+0.06	RBC = +0.09	.827

Figure 5.2 shows the individual item means for the Custom Dialogue scale. The largest Analyser advantage is visible on ‘*Kept conversation flowing*’ (RBC = +0.71, $p_{\text{raw}} = .041$, $p_{\text{FDR}} = .206$), which showed the strongest individual effect among the Proactivity items, though it did not survive FDR correction. The only item where the Baseline clearly outperformed the Analyser was ‘*Didn’t rush topic changes*’ ($d = -0.43$, $p_{\text{raw}} = .088$, $p_{\text{FDR}} = .220$), consistent with the negative trend observed for the Timing index.

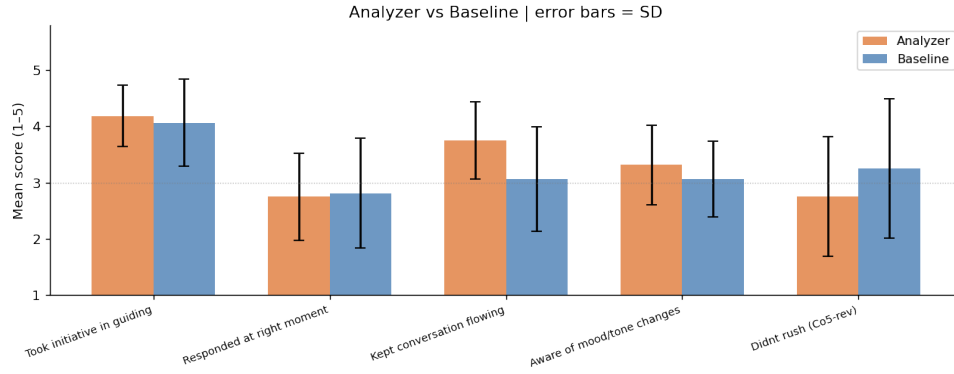


Figure 5.2: Custom Dialogue: Individual Item Analysis

5.3 RQ2: Condition x Topic Interaction

The second research question investigated whether the effect of the Dialogue Analyser was dependent on the emotional content of the conversation.

A 2 x 2 ANOVA test was performed and, from the eight outcomes, only PETS Emotional Response produced a significant interaction after FDR correction ($F(1, 14) = 12.68$, $p_{\text{FDR}} = .025$, $\eta_p^2 = .475$). As shown in Figure 5.3a, in the stressful topic group, the Analyser performed better than the Baseline on perceived emotional responsiveness ($M = 4.00$ versus $M = 3.88$, $\Delta = +0.12$), while in the positive topic group, the Baseline performed slightly better than the Analyser ($M = 3.77$ versus $M = 3.60$, $\Delta = -0.17$).

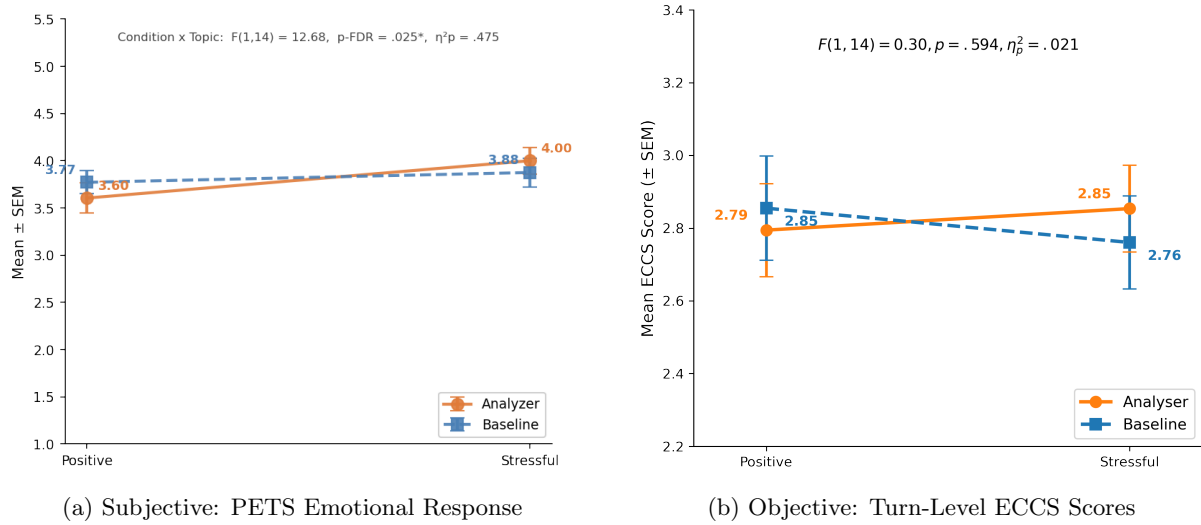


Figure 5.3: Condition x Topic Interaction. The subjective questionnaire finding (a) is mirrored by the objective transcript analysis (b), demonstrating a consistent crossover effect where the Analyser outperforms the Baseline during stressful topics ($N = 16$).

Post-hoc simple effects tested whether the Analyser outperformed Baseline within each topic group separately. As expected given the small per-group sample size ($n \approx 8$), neither post-hoc simple effect (stressful alone or positive alone) independently reached significance for PETS-ER (stressful: $p_{\text{raw}} = .558$, positive: $p_{\text{raw}} = .422$).

Complementing these subjective findings, an LLM-as-a-Judge analysis was done. The LLM-as-a-Judge ECCS score exhibits a similar crossover pattern as the PETS-ER interaction, as can be seen in Figure 5.3b. Mean ECCS scores on emotionally flagged turns were 2.85 versus 2.76 in favour of the Analyser for the stressful topic, and 2.79 versus 2.85 in favour of the Baseline for the positive topic. However, a 2x2 mixed ANOVA on the participant-level mean ECCS scores revealed that neither the main effect of condition ($F(1, 14) = 0.05$, $p_{\text{raw}} = .834$, $\eta_p^2 = .003$), the between-subjects group effect ($F(1, 14) = 0.61$, $p_{\text{raw}} = .448$, $\eta_p^2 = .042$), nor the condition x topic interaction ($F(1, 14) = 0.30$, $p_{\text{raw}} = .594$, $\eta_p^2 = .021$)

reached significance.

During emotionally charged interactions in the stress condition, the Analyser never asked more than one question per response (0 out of 82 emotional interactions, or 0%), whereas the Baseline did so in 8.5% of cases (6 out of 71). This difference was significant after FDR correction (Fisher’s exact test, $p_{\text{FDR}} = .036$). Approximately once in twelve instances where the Baseline robot was confronted with an emotionally charged user statement, it responded by asking two questions at once. In the positive condition, no meaningful difference was found between the two conditions with regard to emotional turns (13.7% versus 11.3% for the Baseline, Fisher’s exact test, $p_{\text{FDR}} = .797$).

The Custom Proactivity finding from RQ1 was replicated within the ANOVA framework, with the Condition main effect again reaching significance ($F(1, 14) = 16.45$, $p_{\text{FDR}} = .009$, $\eta_p^2 = .540$). Among the post-hoc simple effects, Custom Proactivity showed a significant advantage for the Analyser in the stressful topic group ($p_{\text{raw}} = .034$), while no significant difference was found in the positive topic group ($p_{\text{raw}} = .497$). Felt comfort showed a trend towards a condition $\times$ topic interaction ($F(1, 14) = 4.20$, $p_{\text{raw}} = .060$, $p_{\text{FDR}} = .239$), which is worth noting for replication in larger samples.

5.4 Secondary Analyses and Methodological Checks

5.4.1 Forced-Choice Preference

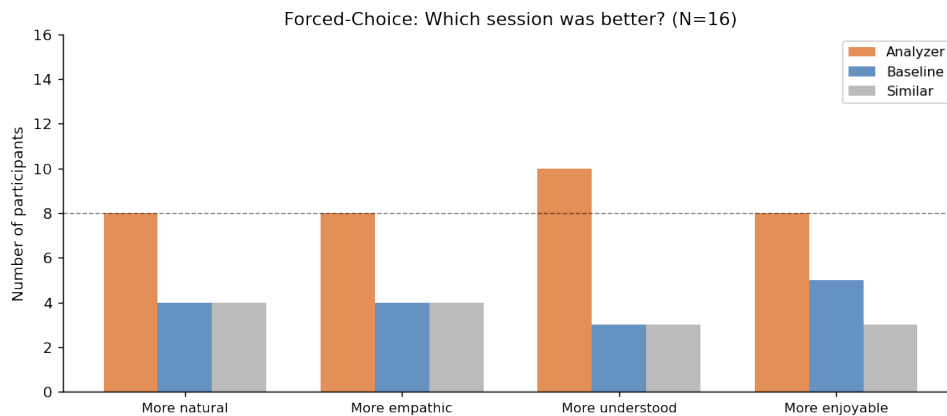


Figure 5.4: Forced-choice preference results ($N = 16$). The dashed line indicates level ($n = 8$).

After completing both sessions, participants were asked which session they felt was more natural, empathic, understood and enjoyable. Binomial tests were used to assess whether the observed preference proportions differed from chance.

The results are shown in Figure 5.4. The only question that reached significance was ‘*More understood*’, where 10 out of 13 participants who expressed a preference chose the Analyser session (77%, $p_{\text{raw}} = .046$). This aligns directionally with the PETS Understanding & Trust scale, which did not reach significance in the paired test, likely due to insufficient power at $N = 16$. For the remaining three questions, the Analyser was favoured in the right direction, with 67% preferring it for naturalness and empathy, and 62% for enjoyment, but none of these reached significance. Additionally, 69% of participants (11 out of 16) reported noticing a difference between the two sessions, suggesting that the manipulation was noticeable even when formal tests did not reach significance.

5.4.2 Within-Person Consistency

To examine how consistently the Analyser effect held at the individual level, the Analyser-Baseline difference score was computed for each participant on each outcome, as can be seen in Figure 5.5. A positive score means the participant rated the Analyser session higher than the Baseline and a negative score means the reverse.

11 out of 16 participants rated the Analyser higher for Custom Proactivity, one rated it lower and four showed no difference. For all other outcomes, participants were roughly evenly split between those who rated the Analyser higher and those who did not. The null results for the other outcomes do not reflect an absence of effect, but rather a pattern where the Analyser helped some participants but had no effect

on others. Individual characteristics such as personality, conversational style and sensitivity to the topic likely determine who benefits and who does not.

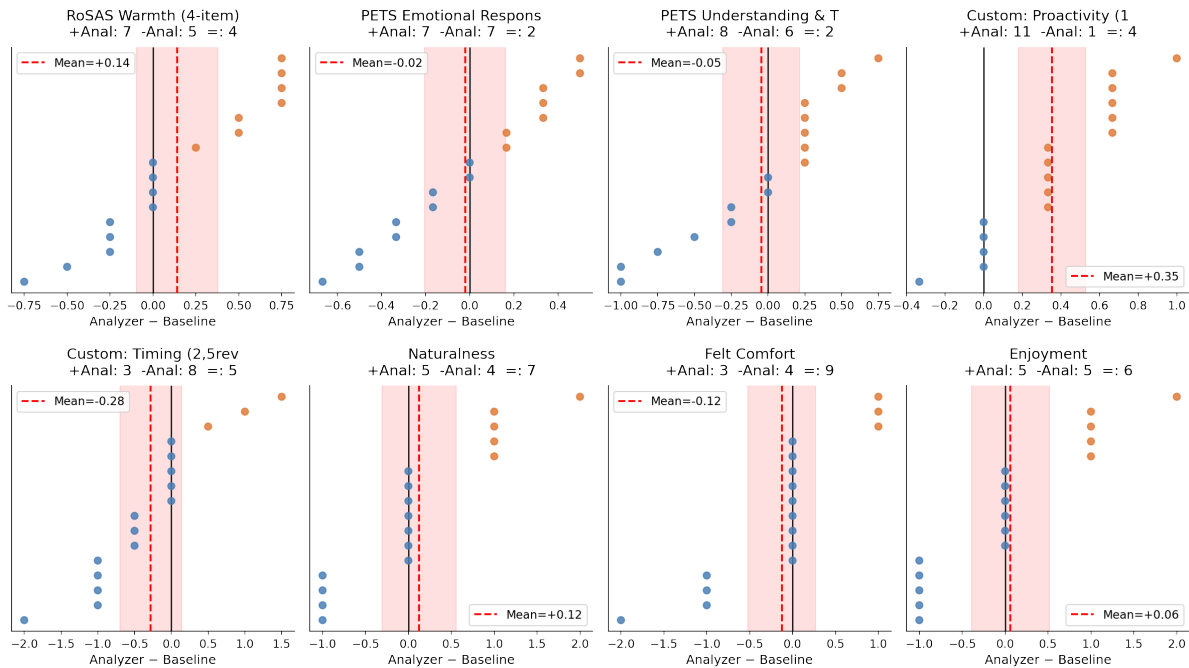


Figure 5.5: Within-person profiles showing the Analyser-Baseline difference score for each participant across all eight outcomes ($N = 16$). Orange dots indicate participants for whom the Analyser scored higher than Baseline, blue dots indicate the reverse. The red dashed line shows the group mean difference. The shaded region represents the 95% confidence interval around the group mean.

5.4.3 Session Order Effects

To verify that the counterbalancing procedure had successfully prevented session position from influencing condition comparisons, a paired t-test or Wilcoxon test was conducted on Session 1 versus Session 2 scores for each outcome, with the results being averaged across conditions. After FDR correction, no significant order effects were found on any of the eight outcomes. Participants did not systematically rate the second session differently from the first, regardless of the condition it contained.

5.4.4 Demographic Covariates

To examine whether individual characteristics moderated the Analyser’s effect, Spearman ρ correlations were computed between each demographic variable (age, tech comfort, talkativeness, AI chatbot experience and prior robot experience) and each Analyser condition outcome score, yielding 56 simultaneous tests. No demographic variable survived FDR correction. The three strongest raw associations were tech comfort and RoSAS Warmth ($\rho = +0.69$, $p_{\text{raw}} = .003$, $p_{\text{FDR}} = .179$), tech comfort and Custom Proactivity ($\rho = +0.65$, $p_{\text{raw}} = .006$, $p_{\text{FDR}} = .179$) and age and Custom Timing ($\rho = -0.62$, $p_{\text{raw}} = .010$, $p_{\text{FDR}} = .184$). Correlations involving prior robot experience are based on only four participants and should be interpreted with particular caution. Given the small sample size and the number of simultaneous tests, these associations should be treated as hypotheses for future research rather than conclusions.

5.5 Conversational Dynamics and Latency

Three behavioural metrics were extracted from the session silence logs to complement the self-report data: mean human turn duration (defined as how long participants spoke for on average per turn), robot-to-human (R→H) latency (defined as the silence between the end of a robot turn and the start of the participant’s response) and human-to-robot (H→R) latency (capturing how long the robot took to begin responding after the participant had finished speaking). This final metric is an indicator of whether the

Dialogue Analyser introduced processing delays into the interaction. All 16 participants were included in this analysis. Of these, 2 participants assigned to Group C, whose first session combined the Analyser with a stressful topic, reported finding the interaction strange. This is likely because they were required to immediately discuss a stressful topic while still habituating to the novelty of a robot, as can be seen in the questionnaire answers, P10: *‘At first, I felt a bit awkward because I am not used to talk with a robot.’* Sensitivity analyses excluding these two participants are reported alongside the main results.

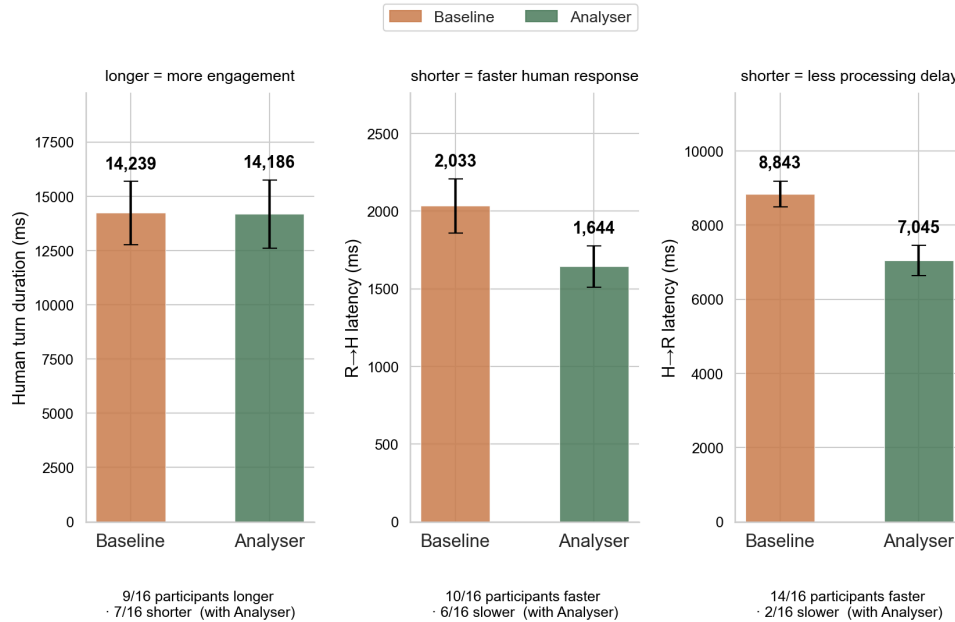


Figure 5.6: Comparison of conversational latency metrics between the Baseline and Analyser conditions ($N = 16$). Bar heights represent the group mean for each condition, with error bars indicating ± 1 Standard Error of the Mean (SEM). The text below each metric details the within-participant trends, showing the number of individuals who exhibited longer/faster times when interacting with the Analyser.

The mean duration of human turns did not differ significantly between conditions ($M = 14,186$ ms versus $M = 14,239$ ms, paired t-test: $p_{FDR} = 0.972$, $d = -0.009$), suggesting that participants spoke for comparable amounts of time regardless of condition. In the sensitivity analysis excluding the two Group C participants, participants spoke on average 1.4 seconds longer per turn in the Analyser condition ($M = 14,637$ ms versus $M = 13,236$ ms, paired t-test: $p_{FDR} = .293$, $d = +0.23$).

Examining the two latency metrics, it was found that participants responded more quickly after the robot had finished speaking in the Analyser condition than in the Baseline condition with a medium to large effect (R→H latency: $M = 1,644$ ms versus $M = 2,033$ ms; paired t-test: $p_{raw} = 0.047$, $p_{FDR} = .071$; $d = -0.625$). The robot itself also responded significantly faster under the Analyser condition than under Baseline (H→R latency: $M = 7,045$ ms versus $M = 8,843$ ms; $p_{FDR} = .009$, $RBC = -0.809$), which was a large effect. Sensitivity analyses excluding the two Group C participants yielded consistent conclusions for Human-to-Robot latency ($M = 7,153$ ms versus $M = 8,906$ ms; $p_{FDR} = .006$) and a slightly stronger effect for Robot-to-Human latency ($M = 1,561$ ms versus $M = 2,041$ ms; $p_{FDR} = .042$).

Discussion

This chapter interprets the findings reported in Chapter 5 in relation to the two research questions that formed the basis of this study:

- **Research Question 1:** Does the Dialogue Analyser enhance the user’s conversational experience compared to a Baseline system?
- **Research Question 2:** Does the effect of the Analyser depend on the emotional load of the conversation?

Overall, the results suggest that the Dialogue Analyser successfully transformed the Baseline LLM from a reactive responder into a proactive conversational partner. The system achieved a significant and consistent increase in perceived conversational proactivity. However, the data also reveal that the benefits of this architecture are context-dependent. The system significantly improved the perception of emotional understanding during stressful conversations, but not during positive ones.

The discussion is organised around three themes. Section 6.1 addresses the effect of the Dialogue Analyser on conversational proactivity and the timing trade-off it introduced. Section 6.2 examines the context-dependent nature of the Analyser’s empathic attunement, supported by the subjective scales, objective LLM-as-a-Judge metrics and forced-choice preference data. Section 6.3 discusses the latency findings.

6.1 Effect on Timing and Proactivity

While the Dialogue Analyser did not produce a broad improvement across all eight outcome measures, it did produce a significant improvement in conversational proactivity and this finding held across both topic conditions. This is important because a proactive companion robot can support a user who is too withdrawn, fatigued, or uncertain to maintain the conversation themselves. This means that the weight of the relationship is not entirely placed on the vulnerable user [37, 68]. The proactivity module, which is described in Section 3.3.4, addressed this issue by monitoring the conversation for moments when a new perspective, an unacknowledged detail or a reference to something the user had previously mentioned could prevent the exchange from becoming repetitive.

From all the outcomes, only the Custom Proactivity Index survived FDR correction across both research questions, demonstrating a large effect size ($d = +0.95$) and exhibiting the most consistent individual profile in the dataset. 11 out of 16 participants rated the Analyser higher, with only one rating it lower. Without an explicit supervisory mechanism, LLMs respond to what the user says, but they cannot anticipate where the conversation needs to go. The Baseline condition demonstrated this limitation in practice, producing fluent and contextually appropriate responses, yet failing to reliably take the initiative or maintain the directional flow of the conversation.

This was also visible in the qualitative feedback. For example, participant 14 noted about the Baseline + Positive topic condition that the robot *‘feels like it gets stuck in the same conversation and doesn’t push the conversation to a next topic or a next level’* and that *‘at the end of the conversation he kept kind of summarising the conversation we just had instead of trying to initiate another topic.’* In their Analyser session on a stressful topic, this participant wrote that, in the second session, the robot *‘seemed more interested to keep the conversation going and tried to help me find solutions to my problems’*, which is a within-person comparison that directly captures the effect that the Proactivity module was designed to produce. Participant 9 noted about the Baseline + Stressful condition that: *‘I felt that in the second conversation, it was more difficult to keep the conversation going, because the robot gave more ‘non*

sequitur’ [i.e. not logically following] *answers*.’ Participant 15 said about their Baseline + Positive session that the robot ‘*always summarised or repeated my answers, and that bothered me*’, again an example of a pattern that is consistent with a system that lacks proactive direction and therefore tends to echo the user’s own words rather than advancing the exchange.

Item-level analysis showed that the dominant driver was the item ‘*Kept conversation flowing*’ (RBC = +0.71), while other proactivity items contributed more modestly.

The timing index trended in the opposite direction ($d = -0.32$) and requires careful analysis. At the item level, ratings for ‘*Responded at the right moment*’ were comparable across conditions, so the negative trend for the Analyser condition was driven entirely by the response to the item ‘*Did not rush topic changes*’. This relates to when the robot redirected, not when it spoke. As described in Section 3.3.4, the Proactivity module operates on a heuristic trigger: a baseline of six turns, with earlier activation if engagement signals drop. There is no mechanism in the module for estimating how long a given participant wishes to remain on a topic before moving on. The module predicts what to raise next, but not when the user is ready for it and in some sessions this redirection occurred before the participant was ready to transition.

Therefore, the answer to the first research question is that, while the Dialogue Analyser did not produce a broad improvement across all outcome measures, participants did notice that the robot guided the conversation more actively and this perception remained consistent across both topic conditions and individuals.

However, while users recognised this increased proactivity, it is equally important to note what the Analyser did not change. Perceived warmth, felt comfort, naturalness and enjoyment were the same with and without the Analyser, indicating that the robot was not more pleasant overall. While several of these outcomes showed small positive directional trends (RoSAS Warmth: $d = +0.27$; Naturalness: RBC = +0.20), none survived FDR correction. As the study was only powered for large effects and sessions lasted approximately eight minutes, it is possible that these dimensions require a longer exposure period to shift meaningfully. These outcomes should be treated as inconclusive rather than null and represent a priority target for replication with larger samples and longer interaction protocols.

6.2 Context-Dependent Emotional Attunement

If overarching metrics like warmth and comfort did not show an overall improvement, where did the Analyser then add value? The data reveals that the empathic benefit of the Dialogue Analyser emerges when the conversation provides the supervisory layer with clear emotional signals to act on and it is absent when these signals are not present. This conclusion is supported by the PETS-ER crossover interaction, the LLM-as-a-Judge transcript analysis and the forced-choice preference data.

The crossover interaction on PETS-ER ($p_{FDR} = .025$, $\eta_p^2 = .475$) reveals that the impact of the Analyser on perceived emotional intelligence depended on the emotional content of the conversation. When participants discussed a stressful topic, the Dialogue Analyser outperformed the Baseline measure in terms of perceived emotional responsiveness. However, when the topic was positive the pattern reversed. The Baseline PETS-ER scores remained stable across both topic conditions. This indicates that GPT-4o, when not using the Analyser, produced a consistent but generic emotional response, regardless of whether the user was expressing distress or positivity. The flat Baseline score is evidence of the system’s context-blindness.

It is worth noting that the Baseline was already quite empathic. It was driven by a companion persona prompt that instructed GPT-4o explicitly to be attentive and friendly. The reasonably high absolute scores ($M = 3.82$) confirm that this worked as intended. Therefore, the Dialogue Analyser did not create empathy where there was none, but added situational awareness to an already empathic system. For instance, a general empathic prompt instructs the robot to appear warm in all situations, but it cannot recognise when a user who has just shared something difficult needs space rather than two questions at once, or when a significant personal moment has passed without proper acknowledgement and deserves to be revisited. This also suggests that the Analyser would be even more beneficial in systems that do not start from an empathic prompt.

The LLM-as-a-Judge transcript analysis provides two complementary signals to explain the mechanism behind this crossover. ECCS scores on emotionally flagged turns show a non-significant directional crossover pattern that is consistent with the PETS-ER interaction. The effect sizes are negligible to small, clustering around the reflection level ($ECCS = 3$) across all conditions (2.85 versus 2.76 for stressful topics and 2.79 versus 2.85 for positive topics for Analyser versus Baseline, respectively). This limits the sensitivity of the measure in detecting differences between conditions. The structural quality metrics

are more revealing. During the stressful topic, the Baseline robot asked two questions in 8.5% of emotionally flagged turns. For the Analyser, this was 0%, suggesting that the Analyser was able to avoid asking a distressed user two questions simultaneously under negative emotional pressure. This means that, on average, the robot replied by asking two questions at once approximately once in every twelve times a participant in the Baseline condition shared something emotionally loaded. As discussed in Section 2.1.2, asking multiple questions simultaneously during emotionally charged conversations feels unnatural and creates an excessive cognitive burden [61]. For a distressed elderly user, being asked, ‘*How long have you been feeling this way? Have you spoken to anyone about it?*’ in a single turn is overwhelming and not supportive. The findings of the ECCS pattern, the multi-question finding and the PETS-ER interaction are therefore all consistent with each other.

The qualitative data and stress trajectory confirm that participants felt this behavioural difference. Participant 7, whose Analyser session covered a stressful topic, reflected that ‘*I feel really understood by the robots. He also shows empathy, he really tries to help me. It’s not that he just answers and then moves on.*’, which is also something that the post-analysis is designed to prevent.

The stress trajectory data reported in Section 5.1.2 provides further supporting evidence. Post-session stress was lower after the Analyser condition on a stressful topic ($M = 3.62$) than after the equivalent Baseline session ($M = 4.38$). This suggests that the module may have partially buffered the emotional impact of the conversation.

The reversal on positive topics is more difficult to interpret. When the conversation was light, the Baseline performed better than the Analyser in terms of PETS-ER, but this difference was not statistically significant ($p_{\text{raw}} = .422$). The Analyser’s pre- and post-processing components primarily manage the emotional content of conversations by detecting signals of distress or missed emotion and they rarely intervened during sessions on positive topics where these signals were largely absent. When the Analyser did engage under these conditions, it was mostly through the proactivity mechanism rather than emotional guidance, which also explains the slightly higher multi-question rate in the positive Analyser condition. The proactivity module introduced new conversational directions through questions, without having an incorporated question limit per turn. In practice, this is less of a concern than equivalent behaviour under stress because in the positive condition, emotionally flagged turns usually reflect joy or excitement rather than distress. So for most turns, the conversation unfolded in the same way as in the Baseline condition. Therefore, the crossover pattern shows that the Analyser’s advantage was concentrated in the stressful topic condition, where there was a clear emotional target, rather than indicating that the module has a negative impact on positive conversations.

The forced-choice preference data reached the same conclusion from a different angle of measurement. When asked directly which session made them feel more understood, 77% of participants who expressed a preference chose the Analyser session ($p_{\text{raw}} = .046$). Unlike the PETS scales, which asked participants to rate specific robot behaviours after each session, the forced-choice question asked them to compare both sessions and name which one made them feel more understood. The fact that the PETS Understanding and Trust scale did not reach significance in the paired test is most likely due to insufficient power at $N = 16$. As noted in Section 4.5.3, a sample of 33 participants is normally needed to detect a medium-sized effect. The fact that a clear preference emerged with only half that number indicates that the Analyser produces a strong effect.

6.3 Latency Dynamics and Architectural Efficiency in Parallel Processing

A valid concern regarding any architecture that incorporates a supervisory processing layer into a real-time dialogue pipeline is whether the additional API calls result in latency, which participants may perceive as unnatural pauses. The results of Section 5.5 show that the Dialogue Analyser did not affect interactional fluency. In fact, both the robot and the participants responded faster under the Analyser condition than under the Baseline condition.

The Human-to-Robot latency showed that on average, the robot responded 1.79 seconds faster under the Analyser condition ($p_{\text{FDR}} = .009$, $\text{RBC} = -0.809$), which is considered a large effect. This finding is counterintuitive because, at worst, an architecture that adds parallel supervisory calls should increase response time, and at best, leave it unchanged. However, the reason for the faster response times could not be determined from the available data. Both the primary model and the Analyser’s internal calls rely on cloud-based API services whose response times are affected by server load and network variability beyond the control of the system. Therefore, the observed difference may partly reflect natural variation

in API performance across sessions rather than being a consistent property of the architecture. For this reason, the Human-to-Robot latency finding should be treated as an observed result of this specific experiment, rather than a reliable architectural property. It is discussed further as a limitation in Section 7.1.1. Nevertheless, the finding does allow us to conclude that the Analyser’s parallel architecture did not introduce the anticipated processing overhead (in this experiment).

The Robot-to-Human latency finding adds a behavioural dimension to this picture. On average, participants responded approximately 400 ms faster after the robot stopped speaking under the Analyser condition. In conversational terms, shorter Robot-to-Human intervals indicate smoother turn transitions and are associated with higher engagement and lower cognitive processing load for the listener [47]. This suggests that participants who respond sooner found the robot’s turn easier to process and engage with. The sensitivity analysis discussed in Section 5.5 showed a trend towards longer human turn duration under the Analyser condition. On average, participants spoke for an additional 1.4 seconds per turn, though this difference was not statistically significant and no meaningful difference in turn duration was observed in the full analysis of $N = 16$ participants. Due to the small sample size, this pattern should be treated as exploratory and should not be used to draw conclusions about engagement.

Despite the absence of processing overhead, perceived naturalness did not improve significantly under the Analyser condition ($RBC = +0.20$, $p_{FDR} = .827$). There are two possible, overlapping explanations. Firstly, perceived naturalness of human-robot interaction cannot be reduced to response timing alone. It depends on whether responses feel contextually appropriate, whether the conversational flow feels continuous and whether topic transitions feel smooth, none of which latency logs can capture. Secondly, the timing trend discussed in Section 6.1 may have offset any naturalness gains from improved fluency. A robot that responds quickly but redirects topics before the participant is ready may seem fast, but not natural. Therefore, the latency and naturalness data are not contradictory.

6.4 Conclusion

These findings offer a partial validation of the supervisory architecture introduced in Chapter 3. Measuring conversational experience is difficult because several factors need to be considered, such as participants’ subjective experience, their behaviour during the interaction and the robot’s actions at the turn level. This study addressed this complexity by using a triangulation approach across subjective self-report (PETS, RoSAS and custom scales), behavioural metrics (response latency and turn duration) and objective turn-level coding (ECCS scores and multi-question rate). The subjective and objective measures did not always agree in magnitude, but they agreed in direction and in the experiment conditions under which effects appeared. The Analyser improved specific dimensions of conversational quality, such as perceived empathy and emotional appropriateness under stress. The spontaneous use of words such as ‘*empathic*,’ ‘*compassionate*’ and ‘*interested*’ across different session orders and topic conditions is consistent with the CASA paradigm discussed in Chapter 2, when a system responds in socially attuned ways, users spontaneously attribute social qualities to it. This social attribution was triggered by the interaction as a whole and not by the Analyser condition specifically, which aligns with CASA’s prediction that minimal social cues are sufficient to elicit this response, regardless of the system’s underlying architecture.

However, the timing trend reveals a specific design limitation. While Section 4.4 hypothesised that naturalness might decrease due to the processing latency of parallel API calls, the system timing data did not show the anticipated overhead. Instead, the subjective questionnaire results indicated that the limitation was the system’s conversational pacing. The Analyser predicts what to say next but not how long the participant wants to stay on a given topic. When it redirected too soon, participants noticed this and rated it negatively, especially when discussing positive topics. Addressing this is the central challenge for future iterations.

Within-person profiles reveal that the Analyser’s benefit was not uniform across individuals. Apart from Custom Proactivity, the Analyser helped some participants but made no difference to others. It works most reliably in emotionally demanding conversations that give the supervisory layer a clear signal to act on.

The LLM-as-a-Judge framework was also important because, although questionnaires can indicate that participants felt understood, they cannot reveal what the robot did differently to produce that feeling. The turn-level analysis showed that the Analyser completely eliminated question stacking during emotionally loaded turns in the stressful condition, while the ECCS scores revealed a trend towards warmer, more reflective responses.

The results show that adding a supervisory LLM layer can improve certain aspects of conversational qual-

ity, such as proactivity, emotional attunement under stressful topics and perceived understanding. These improvements can be detected in brief interactions with a non-target population. However, it remains to be tested whether these improvements scale to longer interactions, older users and the emotionally complex conversations typical of real-world companion robot use.

Conclusion and future work

7.1 Limitations

While this study provides initial evidence for the viability of the Dialogue Analyser architecture, several limitations restrict the conclusions that can be drawn.

7.1.1 Methodological and Experimental Limitations

The first constraint is the sample. The sixteen participants, mostly engineering students with a mean age of 22.2 years, are far removed from the intended deployment population of elderly care home residents. The participants were comfortable with technology, socially fluent and familiar with AI chatbots. These qualities likely made the interaction easier to navigate than it would be for an older, less technologically experienced user. Therefore, replication with the target population is a logical next step.

Statistical power is also a limitation. The study was only adequately powered for large effects ($d \geq 0.75$). Detecting medium effects would have required approximately 33 participants. Non-significant results should therefore be interpreted as inconclusive rather than as evidence of no effect. The small group size of around eight participants in the topic crossover analysis also limited the ability to isolate simple effects within each topic condition, even when the overall interaction pattern was significant and large.

Session length adds a third constraint. Each session lasted approximately eight minutes, which is sufficient to detect differences in the conversational process (e.g. proactivity), but probably not long enough to change participants' general impressions of the robot's emotional state. The absence of significant effects on warmth, comfort and enjoyment does not necessarily indicate a null effect.

Although the manipulation check reached significance with the Wilcoxon test, the effect size was moderate ($RBC = 0.744$) and the absolute stress difference between the different topics was small. The topic factor in RQ2 should thus be interpreted with this in mind.

The low internal consistency of several scales limits statistical precision. All scales, except for PETS Understanding and Trust ($\alpha = .693$), fell below the acceptable threshold of .70. The Custom Dialogue scale ($\alpha = .386$) had to be split into two indexes because its items measured different constructs. The resulting subscales are treated as behavioural indexes (sums of related observable behaviours), rather than as measures of a single underlying trait. This means that their findings should be interpreted on an item-by-item basis, rather than as summary scores.

The reason why the robot responded faster under the Analyser condition could not be determined from the available data. Specific technical factors, such as variability in cloud-based API response times or network conditions during the session, may have contributed to the observed difference.

The LLM-as-a-Judge ECCS was applied to all emotionally flagged turns, regardless of whether they were positive or negative. The Analyser was designed to intervene only on negative emotional content. This mismatch means that the ECCS was partly evaluating the robot's responses to emotional content that the module was not designed to handle. This limits the sensitivity of the ECCS as an objective validation instrument, particularly in the positive topic condition.

Another limitation is that the Pepper platform's speech endpoint detection was sometimes inaccurate, causing the robot to interrupt participants before they had finished speaking. As this occurred in both conditions, the effect is expected to be minimal, but future research could quantify this with greater certainty.

7.1.2 Architectural Shortcomings

The architecture carries several known shortcomings apart from the experimental setup. For example, guidance delivery relies on soft prompt directives, meaning that GPT-4o reads each instruction and exercises its own judgement about whether and how to follow it. Therefore, the model might partially reinterpret it, or prioritise other conversational goals over it.

The proactivity and pacing logic optimises for conversational continuity, but it lacks a mechanism to determine how long a participant wishes to remain on a given topic. This explains the complaints about topic redirection captured in the Timing trend.

The current architecture does not detect indirect or implied distress, such as when elderly users implicitly signal loneliness or sadness (*‘Oh, it’s fine. I’m used to being alone’*).

Hallucinations, where the robot invents a fact that was never given to it, are also not detected. For example, if a robot says, *‘You told me your sister lives in Ghent’* and this is not true, this will not be noticed.

The architecture’s reliance on cloud-based API calls makes response timing unpredictable and outside the system’s control. This is a limitation in a real-world deployment context, where consistent timing is important for perceived naturalness and user trust.

The system consists of three parts with distinct functionalities, but only the entire system was tested. Therefore, it is not possible to determine exactly which part had an effect, and the system may have been overengineered.

Another limitation concerns long-term persona stability. The current implementation provides persona guidance at the level of individual turns, but does not adapt dynamically across multiple sessions. Furthermore, while turn-level persona consistency is monitored, it is unknown whether this is sufficient to prevent persona drift over time. It remains possible that gradual shifts in the robot’s tone or personality could still accumulate across sessions despite these turn-level checks.

7.2 Future Work

For future work, this study should be replicated with the intended target population. Testing with a properly powered sample of at least 33 elderly care home residents would provide the statistical power needed to detect medium effects that this study could not. Longer sessions would allow general impressions, such as warmth and comfort, to emerge more clearly.

At the system level, the Dialogue Analyser’s proactivity module would benefit from an adaptive pacing module. By analysing speech lengths and pause patterns, the system could more accurately determine whether a user wants to stay on a topic before redirecting the conversation. This would address the complaints about redirection captured in the Timing trend. Investigating dwell-time preferences within the target population is also relevant, given that tolerance for topic redirection may differ from that observed in a student sample. In addition to pacing, future iterations could also explore dynamic adaptation based on the user’s cognitive state. For example, vocabulary could be simplified or dwell times could be extended for users exhibiting signs of cognitive decline.

Another future iteration could evaluate the three components separately by deactivating each in turn. This would clarify the contribution of the pre-analysis, post-analysis and proactivity modules to the observed effects and determine whether the system is currently too complex or not.

Latency logging at the turn level, categorised by whether guidance was injected, could enable formal validation of the Human-to-Robot latency reduction, which currently remains a hypothesis. Systematic audio review across conditions would also be useful to quantify interruption frequency and clarify its possible contribution to the Timing trend.

The within-person profiles revealed that the Analyser helped some participants more than others. Demographics such as talkativeness and comfort with technology showed possible associations with Analyser condition outcomes. A future study with sufficient statistical power to test these moderating effects directly would clarify which user profiles would benefit most from the supervisory layer.

In terms of hardware, the system could be further enhanced by incorporating multimodal cues such as facial expression tracking. These cues could provide a fuller picture of user engagement than speech alone can offer. However, continuous visual monitoring of elderly users in care settings raises privacy considerations that would need to be addressed before deployment.

Another avenue to explore is prosodic analysis. Currently, the architecture only processes what a user says, not how they say it. In practice, this distinction could be important, as a user saying ‘no’ in a neutral tone is completely different to the same word delivered with frustration. Incorporating prosodic

features would enable the Dialogue Analyser to distinguish between these cases and adjust its behaviour accordingly.

Another future work to consider is how the system communicates with the wider care environment. For now, the robot operates in a closed loop between the AI and the resident, with no mechanism in place to share information about interactions with the nursing staff. Future work could explore how the Dialogue Analyser could summarise sessions in clinically meaningful terms and securely report these back to caregivers. This would allow interaction patterns, such as sustained frustration or mental health issues (e.g. emotional withdrawal), to become visible to those responsible for the resident's wellbeing, rather than being lost.

7.3 Conclusion

The aim of this thesis was to address a specific structural issue in social robotics. When a robot speaks naturally but cannot remember anything, users notice the difference. When it remembers, but never takes the initiative, the relationship remains superficial. When its personality shifts between sessions, trust erodes. Similarly, when its personality is not adapted to the current emotional tone, the user does not feel heard. These are failures that a base language model cannot fix on its own, requiring an additional structure built around it. The Dialogue Analyser was proposed as an architectural solution operating in parallel with the primary language model. It acts as a real-time supervisory layer, monitoring, correcting and steering without replacing the generative process beneath it.

Results show that the Analyser transformed the baseline system from a reactive responder into a proactive conversational partner. This effect was significant, consistent across individuals and observed in both topic conditions. But, perfecting the timing of these proactive interventions remains challenging, as demonstrated by the need for adaptive pacing. A second finding was that the Analyser improved perceived emotional responsiveness when participants discussed something stressful. This conclusion is supported by the PETS-ER crossover interaction and the forced-choice preference data, which showed that 77% of participants felt more understood in the Analyser session. The LLM-as-a-Judge transcript analysis showed that the Analyser eliminated question stacking entirely during emotionally charged turns in the stress condition. The convergence of subjective, behavioural (latency) and objective measurements strengthens the confidence that the effect is real rather than a measurement artefact. The third finding was counterintuitive. Rather than introducing processing delays, the parallel architecture produced significantly faster robot response times. This suggests that adding a reflective LLM layer does not necessarily result in computing delays.

The Analyser did not change perceived warmth, comfort, naturalness or enjoyment. Given the short sessions and the limited statistical power of the small sample size, these outcomes should be treated as inconclusive rather than null.

As populations age and healthcare systems come under increasing pressure, the need for sustainable long-term care solutions will only grow. Social robots are not a replacement for human caregivers, but they can support the care environment by providing consistent social and cognitive stimulation. By addressing the subtler aspects of conversation, such as personality and proactivity, this work contributes to a future where artificial companions are not just tolerated by users, but welcomed as valuable additions to elderly care.

Societal Reflection

Social robots are already present in care facilities, hospitals and homes, and their role is growing. This thesis focuses on a specific technical issue, being how to bridge the gap between user expectations of social robots and the capabilities of current systems. However, if robots are being built that remember individuals, adapt to their personalities and take the initiative in conversation, it must be considered what it means to deploy them in the real world and who would be responsible if something went wrong.

Most societal concerns focus on the software, such as the language model, the memory system and the data they store. Research shows that deploying a physically embodied robot such as Pepper in a care setting triggers a stronger psychological response than a voice from a speaker or a face on a screen [21, 78]. People instinctively apply social rules to something that occupies space, moves and looks back at them. This makes the interaction more engaging, but it also increases the expectations. For example, a user who might dismiss an odd response from a chatbot as a software glitch may feel unsettled by the same response from a robot sitting opposite them. This also amplifies the ethical concerns raised below. A physically embodied agent has a stronger social influence than a disembodied one [6], meaning a robot that confidently states a false memory is more likely to be believed, which is of importance when the user is elderly or cognitively vulnerable.

The most immediate risk of deployment is that a working companion robot becomes a reason to reduce human contact. Care institutions are under financial pressure and a conversational robot is cheaper than a human caregiver. Yet robots cannot hold someone's hand, share a meal or provide the quiet, comforting resonance of shared silence that exists only between living people. Sharkey and Sharkey [67] raised this concern over a decade ago and it remains relevant today. Robotic companions should be introduced as an addition to, rather than a replacement for, human care.

Another closely related issue is transparency. The system in this thesis uses a persona called Noa, who is framed as a lifelong friend of the user. When asked whether it is an AI, Noa stays in character without directly denying it. This is a deliberate design choice, based on research showing that constantly reminding vulnerable users they are talking to a machine can break the emotional comfort of the interaction [79]. This makes transparency outside the conversation non-negotiable. The robot can remain in character within the room, but there must be no ambiguity outside of it. The people around them, such as their family or caregivers, must fully understand the potential harms of the system, enough to give informed consent on the vulnerable person's behalf.

Another important aspect of this thesis is the risk of hallucinations for users, particularly those with dementia. The documented architecture explicitly acknowledges that the main language model can invent personal facts about the user without triggering an alert. For a cognitively healthy adult, being told '*You mentioned your sister lives in Ghent*' when this is not true is just an irritating mistake. For someone with dementia, however, it can be harmful. False memories can be reinforced and real distress can arise when family members contradict what the robot said. Furthermore, the user has no reliable way to distinguish between fact and fiction [55]. Preventing hallucinations needs to be treated as a safety requirement, not just a limitation to note and move past. Currently, that bar is not met.

The persistent memory that makes this system feel personal also raises serious privacy concerns [86]. The system stores information about daily habits, emotional states, health signals, family relationships and life history across sessions. While the GDPR gives users the legal right to check or delete their data, simply following the law is not enough to protect someone who is too confused to know their privacy is at risk. A resident suffering from cognitive decline cannot log into a database or request to wipe a memory file they do not even realise exists. This means that consent procedures must be adapted to suit the indi-

vidual user rather than being written for a general adult population. Additionally, there is the question of secondary use. The data collected to personalise a conversational companion is also a detailed record of a person's emotional patterns and cognitive habits. It is crucial to provide explicit answers regarding ownership of this data and whether it can be accessed by insurers or healthcare providers [49].

When a robotic companion causes harm by, for example, missing a crisis signal or reinforcing a false memory, it is unclear who should be held responsible. Should it be the developer? The care institution? In a live deployment involving vulnerable users, each of these issues carries a real human cost and someone needs to be held accountable for that before these systems leave the research lab.

In conclusion, the work in this thesis addresses a real problem. Loneliness among the elderly causes measurable harm and technology that can meaningfully reduce it is closer than ever before. As the population of Belgium, and Europe in general, is ageing rapidly [24] alongside a shrinking care workforce, traditional care models are under pressure [2], making the development of supportive technologies necessary. But, a working prototype and a responsibly deployed system are not the same thing. The issues raised here (substitution, transparency, hallucination risk for cognitively vulnerable users, data privacy, and accountability) represent the line between technology that helps people and technology that exploits them.

Bibliography

- [1] Vasu Agrawal et al. *Seamless Interaction: Dyadic Audiovisual Motion Modeling and Large-Scale Dataset*. July 1, 2025. DOI: 10.48550/arXiv.2506.22554. arXiv: 2506.22554[cs]. URL: <http://arxiv.org/abs/2506.22554> (visited on 12/03/2025).
- [2] Johan Albrecht. *Kan het preventiebeleid een zorginfarct voorkomen? Vergrijzing en multimorbiditeit zetten zorgsysteem onder grote druk*. Policy Report. Itinera Institute / Ghent University, 2024. URL: https://www.itinera.team/sites/default/files/content-documents/Kan%20het%20preventiebeleid%20een%20zorginfarct%20voorkomen_0.pdf.
- [3] Theo Araujo and Nadine Bol. “From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents”. In: *Computers in Human Behavior: Artificial Humans* 2.1 (Jan. 2024), p. 100030. ISSN: 29498821. DOI: 10.1016/j.chbah.2023.100030. URL: <https://linkinghub.elsevier.com/retrieve/pii/S2949882123000300> (visited on 11/03/2025).
- [4] Minja Axelsson, Micol Spitale, and Hatice Gunes. “Robots as Mental Well-being Coaches: Design and Ethical Recommendations”. In: *ACM Transactions on Human-Robot Interaction* 13.2 (June 30, 2024), pp. 1–55. ISSN: 2573-9522. DOI: 10.1145/3643457. URL: <https://dl.acm.org/doi/10.1145/3643457> (visited on 04/05/2026).
- [5] Yuntao Bai et al. *Constitutional AI: Harmlessness from AI Feedback*. Dec. 15, 2022. DOI: 10.48550/arXiv.2212.08073. arXiv: 2212.08073[cs]. URL: <http://arxiv.org/abs/2212.08073> (visited on 01/08/2026).
- [6] Wilma A. Bainbridge et al. “The Benefits of Interactions with Physically Present Robots over Video-Displayed Agents”. In: *International Journal of Social Robotics* 3.1 (Jan. 2011), pp. 41–52. ISSN: 1875-4791, 1875-4805. DOI: 10.1007/s12369-010-0082-7. URL: <http://link.springer.com/10.1007/s12369-010-0082-7> (visited on 05/20/2026).
- [7] Christoph Bartneck et al. *Human-robot interaction: an introduction*. Cambridge: Cambridge university press, 2020. ISBN: 978-1-108-67664-9.
- [8] Mylène Bilodeau-Mercure and Pascale Tremblay. “Age Differences in Sequential Speech Production: Articulatory and Physiological Factors”. In: *Journal of the American Geriatrics Society* 64.11 (Nov. 2016). ISSN: 0002-8614, 1532-5415. DOI: 10.1111/jgs.14491. URL: <https://agsjournals.onlinelibrary.wiley.com/doi/10.1111/jgs.14491> (visited on 05/13/2026).
- [9] Kenneth A. Bollen and Richard Lennox. “Conventional wisdom on measurement: A structural equation perspective”. In: *Psychological Bulletin* 110.2 (1991), pp. 305–314. DOI: 10.1037/0033-2909.110.2.305.
- [10] Rishi Bommasani et al. “On the opportunities and risks of foundation models”. In: *arXiv preprint arXiv:2108.07258* (2021).
- [11] Rezvan Boostani, Cosmin Munteanu, and Anastasia Kuzminykh. “Towards Age-Inclusive Conversational interfaces: Understanding Requirements Across Age Groups”. In: *Proceedings of the 7th ACM Conference on Conversational User Interfaces*. CUI ’25: Proceedings of the 7th ACM Conference on Conversational User Interfaces. Waterloo ON Canada: ACM, July 8, 2025, pp. 1–13. ISBN: 979-8-4007-1527-3. DOI: 10.1145/3719160.3736619. URL: <https://dl.acm.org/doi/10.1145/3719160.3736619> (visited on 11/03/2025).
- [12] Sera Buyukgoz et al. “Exploring Behavioral Creativity of a Proactive Robot”. In: *Frontiers in Robotics and AI* 8 (Nov. 26, 2021), p. 694177. ISSN: 2296-9144. DOI: 10.3389/frobt.2021.694177. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8662345/> (visited on 03/29/2026).
- [13] Colleen M. Carpinella et al. “The Robotic Social Attributes Scale (RoSAS): Development and Validation”. In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. HRI ’17: ACM/IEEE International Conference on Human-Robot Interaction. Vienna Austria: ACM, Mar. 6, 2017, pp. 254–262. ISBN: 978-1-4503-4336-7. DOI: 10.1145/2909824.3020208. URL: <https://dl.acm.org/doi/10.1145/2909824.3020208> (visited on 05/14/2026).

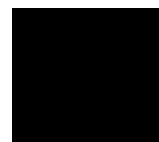
- [14] Elior Carsenti et al. “Designing a Robot’s Proactivity: Adjusting the Robot’s Communication Modality to the Participant’s Cognitive Load”. In: *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*. HRI Companion ’26. New York, NY, USA: Association for Computing Machinery, Mar. 16, 2026, pp. 552–557. ISBN: 979-8-4007-2321-6. DOI: 10.1145/3776734.3794456. URL: <https://dl.acm.org/doi/10.1145/3776734.3794456> (visited on 04/28/2026).
- [15] Yi-Pei Chen et al. “Recent Trends in Personalized Dialogue Generation: A Review of Datasets, Methodologies, and Evaluations”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024. URL: <https://aclanthology.org/2024.lrec-main.1192/>.
- [16] Yi Cheng et al. *AutoPal: Autonomous Adaptation to Users for Personal AI Companionship*. June 7, 2025. DOI: 10.48550/arXiv.2406.13960. arXiv: 2406.13960[cs]. URL: <http://arxiv.org/abs/2406.13960> (visited on 11/03/2025).
- [17] Carlos A. Cifuentes et al. “Social robots in therapy and care”. In: *Current Robotics Reports* 1 (2020), pp. 59–74.
- [18] Chloé Clavel, Matthieu Labeau, and Justine Cassell. “Socio-conversational systems: Three challenges at the crossroads of fields”. In: *Frontiers in Robotics and AI* 9 (Dec. 15, 2022), p. 937825. ISSN: 2296-9144. DOI: 10.3389/frobt.2022.937825. URL: <https://www.frontiersin.org/articles/10.3389/frobt.2022.937825/full> (visited on 04/05/2026).
- [19] I. M. Corbin and A. Dhand. “Unshared minds, decaying worlds: towards a pathology of chronic loneliness”. In: *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 49 (2024), pp. 354–366.
- [20] Samuel Rhys Cox, Grace Cheong, and Wei Tsang Ooi. *The use of deception in dementia-care robots: Should robots tell “white lies” to limit emotional distress?* Sept. 8, 2023. DOI: 10.48550/arXiv.2309.04267. arXiv: 2309.04267[cs]. URL: <http://arxiv.org/abs/2309.04267> (visited on 05/09/2026).
- [21] Kerstin Dautenhahn. “Socially intelligent robots: dimensions of human–robot interaction”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1480 (Apr. 29, 2007), pp. 679–704. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2006.2004. URL: <http://royalsocietypublishing.org/rstb/article/362/1480/679-704/20947> (visited on 05/20/2026).
- [22] J. De Witte and T. Van Regenmortel. *SILVER EMPOWERMENT - A quantitative picture of loneliness among elderly in Belgium and Europe*. Tech. rep. Leuven, Belgium: HIVA - KU Leuven, 2020.
- [23] Joan Domènech-Abella et al. “Loneliness and depression in the elderly: The role of social network”. In: *Social Psychiatry and Psychiatric Epidemiology* 52 (2017), pp. 381–390.
- [24] Eurostat. *Population structure and ageing*. European Commission Statistical Reports. 2025. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_structure_and_ageing.
- [25] Jesse Fox and Andrew Gambino. “Relationship Development with Humanoid Social Robots: Applying Interpersonal Theories to Human–Robot Interaction”. In: *Cyberpsychology, Behavior, and Social Networking* 24.5 (May 1, 2021), pp. 294–299. ISSN: 2152-2715, 2152-2723. DOI: 10.1089/cyber.2020.0181. URL: <https://journals.sagepub.com/doi/10.1089/cyber.2020.0181> (visited on 04/05/2026).
- [26] Darren George and Paul Mallery. *SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 Update*. 4th. Allyn & Bacon, 2003.
- [27] George Mason University. *GMU HRI Scale Database*. 2024. URL: <http://hriscaledatabase.psychology.gmu.edu/Database/> (visited on 05/13/2026).
- [28] *GPT-4o mini: advancing cost-efficient intelligence*. OpenAI. URL: <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/> (visited on 05/09/2026).
- [29] Lucrezia Grassi, Carmine Tommaso Recchiuto, and Antonio Sgorbissa. “Knowledge-Grounded Dialogue Flow Management for Social Robots and Conversational Agents”. In: *International Journal of Social Robotics* 14.5 (July 2022), pp. 1273–1293. ISSN: 1875-4791, 1875-4805. DOI: 10.1007/s12369-022-00868-z. arXiv: 2108.02174[cs]. URL: <http://arxiv.org/abs/2108.02174> (visited on 01/08/2026).

- [30] J. Guarnera, E. Yuen, and H. Macpherson. “The impact of loneliness and social isolation on cognitive aging: A narrative review”. In: *Journal of Alzheimer’s Disease Reports* 7 (2023), pp. 699–714.
- [31] Louise C Hawkey and John T Cacioppo. “Loneliness matters: A theoretical and empirical review of consequences and mechanisms”. In: *Annals of behavioral medicine* 40.2 (2010), pp. 218–227.
- [32] Marcel Heerink et al. “Assessing Acceptance of Assistive Social Agent Technology by Older Adults: the Almere Model”. In: *International Journal of Social Robotics* 2.4 (Dec. 2010), pp. 361–375. ISSN: 1875-4791, 1875-4805. DOI: 10.1007/s12369-010-0068-5. URL: <http://link.springer.com/10.1007/s12369-010-0068-5> (visited on 05/14/2026).
- [33] Julianne Holt-Lunstad et al. “Loneliness and Social Isolation as Risk Factors for Mortality: A Meta-Analytic Review”. In: *Perspectives on Psychological Science* 10.2 (Mar. 2015), pp. 227–237. ISSN: 1745-6916, 1745-6924. DOI: 10.1177/1745691614568352. URL: <https://journals.sagepub.com/doi/10.1177/1745691614568352> (visited on 05/20/2026).
- [34] Aurora An-Lin Hu et al. “From Failing to Exceeding User Expectations: A Large-Scale Study Exploring the Role of Expectation Gap Magnitude and Direction in Robot Adoption Intention”. In: *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion ’26)*. Edinburgh, Scotland, UK: ACM, 2026, pp. 263–267. DOI: 10.1145/3776734.3794395.
- [35] Jie Huang et al. *Large Language Models Cannot Self-Correct Reasoning Yet*. Mar. 14, 2024. DOI: 10.48550/arXiv.2310.01798. arXiv: 2310.01798[cs]. URL: <http://arxiv.org/abs/2310.01798> (visited on 04/29/2026).
- [36] ITU-T. *Recommendation P.800: Methods for subjective determination of transmission quality*. Recommendation. Geneva, Switzerland: International Telecommunication Union, Aug. 1996.
- [37] Ruben Janssens et al. “Online Prediction of User Enjoyment in Human-Robot Dialogue with LLMs”. In: *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE Press, 2025, pp. 1363–1367.
- [38] Benjamin T. Kaveladze et al. “Social interactivity in live video experiences reduces loneliness”. In: *Frontiers in Digital Health* 4 (2022), p. 859849.
- [39] Zoha Khawaja and Jean-Christophe Bélisle-Pipon. “Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots”. In: *Frontiers in Digital Health* 5 (Nov. 8, 2023), p. 1278186. ISSN: 2673-253X. DOI: 10.3389/fdgth.2023.1278186. URL: <https://www.frontiersin.org/articles/10.3389/fdgth.2023.1278186/full> (visited on 05/09/2026).
- [40] Callie Y. Kim, Christine P. Lee, and Bilge Mutlu. “Understanding large-language model (llm)-powered human-robot interaction”. In: *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. New York, NY, USA: ACM, 2024, pp. 371–380.
- [41] Filipa Landeiro et al. “Reducing social isolation and loneliness in older people: a systematic review protocol”. In: *BMJ Open* 7.5 (2017), e013778. DOI: 10.1136/bmjopen-2016-013778.
- [42] Dong Won Lee et al. *The Human Robot Social Interaction (HSRI) Dataset: Benchmarking Foundational Models’ Social Reasoning*. Apr. 7, 2025. DOI: 10.48550/arXiv.2504.13898. arXiv: 2504.13898[cs]. URL: <http://arxiv.org/abs/2504.13898> (visited on 12/03/2025).
- [43] Séverin Lemaignan et al. “Artificial cognition for social human-robot interaction: An implementation”. In: *Artificial Intelligence* 247 (June 2017), pp. 45–69. ISSN: 00043702. DOI: 10.1016/j.artint.2016.07.002. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0004370216300790> (visited on 01/08/2026).
- [44] Pengfei Liu et al. “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing”. In: *ACM Computing Surveys* 55.9 (2023), pp. 1–35.
- [45] *LLM as a Judge*. Arize AI. URL: <https://arize.com/llm-as-a-judge/> (visited on 04/28/2026).
- [46] Wing-Yue Geoffrey Louie 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*. 2014 RO-MAN: The 23rd IEEE International Symposium on Robot and Human Interactive Communication. Edinburgh, UK: IEEE, Aug. 2014, pp. 237–242. ISBN: 978-1-4799-6765-0 978-1-4799-6763-6. DOI: 10.1109/ROMAN.2014.6926259. URL: <http://ieeexplore.ieee.org/document/6926259/> (visited on 04/05/2026).

- [47] Yong Ma et al. “Beyond Words: Measuring User Experience through Speech Analysis in Voice User Interfaces”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '26)*. arXiv:2603.19904. 2026. DOI: 10.1145/3772318.3791747.
- [48] Ester Martinez-Martin and Angel P. del Pobil. “Personal robot assistants for elderly care: an overview”. In: *Personal assistants: Emerging computational technologies*. Springer, 2017, pp. 77–91.
- [49] Brent Daniel Mittelstadt and Luciano Floridi. “The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts”. In: *Science and Engineering Ethics* 22.2 (2016), pp. 303–341. DOI: 10.1007/s11948-015-9652-2.
- [50] E. Morgante et al. “Is it possible for people to develop a sense of empathy toward humanoid robots and establish meaningful relationships with them?” In: *Frontiers in Psychology* 15 (2024), p. 1391832.
- [51] Alice Nardelli, Antonio Sgorbissa, and Carmine Tommaso Recchiuto. *Designing Empathetic Companions: Exploring Personality, Emotion, and Trust in Social Robots*. Apr. 17, 2025. DOI: 10.48550/arXiv.2504.13964. arXiv: 2504.13964[cs]. URL: <http://arxiv.org/abs/2504.13964> (visited on 04/05/2026).
- [52] Bérangère Naudé, Anne-Sophie Rigaud, and Maribel Pino. “Video Calls for Older Adults: A Narrative Review of Experiments Involving Older Adults in Elderly Care Institutions”. In: *Frontiers in Public Health* 9 (2022), p. 751150. DOI: 10.3389/fpubh.2021.751150.
- [53] *NevaMind-AI/memU*. 2025-07-29. Apr. 28, 2026. URL: <https://github.com/NevaMind-AI/memU> (visited on 04/28/2026).
- [54] Samuel A. Olatunji et al. “Considerations for designing socially assistive robots for older adults”. In: *Frontiers in Robotics and AI* 12 (Oct. 16, 2025), p. 1622206. ISSN: 2296-9144. DOI: 10.3389/frobt.2025.1622206. URL: <https://www.frontiersin.org/articles/10.3389/frobt.2025.1622206/full> (visited on 05/13/2026).
- [55] Lucy Osler. “Hallucinating with AI: Distributed Delusions and “AI Psychosis””. In: *Philosophy & Technology* (2026). DOI: 10.1007/s13347-026-01034-3.
- [56] Maike Paetzel-Prüsmann et al. “An Automatic Evaluation Framework for Social Conversations with Robots”. In: *Proceedings of the 2024 International Symposium on Technological Advances in Human-Robot Interaction*. TAHRI 2024: 2024 International Symposium on Technological Advances in Human-Robot Interaction. Boulder CO USA: ACM, Mar. 9, 2024, pp. 56–64. ISBN: 979-8-4007-1661-4. DOI: 10.1145/3648536.3648543. URL: <https://dl.acm.org/doi/10.1145/3648536.3648543> (visited on 01/08/2026).
- [57] Amit Kumar Pandey and Rodolphe Gelin. “A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of Its Kind”. In: *IEEE Robotics & Automation Magazine* 25.3 (2018), pp. 40–48. DOI: 10.1109/MRA.2018.2833157.
- [58] Richard Yuanzhe Pang et al. *Leveraging Implicit Feedback from Deployment Data in Dialogue*. Feb. 1, 2024. DOI: 10.48550/arXiv.2307.14117. arXiv: 2307.14117[cs]. URL: <http://arxiv.org/abs/2307.14117> (visited on 05/09/2026).
- [59] Sung Park and Mincheol Whang. “Empathy in Human–Robot Interaction: Designing for Social Robots”. In: *International Journal of Environmental Research and Public Health* 19.3 (Feb. 8, 2022), p. 1889. ISSN: 1660-4601. DOI: 10.3390/ijerph19031889. URL: <https://www.mdpi.com/1660-4601/19/3/1889> (visited on 04/05/2026).
- [60] Maria J. Pinto and Tony Belpaeme. “Predictive Turn-Taking: Leveraging Language Models to Anticipate Turn Transitions in Human-Robot Dialogue”. In: *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Aug. 2024. URL: <https://ieeexplore.ieee.org/document/10731379/>.
- [61] Maria J. Pinto-Bernal, Kato Boels, and Tony Belpaeme. “Fostering Social Connection and Emotional Engagement to Mitigate Social Isolation in Care Homes Using Empathetic Human-Robot Conversation”. Preprint via Research Square. 2024. URL: <https://doi.org/10.21203/rs.3.rs-6811088/v1>.
- [62] Robert D. Putnam. *Bowling alone: The collapse and revival of American community*. Simon and Schuster, 2000.

- [63] Yifu Qiu et al. *Are Large Language Models Temporally Grounded?* Nov. 16, 2023. DOI: 10.48550/arXiv.2311.08398. arXiv: 2311.08398[cs.CL]. URL: <http://arxiv.org/abs/2311.08398> (visited on 05/12/2026).
- [64] Anne-Sophie Rigaud et al. "Implementation of socially assistive robots in geriatric care institutions: Healthcare professionals' perspectives and identification of facilitating factors and barriers". In: *Journal of Rehabilitation and Assistive Technologies Engineering* 11 (Aug. 2024), p. 20556683241284765. ISSN: 2055-6683, 2055-6683. DOI: 10.1177/20556683241284765. URL: <https://journals.sagepub.com/doi/10.1177/20556683241284765> (visited on 05/09/2026).
- [65] Laura Saad et al. "Choosing the "Perfect" Scale: A Primer to Evaluate Existing Scales in HRI". In: *ACM Transactions on Human-Robot Interaction* 15.2 (Mar. 31, 2026), pp. 1–30. ISSN: 2573-9522. DOI: 10.1145/3772066. URL: <https://dl.acm.org/doi/10.1145/3772066> (visited on 05/14/2026).
- [66] Matthias Schmidmaier et al. "Perceived Empathy of Technology Scale (PETS): Measuring Empathy of Systems Toward the User". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '24: CHI Conference on Human Factors in Computing Systems. Honolulu HI USA: ACM, May 11, 2024, pp. 1–18. ISBN: 979-8-4007-0330-0. DOI: 10.1145/3613904.3642035. URL: <https://dl.acm.org/doi/10.1145/3613904.3642035> (visited on 05/14/2026).
- [67] Amanda Sharkey and Noel Sharkey. "Granny and the robots: ethical issues in robot care for the elderly". In: *Ethics and Information Technology* 14.1 (Mar. 2012), pp. 27–40. ISSN: 1388-1957, 1572-8439. DOI: 10.1007/s10676-010-9234-6. URL: <http://link.springer.com/10.1007/s10676-010-9234-6> (visited on 05/20/2026).
- [68] Chapa Sirithunge, A. G. Buddhika P. Jayasekara, and D. P. Chandima. "An Evaluation of Human Conversational Preferences in Social Human-Robot Interaction". In: *Applied Bionics and Biomechanics* 2021 (Feb. 22, 2021). Ed. by Francesca Cordella, pp. 1–13. ISSN: 1754-2103, 1176-2322. DOI: 10.1155/2021/3648479. URL: <https://www.hindawi.com/journals/abb/2021/3648479/> (visited on 04/05/2026).
- [69] Gabriel Skantze. "Turn-taking in Conversational Systems and Human-Robot Interaction: A Review". In: *Computer Speech & Language* 67 (2021), p. 101178. DOI: 10.1016/j.cs1.2020.101178.
- [70] Gabriel Skantze and Bahar Irfan. "Applying General Turn-taking Models to Conversational Human-Robot Interaction". In: *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2025, pp. 859–868.
- [71] Yenee Soh et al. "Chronic loneliness and the risk of incident stroke in middle and late adulthood: a longitudinal cohort study of U.S. older adults". In: *EClinicalMedicine* 73 (2024), p. 102639. DOI: 10.1016/j.eclinm.2024.102639.
- [72] Carson Stark et al. "Dobby: a conversational service robot driven by gpt-4". In: *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2024, pp. 1362–1369.
- [73] Richard Sutcliffe. *A Survey of Personality, Persona, and Profile in Conversational Agents and Chatbots*. Dec. 31, 2023. DOI: 10.48550/arXiv.2401.00609. arXiv: 2401.00609[cs]. URL: <http://arxiv.org/abs/2401.00609> (visited on 11/03/2025).
- [74] A. R. Sutin et al. "Loneliness and risk of dementia". In: *The Journals of Gerontology: Series B* 75 (2020), pp. 1414–1422.
- [75] TEN Ecosystem. *TEN Framework Documentation: Vision*. Agora. 2024. URL: <https://theten.ai/docs#vision> (visited on 05/13/2026).
- [76] Vera Toepoel. "Ageing, leisure, and social connectedness: How could leisure help reduce social isolation of older people?" In: *Social Indicators Research* 113.1 (2013), pp. 355–372.
- [77] United Nations, Department of Economic and Social Affairs, Population Division. *World Population Ageing 2019*. ST/ESA/SER.A/444. New York: United Nations, 2019.
- [78] Joshua Wainer et al. "The Role of Physical Embodiment in Human-Robot Interaction". In: *RO-MAN 2006 - 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, Sept. 2006, pp. 117–122. DOI: 10.1109/ROMAN.2006.314404. URL: <https://ieeexplore.ieee.org/document/4107795> (visited on 05/26/2026).

- [79] Fan Wang et al. *Social Robots for People Living with Dementia: A Scoping Review on Deception from Design to Perception*. Mar. 14, 2026. DOI: 10.48550/arXiv.2507.00963. arXiv: 2507.00963[cs.HC]. URL: <http://arxiv.org/abs/2507.00963> (visited on 05/20/2026).
- [80] Aimee-Rose Wrightson-Hester et al. “A Rule-Based Conversational Agent for Mental Health and Well-Being in Young People: Formative Case Series During the Rise of Generative AI”. In: *JMIR Formative Research* 9 (Dec. 1, 2025), e69841. ISSN: 2561-326X. DOI: 10.2196/69841. URL: <https://formative.jmir.org/2025/1/e69841> (visited on 05/09/2026).
- [81] S. Zamir et al. “Video-calls to reduce loneliness and social isolation within care environments for older people: an implementation study using collaborative action research”. In: *BMC geriatrics* 18 (2018), pp. 1–13.
- [82] Saizheng Zhang et al. *Personalizing Dialogue Agents: I have a dog, do you have pets too?* Sept. 25, 2018. DOI: 10.48550/arXiv.1801.07243. arXiv: 1801.07243[cs]. URL: <http://arxiv.org/abs/1801.07243> (visited on 11/03/2025).
- [83] Zhehao Zhang et al. *Personalization of Large Language Models: A Survey*. July 13, 2025. DOI: 10.48550/arXiv.2411.00027. arXiv: 2411.00027[cs]. URL: <http://arxiv.org/abs/2411.00027> (visited on 11/03/2025).
- [84] Wayne Xin Zhao et al. “A survey of large language models”. In: *arXiv preprint arXiv:2303.18223* (2023).
- [85] Lianmin Zheng et al. “Judging LLM-as-a-judge with MT-Bench and Chatbot Arena”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023, pp. 46595–46617.
- [86] Xing Zhou et al. *MemTrust: A Zero-Trust Architecture for Unified AI Memory System*. Jan. 11, 2026. DOI: 10.48550/arXiv.2601.07004. arXiv: 2601.07004[cs.CR]. URL: <http://arxiv.org/abs/2601.07004> (visited on 05/20/2026).



Questionnaires

A.1 Per- and Post-Session Questionnaire

Per-Session Questionnaire

Block A: Robot Social Attributes (Warmth)

Instructions: “To what extent does each word describe the robot you just spoke with? Rate from 1 (Not at all) to 5 (Extremely).”

Item	Not at all	Slightly	Moderately	Very	Extremely
The robot seemed happy.	☐	☐	☐	☐	☐
The robot seemed capable of feeling.	☐	☐	☐	☐	☐
The robot acted sociable.	☐	☐	☐	☐	☐
The robot seemed compassionate.	☐	☐	☐	☐	☐
The robot appeared emotional.	☐	☐	☐	☐	☐

Block B: Robot Social Attributes (Discomfort)

Instructions: “Same scale as above. To what extent does each word describe the robot you just spoke with?”

Item	Not at all	Slightly	Moderately	Very	Extremely
The robot was scary.	☐	☐	☐	☐	☐
The robot seemed strange.	☐	☐	☐	☐	☐
The robot’s behavior was awkward.	☐	☐	☐	☐	☐
The robot seemed dangerous.	☐	☐	☐	☐	☐
The robot was awful.	☐	☐	☐	☐	☐
The robot acted aggressively.	☐	☐	☐	☐	☐

Block C: Perceived Empathy of Technology (PETS)

Instructions: “Think about how the robot responded to you during this conversation. Rate your agreement from 1 (Strongly Disagree) to 5 (Strongly Agree).”

Emotional Response (6 items)

Item	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
The robot considered my mental state.	☐	☐	☐	☐	☐
The robot seemed emotionally intelligent.	☐	☐	☐	☐	☐
The robot expressed emotions.	☐	☐	☐	☐	☐
The robot sympathized with me.	☐	☐	☐	☐	☐
The robot showed interest in me.	☐	☐	☐	☐	☐
The robot supported me in coping with an emotional situation.	☐	☐	☐	☐	☐

Understanding & Trust (4 items)

Item	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
The robot understood my goals.	☐	☐	☐	☐	☐
I trusted the robot.	☐	☐	☐	☐	☐
The robot understood my needs.	☐	☐	☐	☐	☐
The robot understood my intentions.	☐	☐	☐	☐	☐

Block D: Conversational Dynamics

Instructions: “These statements are about how the conversation unfolded. Rate your agreement from 1 (Strongly Disagree) to 5 (Strongly Agree). Items marked (R) are reverse-scored: a low rating is a positive sign.”

Item	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
The robot took initiative in guiding the conversation.	☐	☐	☐	☐	☐
The robot responded at the right moment: not too early, not too late.	☐	☐	☐	☐	☐
The robot kept the conversation flowing naturally.	☐	☐	☐	☐	☐
The robot seemed aware of changes in my mood or tone.	☐	☐	☐	☐	☐
The robot sometimes changed topic or moved on before I was ready. (R)	☐	☐	☐	☐	☐

Block E: Conversation Naturalness

Instructions: “Please rate the overall naturalness of the conversation you just had.”

Item	Very unnatural	Unnatural	Neutral	Natural	Very natural
How natural did this conversation feel?	☐	☐	☐	☐	☐

Block F: Felt Comfort

Instructions: “Please rate how comfortable you felt during the conversation you just had.”

Item	V. Unconf.	Unconf.	Neutral	Conf.	V. Conf.
How comfortable did you feel during this conversation?	☐	☐	☐	☐	☐

Instructions: “Two short questions about your overall experience and current state.”

Item	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Interacting with the robot was enjoyable.	☐	☐	☐	☐	☐

[illegible]

Please answer the following questions in your own words. There are no right or wrong answers.

1. Q1. How did you feel during this conversation?
2. Q2. Was there a moment where the robot's response surprised you (positively or negatively)? Describe it.
3. Q3. Is there something else you would like to share about your experience during this session?

Post-Experiment Questionnaire

Block 1: Comparative Judgements

Instructions: “Now that you have completed both conversations, please answer these comparison questions. Sessions are referred to by their order, not by what was different about them.”

- Q1. Which conversation felt more natural?
☐ Session 1 ☐ Session 2 ☐ They felt similar
- Q2. In which conversation did the robot seem more empathic?
☐ Session 1 ☐ Session 2 ☐ They felt similar
- Q3. In which conversation did you feel more understood?
☐ Session 1 ☐ Session 2 ☐ They felt similar
- Q4. Which conversation did you enjoy more?
☐ Session 1 ☐ Session 2 ☐ They felt similar

Block 2: Manipulation Check

Instructions: “This section helps us check whether the differences between sessions were noticeable.”

- Q5. Did you notice differences in how the robot responded between the two conversations?
☐ Yes ☐ No ☐ Not sure
- Q6. (Only if “Yes” or “Not sure” above) Could you describe what was different? Please write a few sentences in your own words.

Block 3: Final Reflections

- Q7. How would you describe your overall experience talking with the robot?
- Q8. If you had to pick one word to describe how the robot behaved across both conversations, what would it be?
- Q9. Was there anything that bothered you, or anything you really liked, about the robot’s behaviour?
- Q10. Is there anything else you would like to mention?

A.2 Demographics Questionnaire

Demographics Questionnaire

How old are you?

What is your native language?

Gender

- ☐ Male
- ☐ Female
- ☐ Non-binary / third gender
- ☐ Prefer not to say

Year of study

- ☐ 1st Bachelor
- ☐ 2nd Bachelor
- ☐ 3rd Bachelor
- ☐ 1st Master
- ☐ 2nd Master
- ☐ PhD
- ☐ Other

What is your field of study (e.g. Biomedical engineering)

How talkative would you rate yourself? (10=very talkative)

- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ 6
- ☐ 7
- ☐ 8
- ☐ 9
- ☐ 10

How stressed do you feel right now? (10=very stressed)

- ☐ 1

- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ 6
- ☐ 7
- ☐ 8
- ☐ 9
- ☐ 10

Prior experience with social robots

- ☐ None
- ☐ Seen one in person
- ☐ Interacted briefly
- ☐ Interacted extensively

Prior experience with AI chatbots (e.g. ChatGPT, Gemini, Claude)

- ☐ None
- ☐ Rarely
- ☐ Sometimes (a few times/month)
- ☐ Regularly (weekly or more)

General comfort with technology (1 = very uncomfortable, 5 = very comfortable)

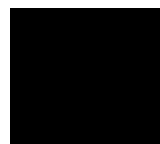
- ☐ Extremely uncomfortable
- ☐ Somewhat uncomfortable
- ☐ Neither comfortable nor uncomfortable
- ☐ Somewhat comfortable
- ☐ Extremely comfortable

Any hearing or speech difficulties?

- ☐ Yes
- ☐ No

If yes, briefly describe

--



Use of Generative AI

In accordance with the Faculty of Engineering and Architecture’s documentation requirements, this section outlines the use of generative AI tools throughout this thesis. In all cases, ownership of the content, reasoning and decisions was retained by the author.

Writing and text editing: For the introduction and literature review, the author provided the source material and citations. Claude Sonnet 4.6 was then used to improve the quality and coherence of the text. For all other chapters, the author provided core information for each section in note form. Claude then generated a first draft, which was reviewed, edited and rewritten where necessary. During restructuring of the flow of the text, Claude was also used to convert existing paragraphs into bullet point summaries to support content reorganization.

Code development: GitHub Copilot was used throughout the implementation and analysis phases. For each function, the author defined the required behaviour, underlying reasoning and intended statistical test for the analysis phase before Copilot generated the corresponding code. All suggested code was reviewed and verified before acceptance and statistical outputs were verified against expected values and cross-checked manually where appropriate.

LLM-as-judge evaluation: Claude Opus 4.7 was used as an automated evaluator in the LLM-as-judge pipeline, with the task of assessing robot dialogue turns against predefined criteria.

Diagram reproduction: The author designed conceptual diagrams in Draw.io. When the export resolution was insufficient, Claude was used to recreate the diagrams in PNG format based on screenshots of the originals. No design decisions were delegated to the model.

Table formatting & BibLaTeX: Several results tables were formatted in LaTeX with Gemini’s help. Gemini reproduced tables from Python notebook screenshots in LaTeX format. Gemini was also used to convert reference information into BibLaTeX format for the bibliography.

Audio transcription: The built-in AI transcription feature of a Google Pixel 6a was used to automatically transcribe the audio recordings of the sessions. All transcripts were reviewed by the author for accuracy before use.

Conceptual support and prompt chaining: Gemini was occasionally consulted to clarify concepts or methods during research. In some cases, Gemini was used to formulate prompts that were then used in Claude for a specific task.