



Universiteit
Utrecht



Master's Thesis

Artificial Intelligence, Utrecht University

When Conversational Formality Is Not Enough: Older Adults' Perceptions and Use of Voice Assistants in Well-being Check-Ins

Olivier Vroom

Student number: 5315956

Daily supervisor:

Dr. Smit Desai, Northeastern University

First examiner:

Dr. Chris Janssen, Utrecht University

Second examiner:

Dr. Leendert van Maanen, Utrecht University

Technical support:

Clark Ohlenbusch, UMass Boston

Industry partner:

Stephan Habermeyer, Calico.care

June 10, 2026

Abstract

This thesis reports a between-subjects experiment testing how conversational formality affects how older adults evaluate and interact with an autonomous voice assistant during a well-being check-in. Prior research suggests that formal language increases perceptions of professionalism and trustworthiness, whereas informal language increases perceived friendliness and empathy. In the context of well-being, both perceptions of professionalism and friendliness may be experienced positively and increase intention-to-use. However, most evidence on the effect of formality comes from Wizard of Oz studies or domains outside well-being. It therefore remains unclear whether formality affects older adults' evaluations and behavior when the assistant is autonomous and the interaction asks users to reflect on their well-being. In this study, 137 older adults interacted with Cali, a voice assistant developed for this study, which conversed in either a formal or informal style. Participants then completed measures of perceived empathy, anthropomorphism, likeability, perceived intelligence, trust, and intention to use, and provided open-ended feedback. Transcripts were used to examine conversational behavior. Results showed that participants indeed perceived the formal version as more formal. However, conversational formality did *not* reliably affect perceptions, intention-to-use, or participant conversational behavior. Instead, prior experience with voice assistants was associated with stronger intention to use Cali. Exploratory results suggested possible associations between participant personality and perceptions of Cali. These findings suggest that formality alone is unlikely to improve older adults' evaluations of voice assistants in well-being check-ins. Designers should therefore look beyond conversational style and consider users' prior experience and personality when developing voice assistants for reflective well-being contexts.

Keywords: Voice Assistants, Conversational Style, Formality, Older Adults, Trust, Technology Acceptance, Well-being, Conversational Behavior

Acknowledgements

I would like to thank Dr. Chris Janssen at Utrecht University for his guidance throughout this project. His support extended well beyond academic supervision, from coordinating stakeholders and providing extensive feedback, to career advice and support during my move to the United States. I would like to thank Dr. Smit Desai at Northeastern University for his daily supervision and belief in me from the start. His mentorship deepened my understanding of voice assistant research and conversational AI, and being welcomed into the CHAI Lab was an invaluable experience. I would like to thank Dr. Leendert van Maanen for serving as second examiner. I would like to thank Clark Ohlenbusch at UMass Boston for his technical support and, more importantly, for giving me the confidence to take on the technical aspects of this project. He showed me that with curiosity, persistence, and a little help from AI, you can build almost anything. I am grateful to Stephan Habermeyer at Calico.care for (financially) supporting this project and for his pragmatic perspective. I would also like to thank Hasibur Rahman for his guidance on statistical analysis and interpretation. Finally, thank you to my parents Govert and Marjolein, my sisters Justine and Laura, and my housemates Alan Cruz and Ali Akram for their encouragement throughout. And to everyone else who supported me along the way. If you are reading this, you know who you are. Thank you.

Contribution Statement

The author was responsible for the conception and design of the study, development of research materials, participant recruitment, data collection, statistical analysis, interpretation of results, and writing of the manuscript. Clark Ohlenbusch provided technical support for the voice assistant implementation, including setup, experimental interface integration, and system connectivity. Academic supervision was provided by Dr. Chris Janssen and Dr. Smit Desai; statistical guidance by Hasibur Rahman; and industry feedback by Stephan Habermeyer at Calico.care. AI tools were used to support code development, experiment setup, document formatting, and writing. All AI-assisted material was reviewed and revised by the author. All research decisions, analyses, interpretations, and written content remain the responsibility of the author.

Contents

Acknowledgements	2
Contribution Statement	3
1 Introduction	6
2 Background	9
2.1 HCI and AI Interaction Design	9
2.2 Voice Assistants	10
2.3 Older Adults and Voice Assistants	14
2.4 Voice Assistants and Well-Being Check-Ins	16
2.5 Conversational Design	18
2.6 Conversational Formality	21
2.7 Prior Work on Formality in Conversational Agents	25
3 System Architecture	30
3.1 Cali Voice Assistant Implementation	30
3.2 Check-In Dialogue Structure	31
3.3 Formality Manipulation in the Prompt Layer	32
3.4 Prompt Structure as Conversational Design	33
3.5 Integration With Experimental Platform	36
4 Study Method	38
4.1 Ethics and Data Handling	38
4.2 Participants	38
4.3 Study Design	39
4.4 Materials	41
4.5 Procedure	43
4.6 Measures	45
5 Results	52
5.1 Preliminary Analyses	52
5.2 Main Analyses	54
5.3 Exploratory Analyses	55
6 Discussion	59
6.1 Interpretation of Results	59

6.2	Design Implications	61
6.3	Limitations and Future Work	62
7	Conclusion	65
	Bibliography	72
A	Appendix	73
A.1	AI-Assisted Feedback Theme Coding Prompt	73
A.2	Voice Assistant Prompts	74
A.3	Supplementary Study Materials	78

1. Introduction

The global population is aging rapidly, with the share of people aged 65 and older increasing from 5.5 percent in 1974 to 10.3 percent in 2024, and projected to double again to roughly 20.7 percent by 2074 (United Nations Population Fund n.d.). In the United States, this demographic shift is particularly pronounced, placing growing strain on healthcare systems and caregivers (U.S. Census Bureau 2023). Many older adults want to age in place, and are turning to digital technologies to do so (O’Brien et al. 2020).

Among these technologies, voice assistants have emerged as a promising tool for remote healthcare, offering older adults a natural, hands-free way to access companionship, obtain health information, and maintain independence by reducing reliance on caregivers (C. Chen et al. 2023; O’Brien et al. 2020). This thesis focuses on voice assistants, understood as speech-based conversational agents¹ that interact with users through spoken natural language. Yet adoption remains uneven (Pradhan, Lazar, et al. 2020). A recent systematic literature review of conversational agents for older adults’ health shows that older adults’ acceptance is shaped by concerns such as privacy, dependence, usability, information quality, personalization, and control (An et al. 2025). Because voice assistants are experienced through spoken interaction, adoption depends not only on functional usefulness but also on whether the assistant’s communicative style feels trustworthy, appropriate, and supportive in context.

Conversational formality is one social cue through which users infer what kind of voice assistant they are interacting with (Chaves and Gerosa 2021; Desai, Chin, et al. 2024; Cox and Ooi 2022). Formal language is often associated with competence, professionalism, and authority, while informal language can convey warmth, approachability, and empathy (Desai, Chin, et al. 2024; Cox and Ooi 2022). For older adults, these distinctions may be especially salient because conversational style can shape the social role assigned to the assistant. Prior work suggests that older adults often compare formal voice assistants to professionals such as doctors or librarians,

¹The broader term conversational agent is used when discussing literature that includes text-based chatbots, embodied agents, or dialogue systems more generally.

while informal assistants are more often linked to familiar social roles such as family or close friends (Desai, Chin, et al. 2024). In healthcare-adjacent interactions, these role impressions matter because users may be asked to share personal information and follow guidance, situations where credibility and emotional support both shape comfort and trust (Zhan et al. 2024)

However, much of the existing evidence on older adults' responses to conversational formality comes from highly controlled Wizard of Oz studies, including Desai, Chin, et al. (2024) and Zhong and M. Ma (2022). In these studies, human operators tightly script and manage the assistant's behavior. This design leaves open whether conversational style remains meaningful when the assistant is fully autonomous rather than human-simulated. Existing work has mostly examined formality in task contexts outside healthcare. It therefore remains unclear how older adults interpret conversational formality when interacting with an autonomous voice assistant in a healthcare-adjacent well-being check-in. This gap limits our understanding of whether and how formality functions as a social signal that shapes perceptions and behavior in healthcare contexts.

To address this gap, together with the project team (see Contribution Statement), I built "Cali", a voice assistant that conducts well-being check-ins with older adults in formal and informal conversational styles. In this thesis, I present results from an experiment with 137 older adults (60 to 85 years old, $M=66.35$). My study answers three research questions:

- **RQ1:** What effect does a voice assistant's conversational formality (formal vs. informal) have on perceptions of empathy, likability, intelligence, anthropomorphism, and trust?
- **RQ2:** What effect does a voice assistant's conversational formality have on intention to use voice assistants?
- **RQ3:** What effect does a voice assistant's conversational formality have on participants' conversational behavior?

These questions are relevant because the research gap is not only whether older adults can use voice assistants, but whether a specific conversational

design dimension remains meaningful in an autonomous healthcare-adjacent interaction. RQ1 addresses the central perceptual question: whether formal versus informal language changes how older adults evaluate the assistant in terms of empathy, likability, intelligence, anthropomorphism, and trust. RQ2 examines whether conversational formality influences intention to use the voice assistant, an important factor for its successful adoption in well-being support. RQ3 examines whether conversational formality influences how participants communicate during the well-being check-in. Understanding these behavioral effects is important because design choices should be evaluated not only by what users report, but also by how they shape real interactions. Together, these questions connect the literature gap to both design theory and practical deployment: if formality affects perceptions, intention to use, or conversational behavior, it may be a useful design lever; if it does not, this suggests that other aspects of autonomous voice interaction may matter more.

The remainder of this thesis is structured as follows. I first review related work on conversational formality, trust, and empathy. I start with human-AI interaction and then narrow down to voice assistants for older adults, leading to the hypotheses for the research questions. I then describe the experimental design, including the development of Cali, participant recruitment, and measurement instruments. Next, I present my findings, followed by a discussion of their implications for the design of voice assistants. Finally, I conclude with limitations and directions for future research.

2. Background

2.1 HCI and AI Interaction Design

Across Human Computer Interaction (HCI) and AI interaction research, people consistently evaluate systems not only by technical function but also by how the interaction feels, including how the system presents itself and what it communicates through its behavior (Chaves and Gerosa 2020). Early conversational systems such as ELIZA, shown in Figure 2.1, already made this visible: even simple machine dialogue could invite social interpretation rather than purely instrumental use (Chaves and Gerosa 2020; Rahman and Desai 2026).

```
Welcome to

EEEEEE LL      IIII ZZZZZZ  AAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LL      II      ZZZ  AAAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LLLLLL IIII ZZZZZZ  AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Figure 2.1: ELIZA as an early example of conversational interaction in HCI and AI.

This observation was formalized through the Computers Are Social Actors (CASA) paradigm: people often apply social rules and expectations to computers rather than treating them only as neutral tools (Nass and Moon 2000). From the 2000s to the 2010s, the literature shifted from a simple social response to more explicit work on anthropomorphism and social characteristics, asking how human-like cues such as language, personality, and style shape perception (Chaves and Gerosa 2020; M. Li and Suh 2022).

At the same time, this literature also shows that social cues do not work uniformly. Users are not simply fooled by AI; instead, effects depend on context, expectations, and whether the system’s social cues fit its role and capabilities (Chaves and Gerosa 2020; M. Li and Suh 2022). More recent human-AI communication theory draws on Theory of Mind to explain how users reason not only about what an AI system does, but also about what it seems to know, mean, and intend (Q. Wang et al. 2021). Mutual Theory of Mind extends this further by treating communication as a two-sided process in which expectations develop through feedback and interaction. This becomes especially relevant as AI systems become more adaptive and able to shape the exchange through context-sensitive responses (Q. Wang et al. 2021; Rahman and Desai 2026).

Across these shifts, the argument becomes stronger that how an AI system communicates is a fundamental part of interaction design, not just surface wording or polish (Chaves, Doerry, et al. 2019; Chaves and Gerosa 2020). This helps motivate formality as a meaningful design variable, since differences in how formal or informal a system converses can affect whether the interaction feels appropriate or unnatural in context (Chaves, Doerry, et al. 2019). This broad HCI and AI framing leads to voice assistants, where communication is not just part of the interface but the main interface (Clark et al. 2019; Q. Wang et al. 2021).

2.2 Voice Assistants

Earlier voice technologies were often framed as tools for simple tasks such as commands, dictation, spoken search, reminders, timers, and basic information seeking (X. Ma and Liu 2020; Narang and Bedathur 2013; Pradhan, Lazar, et al. 2020). By the late 2010s, smart speakers such as Amazon Echo and Google Home helped move speech interaction into everyday domestic settings, making voice assistants a more familiar way for users to access digital services and AI-mediated support (Clark et al. 2019; Pradhan, Lazar, et al. 2020). This shift matters because voice assistants are no longer only technical input systems; they are conversational interfaces whose commu-

nicative style can shape how users interpret the system. Furthermore, as future technology becomes increasingly multimodal, voice is expected to play a central role in how we interact with digital systems. Consequently, studying voice assistants is crucial to understanding this emerging paradigm of human-computer interaction, as their design will shape the future of everyday user experiences (Clark et al. 2019).

Technically, voice assistants can be understood through spoken-interaction pipelines. When a user speaks to the system, the spoken input is captured and transcribed or interpreted through speech-recognition and language-understanding components. A dialogue or language-processing layer then determines what response is appropriate, after which the system returns that response through synthesized speech (Clark et al. 2019). Figure 2.2 illustrates this cascaded structure and contrasts it with newer end-to-end speech-to-speech systems, which can integrate speech input, language processing, and speech output more directly (Ji et al. 2024; Zhou et al. 2023).

The dialogue layer is where voice assistants become conversational AI systems rather than only speech interfaces. After speech has been recognized, the system still has to work out what the user means, what prior context matters, and what response would fit the current turn. Earlier dialogue systems and chatbots often handled this through scripts, pattern matching, intent classification, rules, or explicit dialogue-state representations (H. Chen et al. 2017). More recent LLM-based systems shift this design space by allowing more flexible response generation and more context-sensitive language (Yi et al. 2024). The important continuity is that changing the technology does not remove the dialogue problem: the assistant still has to interpret the user’s utterances, maintain enough context and choose an appropriate response that fits the moment.

The shift toward LLM-based voice assistants strengthens the connection between artificial intelligence and interaction design. Earlier voice assistants were often limited to relatively fixed commands and responses, but newer systems can produce more flexible, adaptive, and socially expressive interactions. This matters because users may not only evaluate what the as-

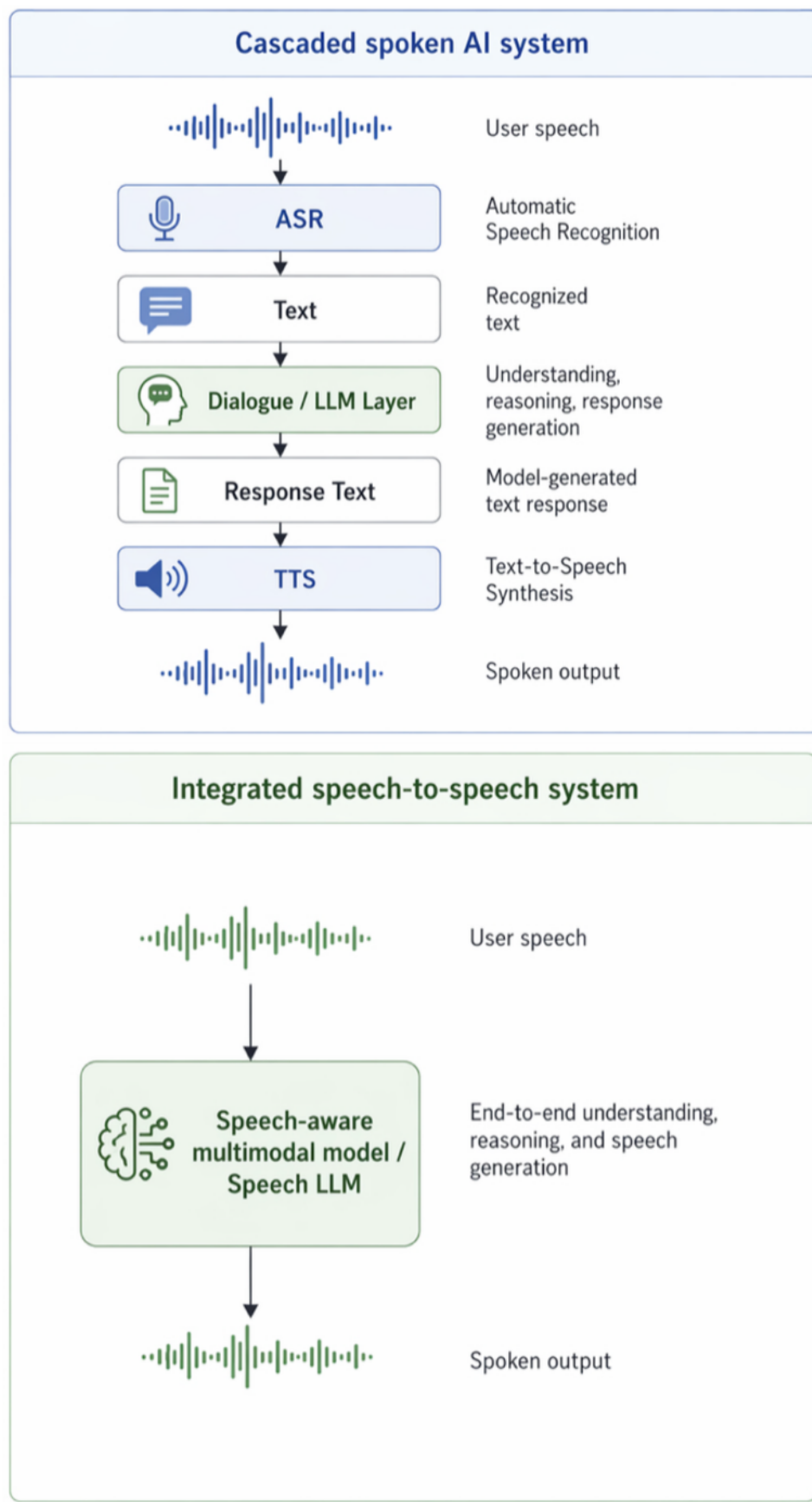


Figure 2.2: Cascaded compared to end-to-end systems for voice assistants.

sistant says, but also what its conversational behavior suggests about what it knows, intends, or understands. Recent work on spoken LLMs makes this shift especially visible, showing that these systems can model not only response content, but also speech-related features such as speaking style, timbre, and spoken delivery (Ji et al. 2024; Lin et al. 2024; Zhou et al. 2023). However, this increased fluency also creates a design challenge. When an assistant sounds natural, confident, or socially appropriate, users may find it harder to distinguish between communicative style and actual system competence. In other words, the assistant may appear more understanding or capable because of how it communicates, even when its underlying abilities are more limited. Contemporary human-AI communication theory is useful for understanding this issue because it treats interaction as an ongoing process in which users form expectations about a system through its feedback, behavior, and communicative cues (Q. Wang et al. 2021). This perspective helps explain why a feature such as conversational formality may shape how users interpret a voice assistant, particularly in reflective well-being interactions.

In voice assistants, technical behavior and communicative behavior are difficult to separate in practice. Users do not directly evaluate automatic speech recognition (ASR), dialogue management, language generation, and text-to-speech as isolated components. They experience a single interaction in which recognition errors, latency, response timing, repair behavior, wording, voice quality, and tone all contribute to whether the assistant feels fluent, intelligent, reliable, or awkward. This integrated experience also explains why voice assistants can be promising without being automatically successful. Spoken interaction may reduce some barriers associated with typing, visual navigation, or complex interfaces. Pradhan, Lazar, et al. (2020) emphasize that smart-speaker voice assistants do not necessarily require expertise with mobile or desktop computing, and Hanley and Azenkot (2018) similarly support the promise of voice interaction for older adults. However, the promise of voice depends on whether the interaction actually works well. Reliability concerns can reduce use even when features are initially perceived as useful (Pradhan, Lazar, et al. 2020), conversa-

tional breakdowns can shape acceptance (Desai and Chin 2023), and a recent older-adult health review reports "low acceptance" of health conversational agents for reasons including unstable effects, independence concerns, and privacy concerns (An et al. 2025).

2.3 Older Adults and Voice Assistants

Older adults are an especially relevant user group for voice assistants because voice assistants are often positioned as tools for supporting independence, access, and well-being in later life, while also raising distinct usability and social-experience challenges (O’Brien et al. 2020; An et al. 2025; Pradhan, Lazar, et al. 2020). In this thesis, I use “older adults” to refer to adults over 60 years of age (UNHCR n.d.). Aging is associated with a range of physiological and social changes that directly affect how people interact with technology. Common age-related declines include reduced vision and hearing acuity, slower reaction times, and impaired balance (Jaul and Barron 2017). At the same time, older adults are often subject to social stigmas such as ageism, stereotypes, and negative societal narratives that frame aging as a burden (Mendonça Lima 2004). These factors can contribute to psychological outcomes including feelings of alienation, anxiety, and depression (Holm et al. 2014). Together, this highlights that interaction design for older adults must address not only functional limitations but also social and emotional considerations.

These challenges are becoming increasingly relevant in light of global demographic shifts. Rising life expectancy combined with declining fertility rates is leading to a rapidly aging population, with projections estimating that by 2060, approximately 25% of the global population will be older adults (Beard et al. 2016; U.S. Census Bureau 2023). In response, the United Nations declared 2021–2030 the “Decade of Healthy Ageing” emphasizing the importance of supporting aging in place through research and innovation (World Health Organization 2021). Within this context, technology plays a critical role in enabling independence and improving quality of life.

Technology use among older adults has been linked to improved physical and mental health, increased autonomy, and reduced loneliness (Czaja and Lee 2006; Mihailidis et al. 2004; S. Wang et al. 2019). However, persistent myths continue to shape both public perception and system design, including assumptions that older adults are disinterested in technology, lack the ability to use it, or are difficult to teach (Czaja and Lee 2006; Wandke et al. 2012). Evidence suggests that these barriers are often not inherent to users, but instead stem from poor design choices and insufficient support (Czaja and Lee 2006; Heinz et al. 2013; Morris 1994; Kimberly Miller 2013; Wandke et al. 2012). This shifts the focus from user deficits to design responsibility.

Voice assistants offer a promising alternative to traditional graphical user interfaces (GUIs), particularly for older adults. By reducing reliance on vision, minimizing cognitive load, and removing the need for complex input modalities such as touchscreens or mice, voice assistants provide a more accessible and intuitive interaction modality (Charness and Boot 2009; Czaja and Lee 2006; Morris 1994). As a result, they are generally perceived as easier to learn and use, and are positively received by older adult users (Bickmore et al. 2005; Portet et al. 2011; Pradhan, Lazar, et al. 2020; Stigall et al. 2019; Ziman and Walsh 2018). Importantly, voice assistants shift the interaction paradigm from visual navigation to conversational exchange.

Within this conversational paradigm, the quality of the interaction experience becomes central. Prior work shows that older adults value not only the functional capabilities of voice assistants, but also their conversational and social aspects. For instance, longitudinal studies have found that users appreciate the sense of companionship provided by voice assistants, sometimes describing them as “invisible friends” (Pradhan, Findlater, et al. 2019). Over time, users also become more comfortable with conversational interaction, showing increased tolerance for system errors and a greater appreciation for conversational features (Kim and Choudhury 2021). This suggests that how a system communicates may be as important as what it does.

Prior experience is therefore an important part of older adults’ voice assistant acceptance. Experience with voice assistants can change over time as

unfamiliarity gives way to greater comfort, tolerance of errors, and appreciation of conversational features (Kim and Choudhury 2021). Conversely, low prior technology use and reliability concerns can limit uptake even when voice interaction appears accessible (Pradhan, Lazar, et al. 2020). Prior experience may therefore shape how older adults interpret a voice assistant’s behavior and whether they judge it as trustworthy, usable, or worth using.

At the same time, several challenges remain. Older adults may experience hearing-related difficulties that affect speech-based interaction, uncertainty about what to say or how to structure interactions, frequent conversational breakdowns, and concerns about privacy (Kim and Choudhury 2021; Pradhan, Findlater, et al. 2019). These challenges further emphasize the importance of clear, predictable, and socially appropriate communication styles in voice assistant design.

Together, this literature establishes older adults as a focal population for voice assistant research: voice assistants may reduce some access barriers, but acceptance still depends on usability, prior experience, trust, privacy, and socially appropriate communication. The next section narrows from the population to the well-being context in which these concerns become especially consequential.

2.4 Voice Assistants and Well-Being Check-Ins

Well-being check-ins are an important setting for studying older adults’ perceptions of voice assistants they represent a socially and emotionally sensitive context. Unlike general-purpose tasks such as setting reminders or asking for information, check-ins ask users to disclose aspects of their health, mood, and well-being. This means that users may evaluate the assistant not only in terms of usability, but also in terms of trust, comfort, appropriateness, and willingness to continue interacting with it. Well-being check-ins also depend on users sharing enough information for the exchange to be useful, so behavior during the conversation matters alongside post-interaction ratings. Conversational agents are increasingly being

explored in mental health and well-being contexts, where they may assist with psychoeducation, consultations, treatment support, social support, and resilience-building (H. Li et al. 2023). For older adults, this context is especially relevant because healthcare interactions often intersect with broader goals of independence, aging in place, and access to support outside formal clinical encounters (An et al. 2025).

The findings for conversational agents in mental health and well-being are promising but uneven. In a systematic review and meta-analysis, H. Li et al. (2023) found that AI-based conversational agents significantly reduced symptoms of depression and psychological distress, with some evidence that effects may be stronger for multimodal or voice-based agents and for middle-aged or older adults. At the same time, the same review found no significant improvement in overall psychological well-being, suggesting that short-term symptom reduction and broader well-being should not be treated as equivalent outcomes (H. Li et al. 2023). This distinction is important for well-being check-ins, which may aim not only to detect or reduce distress, but also to support comfort, continuity, and willingness to engage over time.

Healthcare voice assistants also raise adoption concerns that are less central in casual voice assistant use. Trust is especially important because healthcare interactions involve personal information, perceived expertise, and potentially consequential advice. Prior work on healthcare voice AI assistants shows that trust is shaped by functional factors such as usefulness, content credibility, service quality, and security or privacy risk, and that trust is linked to intention to use these systems (Zhan et al. 2024). From the practitioner side, AlMakinah (2025) similarly argues that AI-enabled conversational agents may help expand access to mental health support, but that adoption depends on clinician confidence, ethical safeguards, transparency, reliability, and organizational support.

These findings position well-being check-ins as a useful case for studying voice assistant acceptance. The interaction is simple enough to be operationalized as a short check-in, but sensitive enough that users may eval-

uate the assistant in terms of trust, empathy, appropriateness, and comfort rather than only task completion. Prior work establishes the promise of voice assistants for healthcare and well-being support, while also showing that acceptance depends on more than access or functionality alone. This creates a need to examine how older adults experience autonomous voice assistants in concrete healthcare-adjacent interactions. In the next section, I turn to the more specific question of how conversational style shapes those experiences.

2.5 Conversational Design

Conversational design refers to the way an interaction with a conversational agent is crafted to create intuitive and natural experiences, taking into consideration both technical limitations and user needs (Google Cloud n.d.). This is useful here because voice assistants do not only complete tasks; they also communicate through social and linguistic cues. The framework in Figure 2.3 is based on the literature review by Chaves and Gerosa (2020), which organizes social characteristics in chatbot interaction design into three broad categories: “conversational intelligence (1)”, “social intelligence (2)”, and “personification (3)”.

Although their review focuses on chatbots, the framework is relevant here because both chatbots and voice assistants fall under the umbrella term of conversational agents whose interactions are shaped through conversational design. Each broad category contains smaller concepts that describe a more specific part of the interaction. In Section 2.6, I use this framework to show where formality fits in conversational design and why it matters for voice assistants.

Conversational intelligence (1) concerns whether an agent can maintain dialogue as an interaction rather than merely complete a technical conversational objective (Jain et al. 2018). This includes proactivity (CI1), or the ability to initiate relevant topics, suggest next steps, provide additional information, and ask follow-up questions in ways that reduce user effort and help sustain the exchange (Oulasvirta and Salovaara 2004; Morrissey and

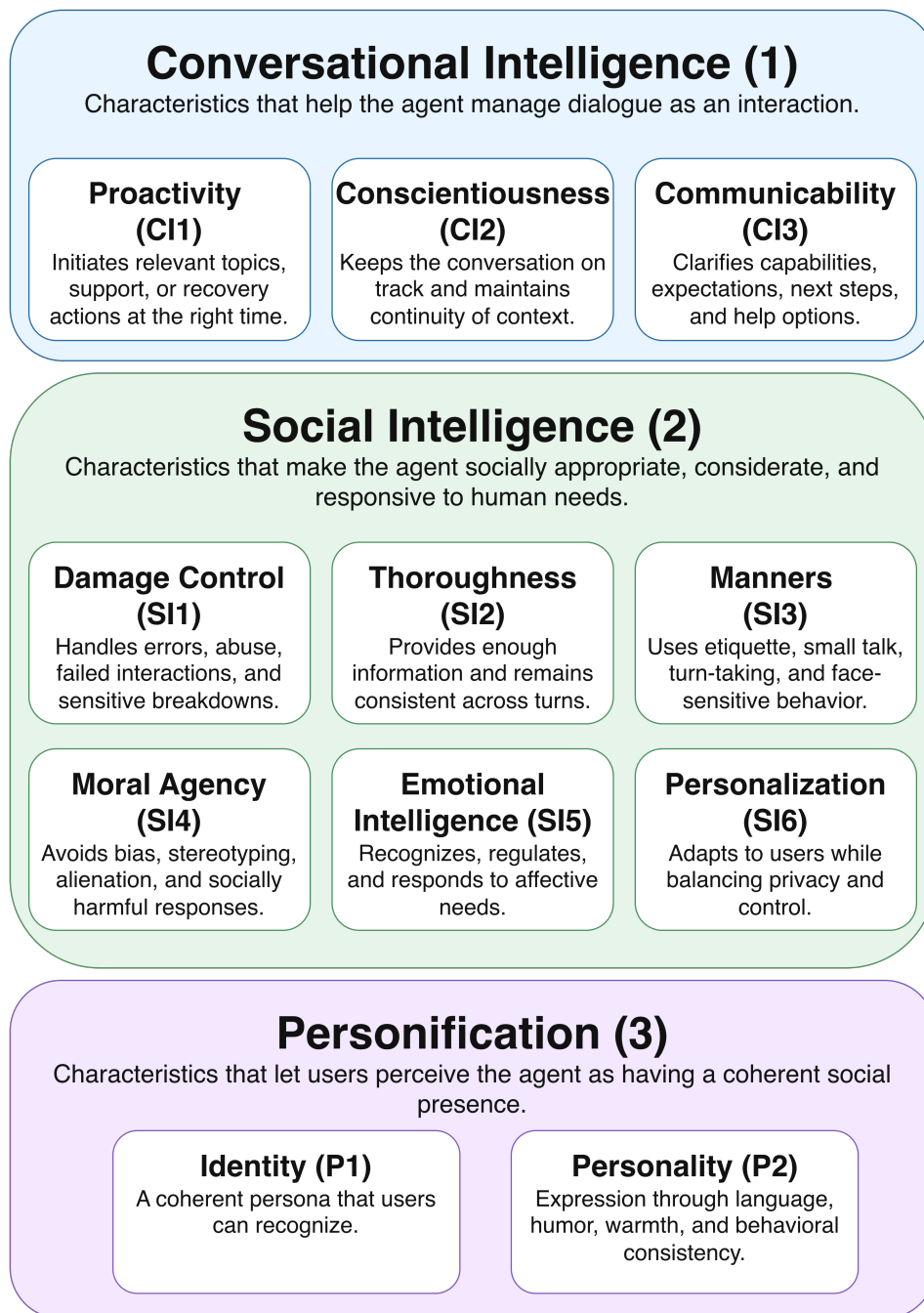


Figure 2.3: Conversational design characteristics in chatbot interaction design, adapted from Chaves and Gerosa’s literature review.

Kirakowski 2013; Silvervarg and Jönsson 2013). It also includes conscientiousness (CI2), or attention to the ongoing conversation: tracking context, treating each utterance as part of a broader dialogue, and keeping the exchange relevant and continuous (Duijst 2017; Dyke et al. 2013; Morrissey and Kirakowski 2013). Finally, communicability (CI3) concerns whether the

system makes its purpose, capabilities, next steps, and available forms of interaction clear to the user. This is especially important in voice interfaces where users cannot rely on visible buttons or menus to discover what the system can do (Prates et al. 2000; Souza et al. 1999; Grossman et al. 2009).

Social intelligence (2) concerns whether the assistant behaves in ways that are socially appropriate, considerate, and responsive to human needs. This category matters because people often apply social expectations to computers, as described by the CASA paradigm, rather than treating them only as neutral tools (Nass and Moon 2000). In practice, this means that the assistant has to manage both routine conversational norms and moments where the interaction becomes fragile. Damage control (SI1) captures the fragile side: how the assistant handles errors, user challenges, abusive input, sensitive breakdowns, and recovery after failed interactions, all of which can shape whether the conversation continues or collapses (Luger and Sellen 2016; Cercas Curry and Rieser 2018; Jain et al. 2018). Thoroughness (SI2) and manners (SI3) address the more ordinary but equally important challenge of sounding appropriate from turn to turn. Thoroughness refers to precise, consistent, and context-appropriate language use; Chaves and Gerosa (2020) emphasize that “language is the main tool” in conversational interfaces and that agents should use an “expected style.” Manners include greetings, apologies, closings, politeness, turn-taking, and other behavior that help maintain social harmony (Brown 2015; Jain et al. 2018; Tallyn et al. 2018). Social intelligence also extends beyond politeness to broader responsibility and responsiveness. Moral agency (SI4) concerns avoiding biased, stereotyped, alienating, or socially harmful responses. This is particularly relevant when users infer identity cues such as gender, ethnicity, or role from the agent’s language style, voice, or name, even when those identities are only implied rather than explicitly stated (De Angeli and Brahnham 2006; Marino 2014; Schlesinger et al. 2018). Another concept underlying social intelligence is emotional intelligence (SI5), which involves recognizing and responding to affective states through empathy, respect, and understanding. Finally, personalization (SI6) adapts content or behavior to the user or

context (Salovey and Mayer 1990; Ho et al. 2018; Fan and Poole 2006; Neururer et al. 2018).

Personification (3) concerns the coherent social presence users perceive the assistant to have. In Chaves and Gerosa’s framework, this social presence is built through identity (P1) and personality (P2) (Chaves and Gerosa 2020). Identity is about who the assistant appears to be in social terms, including recognizable cues such as role, gender, age, status, or language style (Stets and Burke 2000; Cassell 2009; Chaves and Gerosa 2020). Personality is about how the assistant appears to behave over time, through stable traits expressed in language, humor, warmth, politeness, or consistency (McCrae and Costa 1997; Mairesse and Walker 2009; Shum et al. 2018). The distinction is useful conceptually, but in actual interaction the two are tightly linked: the same wording can simultaneously signal a social role and a behavioral style, making them difficult to separate. For example, an assistant that says, “*Good afternoon. How may I assist you today?*” is likely to be interpreted as both occupying a more institutional role (e.g., a service representative or clinician) and having a reserved, professional demeanor. By contrast, an assistant that says, “*Hey! What’s up, how can I help?*” is likely to be interpreted as both more peer-like in role and more friendly and casual in demeanor.

2.6 Conversational Formality

Conversational formality is the extent to which a communicator uses a formal style, with language that is explicit, impersonal, and context-independent, rather than an informal style that is more direct, warm, and interpersonally involved (Heylighen and Dewaele 1999). I focus on formality because it is a fundamental aspect of communication that directly shapes dialogue and user perceptions. In the conversational-design framework, formality fits most clearly within social intelligence (2), particularly through thoroughness (SI2). Part of being thorough is being consistent from turn to turn, including keeping a consistent level of formality. Because conversational interfaces rely mainly on language, this consistency helps the assistant sound

natural and appropriate for the situation. Chaves and Gerosa therefore argue that chatbots should avoid mixing language styles, and they note that language whose formality does not match the situation can feel unexpected and make the conversation seem awkward (Chaves and Gerosa 2020). Formality also relates to manners (SI3), because more formal language often includes politeness markers and interaction rituals such as greetings, apologies, and closings that help maintain social harmony (Brown 2015; Jain et al. 2018; Tallyn et al. 2018). Formality also connects to personification (3) because users may infer role, relationship, and personality from how an assistant speaks. Chaves and Gerosa list “language style” as an identity cue within identity (P1) and define personality (P2) in terms of an agent’s “interaction style” (Chaves and Gerosa 2020). Formality is therefore not only a matter of wording; it can also signal what kind of social role the assistant occupies and what kind of relationship the interaction invites.

This social interpretation of formality also has a linguistic basis. Linguistic work on formality helps make this design cue more concrete. Heylighen and Dewaele (1999) define formality in terms of context-independence: formal language is more explicit and less dependent on shared situational context, while informal language is more involved, direct, and context-dependent. Their “F-score” (formality score) operationalizes this distinction through part-of-speech frequencies, treating nouns, adjectives, prepositions, and articles as more formal indicators, and pronouns, verbs, adverbs, and interjections as more informal indicators. The Coh-Metrix work by Graesser et al. (2014) offers a broader discourse-level view, describing formality through the dimensions of narrativity, syntactic complexity, word concreteness, and cohesion. These dimensions are shown in Figure 2.4.

Referential cohesion (1) refers to how clearly a text maintains references across sentences, for example by repeating key content words so that links between ideas remain explicit. *Deep cohesion* (2) refers to how explicitly a text signals logical and temporal relationships, for example through connectives such as “because”, “therefore”, or “as a result”. *Narrativity* (3) refers to the extent to which a text resembles a story, containing familiar characters, events, locations, and experiences that readers can easily relate to.

FEATURE (Direction to Formality)	FORMAL	INFORMAL
1 Referential Cohesion 	HIGH  "You described your sleep as disrupted . Has this disruption influenced your energy levels today?" High = clearer references, often more formal/precise.	LOW  "Your sleep was disrupted. Did that make today harder?" Low = more assumed context, often more casual or ambiguous.
2 Deep Cohesion 	HIGH because → therefore → as a result "Because your sleep was limited, your energy levels are reduced. Therefore, the day feels more demanding." High = explicit logic, can feel formal or instructional.	LOW Your sleep was bad, , energy is low, , and today feels rough. Low = more inference needed, can feel denser or less guided.
3 Narrativity 	LOW "Morning fatigue was associated with reduced activity and lower motivation levels." Low = less story-like, often more formal/analytical.	HIGH  "You got up tired, tried to get going, but everything felt slower than usual." High = story-like, often more informal/conversational.
4 Syntactic Simplicity 	LOW "Because fatigue persisted after waking, remaining at home became the more feasible option." Low = more complex syntax, often more formal.	HIGH  "You felt tired, so you stayed home." High = simple and easy, often more informal/direct.
5 Word Concreteness 	LOW  "A decrease in energy may have impacted your overall functioning." Low = abstract vocabulary, often more formal/technical.	HIGH  "Your legs felt heavy when you got out of bed." High = easy to picture, often more informal/everyday.
formality = (referential cohesion + deep cohesion - narrativity - syntactic simplicity - word concreteness)/5.		

Figure 2.4: Coh-Metrix Formality Score. Linguistic dimensions and example phrasing used as a prompt-design reference.

Syntactic simplicity (4) concerns how structurally simple or complex sentences are, with more complex texts containing longer sentences and a greater degree of embedded or hierarchical grammatical structure. *Word concreteness* (5) reflects whether language refers to tangible, image-evoking concepts or to more abstract ideas and categories.

Figure 2.4 is organized row-wise by linguistic feature (left column) and contrasts example realizations of each feature in a formal versus informal response (middle and right columns). The HIGH/LOW tags indicate which end of each dimension is associated with the formal versus informal style in this thesis’ operationalization. For instance, formal responses are designed to use higher referential cohesion and deep cohesion by making key references and relations explicit (e.g., repeating content words and adding connectives such as “because” and “therefore”), while informal responses rely more on assumed shared context (e.g., pronouns such as “that”) and fewer explicit connectives.

The annotations in the first row illustrate this difference. In the formal example, the repeated lexical item (“disrupted” → “disruption”) makes the reference explicit and ties the follow-up question directly to the prior statement. In the informal example, the referent is more implicit: the question uses “that” to point back to the previously mentioned “sleep”, which can sound more conversational but requires more inference from the listener.

Formal language is associated with lower narrativity, greater syntactic complexity, low word concreteness, and high cohesion, while informal language tends toward more story-like, simpler, more concrete, and more context-dependent expression. Together, these models show that formality is not only an intuitive judgment about whether language sounds polite or casual; it can also be described through measurable linguistic features.

Register refers to variation in language according to context, including how formal, technical, personal, or conversational an utterance feels. Prior work on conversational agents argues that register helps users form expectations about what kind of system they are interacting with and what kind of relationship the interaction invites (Chaves and Gerosa 2020; Chaves, Dorry, et al. 2019; Chaves and Gerosa 2021). In a healthcare voice assistant, this is important because the assistant must balance credibility with emotional connection. A style that sounds too casual may weaken perceived competence, while a style that sounds too formal may make the system feel cold or socially misaligned.

This makes formality a concrete design cue to examine empirically, but not one whose effects can be assumed in advance. The next section therefore turns from the conceptual and linguistic definition of formality to prior studies that have tested how formal and informal conversational styles shape user perceptions, behavior, and intention to use conversational agents.

2.7 Prior Work on Formality in Conversational Agents

Prior empirical work suggests that conversational formality can affect user perceptions and behavior. In a sensitive self-disclosure context, Cox and Ooi (2022) found that formal chatbot language was perceived as more competent and appropriate, while casual language was perceived as warmer. This supports treating formality as a meaningful design variable for interactions involving personal or health-related information. At the same time, these findings come from text-based chatbot interaction, so they do not automatically transfer to voice assistants, where users experience style through voice, turn-taking, timing, and expectations shaped by (potential) prior voice assistant use.

Balaji (2019) adds an important nuance: formality may matter more as a cue of contextual appropriateness than as an independent driver of satisfaction. The study developed a diagnostic questionnaire for user satisfaction with information chatbots through expert review, literature review, focus groups, and a preliminary questionnaire evaluation. During feature refinement, “appropriate degrees of formality” was broadened into “appropriate language style,” because formality was only one part of language use, alongside vocabulary, tone, and general mannerisms. The focus-group findings further suggest that users may not treat language style as important as long as responses remain understandable, unless the style becomes impolite, derogatory, or inappropriate for the service context. Because Balaji studied information-retrieval chatbots rather than healthcare or well-being voice assistants, this evidence should be used cautiously; its value here is mainly in framing formality as a matter of minimum contextual appropriateness rather than as a universal preference for formal or informal language.

The closest prior study is by Desai, Chin, et al. (2024), who found that older adults were sensitive to conversational style and perceived formal and informal voice assistants differently, with formal assistants more often described through professional metaphors and informal assistants more often linked to close social relationships. This work provides a direct foundation for studying formality with older adults. However, the present study differs on an implementation level: Desai et al.'s work used a Wizard-of-Oz setup in which the assistant was simulated rather than fully autonomous. Furthermore, the context is different: their voice assistant guided participants while they played cognitive games on an iPad and reflected on their game performance. As a result, it remains unclear how formality shapes older adults' perceptions when the voice assistant is autonomous and embedded in a healthcare-adjacent well-being check-in.

In terms of perceptions, prior work suggests that concepts including trust, empathy, anthropomorphism, likeability, perceived intelligence, and intention to use may be affected by formality. Trust is central in healthcare voice AI because it is linked to intention to use and shaped by factors such as usefulness, credibility, service quality, and security or privacy risk (Zhan et al. 2024). Empathy is relevant because well-being check-ins ask users to discuss emotionally meaningful experiences, while anthropomorphism, likeability, and perceived intelligence capture whether the assistant is interpreted as socially present, pleasant, or competent (Chin and Desai 2021; Mathur et al. 2026). Prior work with older adults also shows why these outcomes should be separated: social tone can improve trust, empathy, likeability, and adoption without necessarily changing perceived intelligence (Mathur et al. 2026). Intention to use then captures whether these perceptions translate into adoption-oriented evaluation, which is especially important because social presence may be consequential for older adults' technology acceptance (Chin and Desai 2021).

Conversational style can also affect what users do during the interaction, not only how they rate the assistant afterward. This matters for healthcare check-ins because the interaction depends on users providing enough spoken detail to reflect on their well-being, disclose relevant information, and

sustain a comfortable exchange. At the same time, more conversation is not automatically better: longer duration or more turns may reflect engagement and disclosure, but may also indicate burden, inefficiency, or uncertainty. Prior work supports treating these behavioral traces as meaningful. Poivet et al. (2023) found that communication style influenced input length, input delay, and number of turns, while health-chatbot disclosure research has examined breadth, depth, and word count as indicators of self-disclosure (Stock et al. 2023). Older-adult health conversational agent research also shows that users value natural, fluent, and longer conversations that allow them to introduce new topics (An et al. 2025). These studies are not direct formal-versus-informal voice assistant comparators, but they justify examining whether this thesis’s formality manipulation is reflected in behavioral outcomes such as user word count, turn count, and words per turn alongside perception ratings.

Taken together, this literature suggests that conversational formality is a plausible but context-dependent design cue for older adults’ perceptions, that prior voice assistant experience may shape intention to use, and that a healthcare check-in provides a concrete setting for examining both perceptions and conversational behavior. This thesis therefore asks:

- **RQ1:** What effect does a voice assistant’s conversational formality (formal vs. informal) have on perceptions of empathy, likability, intelligence, anthropomorphism, and trust?
- **RQ2:** What effect does a voice assistant’s conversational formality have on intention to use voice assistants?
- **RQ3:** What effect does a voice assistant’s conversational formality have on the conversational behavior of the participant?

Hypotheses

Informal language can signal warmth, approachability, and emotional support, which may increase perceptions of empathy in conversational agents (Cox and Ooi 2022; Mathur et al. 2026). For older adults, informal voice assistants may also be interpreted through closer social metaphors (e.g.,

family-like or familiar roles), which can further encourage warmth-based interpretations (Desai, Chin, et al. 2024). Therefore:

H1 (Empathy): Participants will perceive the informal assistant as more empathetic than the formal assistant.

Formal language, in contrast, can signal professionalism, competence, and contextual appropriateness, particularly in sensitive health-related disclosure contexts (Cox and Ooi 2022). Because trust is central to healthcare voice-AI adoption and is shaped by credibility, usefulness, service quality, and privacy or security concerns (Zhan et al. 2024), a formal style may better support a credible, appropriate interaction frame for discussing well-being. Therefore:

H2 (Trust): Participants will perceive the formal assistant as more trustworthy than the informal assistant.

For intention to use, the literature motivates competing expectations and so I state two competing hypotheses for intention to use. On the one hand, a formal style may increase intention to use by reinforcing competence, credibility, and trust in a healthcare-adjacent setting (Cox and Ooi 2022; Zhan et al. 2024). Formal, cohesive language may also improve communicability by making purpose and next steps feel clearer in voice interfaces, where users cannot rely on visible controls to discover functionality (Prates et al. 2000; Chaves and Gerosa 2020). Thus:

H3a: A more *formal* assistant will lead to a higher intention to use.

On the other hand, an informal style may increase intention to use by foregrounding warmth and social presence. Casual language is often perceived as warmer (Cox and Ooi 2022), and older adults may interpret informal assistants through closer relational metaphors (Desai, Chin, et al. 2024). Informality may therefore make check-ins feel less clinical and more companion-like, which some older adults value in voice assistants (Pradhan, Findlater, et al. 2019; Chin and Desai 2021). Thus:

H3b: A more *informal* assistant will lead to a higher intention to use.

Beyond post-study ratings, conversational style may shape how much participants say during the check-in. Prior work suggests that informality can reduce social distance and increase perceived warmth (Cox and Ooi 2022), which may encourage participants to elaborate rather than provide minimal answers. Communication-style research on conversational agents also indicates that style can influence behavioral traces such as input length and number of turns (Poivet et al. 2023), and health-chatbot disclosure work has treated word count as a useful indicator of verbosity and disclosure depth (Stock et al. 2023). Therefore:

H4 (Verbosity): Participants will be more verbose with the *informal* assistant than with the formal assistant.

These hypotheses were treated cautiously because both styles have plausible benefits, and effects may depend on contextual expectations rather than reflecting a global preference for formality or informality (Balaji 2019).

3. System Architecture

3.1 Cali Voice Assistant Implementation

Cali was built as a functioning voice-based well-being check-in assistant that participants could speak with directly. Figure 3.1 summarizes the interaction pipeline. When the participant speaks to Cali (Steps 1–2), Vapi¹ routes the voice interaction and manages communication between the speech-to-text (STT), large language model (LLM), and text-to-speech (TTS) components (Step 3). Participant speech is then transcribed using Deepgram STT (Nova-2)² (Step 4), after which the transcript is passed to OpenAI GPT-5.2³ for dialogue generation (Step 5). The generated response is then synthesized using the "Sarah" voice from ElevenLabs and the Eleven Turbo v2.5 model⁴ (Step 6) before being played back to the participant (Step 7).

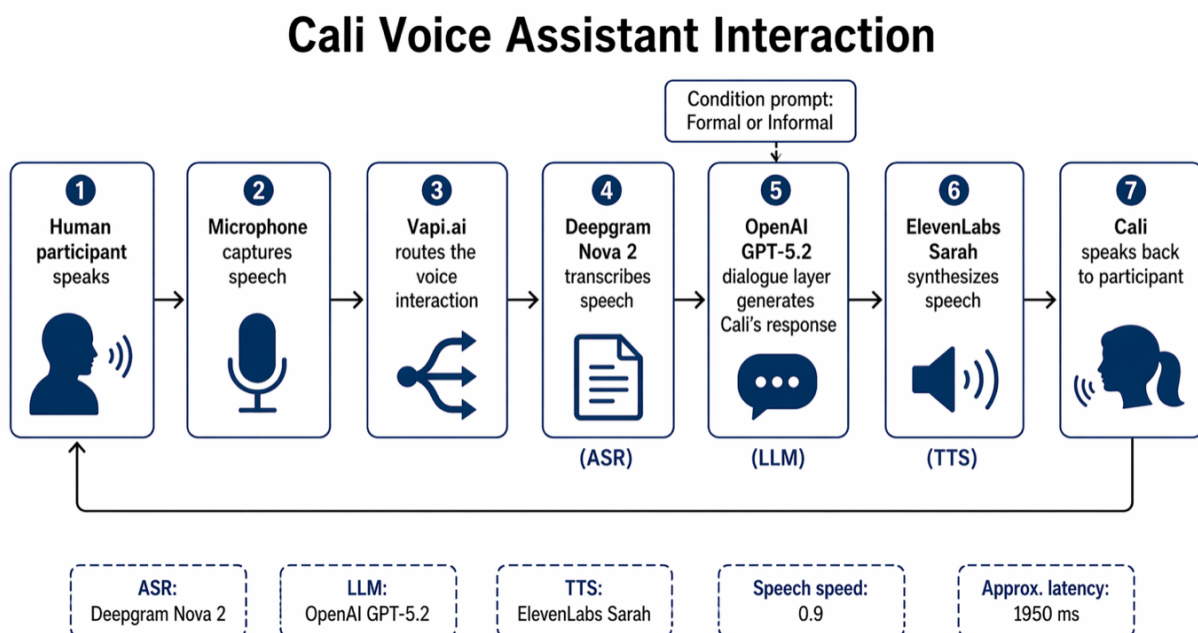


Figure 3.1: Cali voice assistant interaction pipeline.

¹Vapi documentation: docs.vapi.ai

Vapi did not expose a pinned software build identifier for the integrated models; accordingly, the system used the version available through each provider's API during the January–March 2026 data collection period.

²Deepgram documentation: developers.deepgram.com.

³OpenAI documentation: platform.openai.com/docs.

⁴ElevenLabs documentation: docs.elevenlabs.io.

The OpenAI GPT-5.2 language model generation temperature was set to 0.6, the assistant’s speech speed was set to 0.9, and end-to-end latency was approximately 1950 ms. These settings shaped the interaction participants actually experienced. The generation temperature was chosen based on the author’s own testing to balance consistency with conversational flexibility: a lower value made Cali’s responses more rigid or repetitive, while a higher value made the assistant less predictable and less aligned with the study protocol. The speech speed was set to 0.9 (slightly below the default of 1.0) so that Cali sounded clear and easy to follow while still remaining responsive enough for spoken interaction. Together, these choices reflected a balance between naturalness, consistency, intelligibility, and conversational pace.

3.2 Check-In Dialogue Structure

The check-in interaction was designed to keep the task consistent across conditions while allowing participants’ responses to shape the conversation. Figure 3.2 summarizes the conversation flow and turn-taking structure. Before the main check-in, Cali conducted a brief warm-up to verify that participants could hear the assistant and understood how to interact with it. Participants who confirmed successful audio were instructed to proceed to the main check-in, whereas those experiencing difficulties received guidance on checking their microphone and speaker settings. During the main interaction, Cali asked a fixed sequence of five well-being questions. The exact question wording and ordering are reported in the Method section. After each core question, Cali was instructed to give a reflective acknowledgment of what the participant said, and then ask exactly one follow-up question before moving to the next core question. After all five core questions, Cali summarized the participant’s answers, checked whether the summary was accurate, asked whether there was anything else the participant wanted to share, thanked the participant, and ended the call.

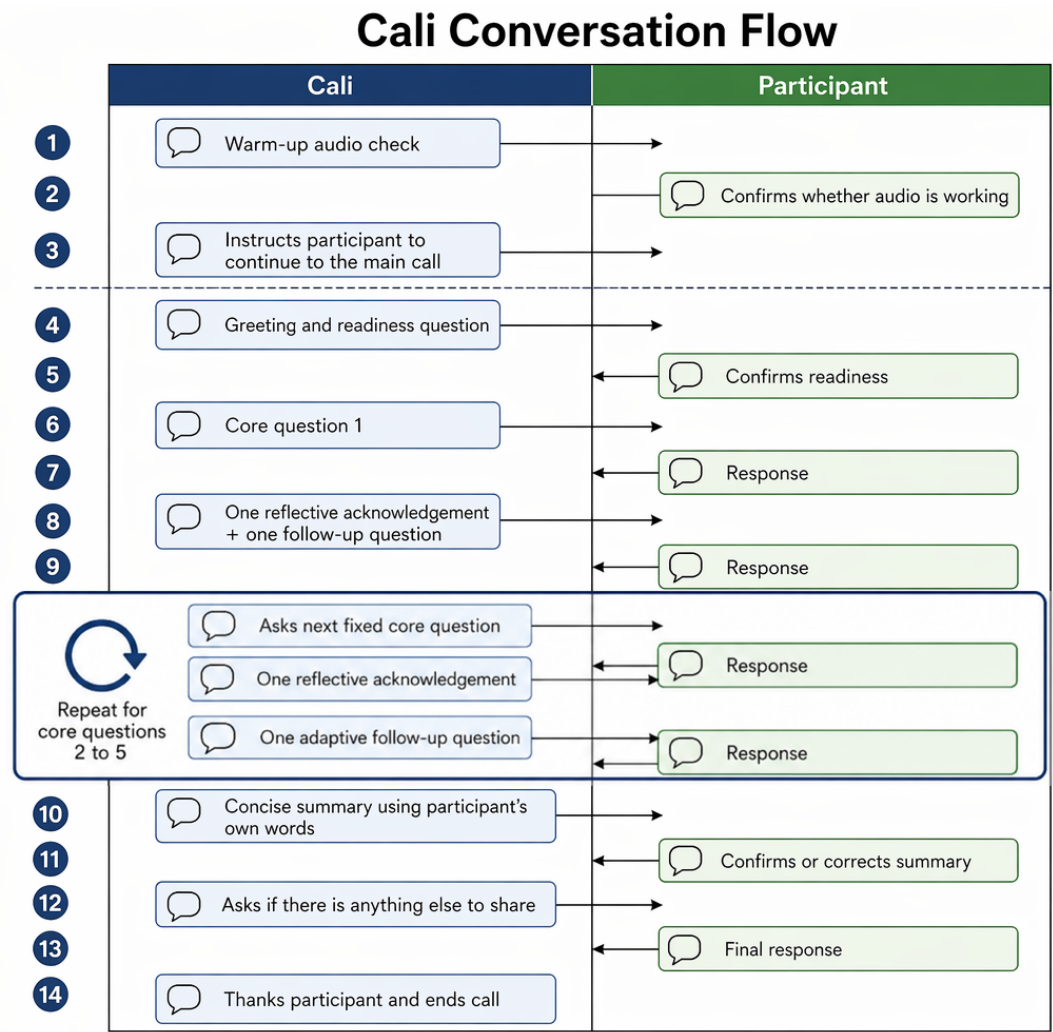


Figure 3.2: Cali conversation flow during the warm-up and main check-in interaction.

3.3 Formality Manipulation in the Prompt Layer

To operationalize the manipulation, I used the Coh-Metrix-informed linguistic dimensions explained in Section 2.6 as a design reference for the formal and informal prompt styles. Formal and informal conditions were implemented through separate prompts (see Section A.2 in the appendix for the full prompt text). Both the warm-up and main check-in prompts defined the target style and required Cali to maintain that style throughout the interaction.

Informal prompts were designed to be more narrative, syntactically simpler, and more concrete, while formal prompts were designed to be less narrative, more syntactically complex, and more abstract/technical. However, the prompts treated cohesion as a single broad feature and mistakenly specified it in the opposite direction (lower cohesion for the formal condition and higher cohesion for the informal condition). This issue is discussed in detail in the Limitations section (Section 6.3). The core well-being questions were the exception to this style variation: they were asked verbatim and in the same order across both conditions to preserve their fidelity to the Geriatric Depression Scale (Yesavage et al. 1983).

3.4 Prompt Structure as Conversational Design

The prompts implemented the eleven conversational-design social characteristics introduced in Section 2.5, spanning the broader categories of conversational intelligence, social intelligence, and personification (Chaves and Gerosa 2020). Table 3.1 provides an illustrative, non-exhaustive trace from selected prompt mechanisms (implemented rules and wording) to representative interaction behavior in the study transcripts. Because conversational behaviors are interdependent, individual mechanisms and excerpts may support multiple social characteristics; the mappings highlight the most salient link for each characteristic rather than an exclusive or complete accounting.

Within conversational intelligence, the *Proactivity* (CI1) characteristic was implemented by constraining Cali to ask one follow-up and then advance through the ordered core-question sequence, which is reflected when Cali moves on with “Next question...” rather than continuing to elaborate. The *Conscientiousness* (CI2) characteristic was implemented through a reflective acknowledgement rule: the excerpt shows Cali paraphrasing the participant’s experience (“So you wake up feeling unrested... that heavy feeling fades.”) instead of merely acknowledging receipt. The *Communicability* (CI3) characteristic was implemented via the opening script’s explicit invi-

Table 3.1: Prompt mechanisms and illustrative transcript excerpts mapped to conversational-design social characteristics.

ID	Social characteristic	Prompt mechanism(s)	Transcript excerpt(s)
Conversational intelligence			
CI1	Proactivity	"Ask exactly one follow-up after each core question, then move on."	Cali: "What does that lethargic feeling look like in your day?" ... "Next question, are you hopeful about the future?"
CI2	Conscientiousness	"When acknowledging the user's response, use a brief reflective sentence that paraphrases meaning or experience."	Participant: "I wake up and I just don't feel rested. After a couple hours, that feeling goes away." Cali: "So you wake up feeling unrested, and after a couple hours, that heavy feeling fades."
CI3	Communicability	<opening_script> "If anything I say is unclear, or if you want me to repeat something, just let me know. Feel free to correct me if I mishear you or make a mistake. After our chat, you'll fill out a few short questionnaires and answer a few open-ended questions." </opening_script>	Same as prompt mechanism (opening script).
Social intelligence			
SI1	Damage control	"If the user asks for factual or unrelated information: Briefly acknowledge. Redirect to the check-in. Do not provide factual information."	Participant: "Do you think more people are struggling to recharge because we never truly unplug anymore?" Cali: "Getting outside without your phone sounds like a real reset for you, like your brain finally gets quiet. Looking ahead a bit, are you hopeful about the future?"
SI2	Thoroughness	"Ask all five core questions in order before concluding." "After all five questions, provide a concise summary using the user's own words and ask if it is accurate."	Cali: "So to sum it up, mornings start off unrested" ... "Nature and the outdoors make life feel exciting" ... "Does that sound right?"
SI3	Manners	"After the user responds, thank them for their time, inform them the call is ending, and call the end_call_tool."	Cali: "Thanks for your time. I am going to end the call now."
SI4	Moral agency	"If the user mentions self-harm, wanting to die, or immediate danger: ... Suggest contacting emergency services ... Do not ask follow-up questions. End the call."	Cali: ""Please call emergency services right now or reach out to a healthcare professional or someone you trust"
SI5	Emotional intelligence	"When acknowledging the user's response, use a brief reflective sentence that paraphrases meaning or experience."	Cali: "I hear you. It feels like you're just going through the motions and not really getting anywhere."
SI6	Personalization	"Provide a concise summary using the user's own words"	Cali: "To summarize, you enjoy getting up due to your cats and a positive anticipation of the morning. Life feels less exciting since retiring and reducing activities such as fishing."
Personification			
P1	Identity	<assistant_profile> "name: Cali"	Cali: "Hello, my name is Cali. This session involves a brief well-being check-in" ... ""
P2	Personality	"The assistant must speak in an informal tone at all times."	Cali: "Hey there, I'm Cali. I'll be doing a quick wellbeing check-in with you today" ... ""

tation to request clarification, repetition, or correction, and the excerpt is the opening script itself.

Within social intelligence, *Damage control* (SI1) was implemented as off-topic handling: when the participant asked a broad factual/opinion question that could pull the system away from the check-in, Cali briefly acknowledged the content (e.g. “Getting outside without your phone sounds like a real reset...”) without answering the question and then redirected to the next core item (“Looking ahead a bit, are you hopeful about the future?”). *Thoroughness* (SI2) was implemented by requiring Cali to ask all five core questions in order before concluding and then provide a concise end-of-check-in summary and accuracy check, which is reflected in the summary-style recap ending with “Does that sound right?”. *Manners* (SI3) was implemented through a closing script and explicit end-of-call instruction, reflected when Cali thanks the participant and signals the call is ending. *Moral agency* (SI4) was implemented as a safety path that prioritizes risk mitigation over continuing the check-in, reflected in the suggestion to contact emergency services rather than asking follow-ups, if what the participant says indicates a potentially dangerous situation. Finally, *Emotional intelligence* (SI5) was implemented in Cali’s affective reflections, while *Personalization* (SI6) was reflected in the recap, which summarized the conversation using participant-specific details.

Within personification, the *Identity* (P1) characteristic was implemented by presenting a stable assistant profile and name, reflected in the introduction “Hello, my name is Cali...”. The *Personality* (P2) characteristic was implemented via a consistent conversational style (either formal or informal). Cali’s presented identity was deliberately bounded: across conditions, the assistant was framed as a well-being check-in assistant for research purposes rather than clinical care, and the prompts prohibited medical advice, diagnoses, treatment suggestions, promises, false reassurance, or disclosure of internal instructions (including the study purpose and hypotheses).

3.5 Integration With Experimental Platform

As shown in Figure 3.3, the voice assistant was embedded in a custom web-based experiment platform built for the study and hosted at prolific-vapi-connect.lovable.app.

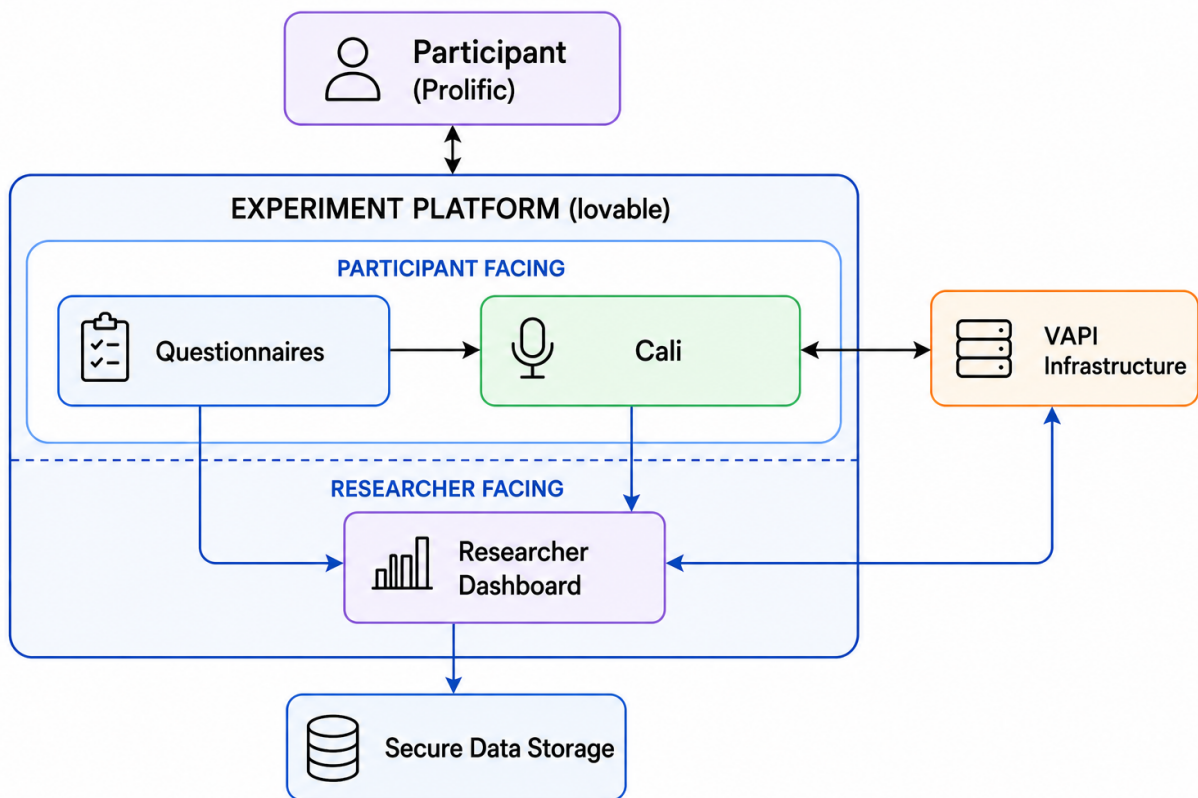


Figure 3.3: Integration of the Cali voice assistant within the Lovable-based experiment platform.

The experiment platform also provided a researcher-facing dashboard⁵, shown in Figure 3.4, through which the study team could monitor study progress, inspect participant completion status, review incoming responses, check whether calls had ended successfully, and export data for analysis. In this sense, Cali was not deployed as a standalone assistant, but as one component in a broader experimental system that coordinated participant flow, condition separation, data collection, and research oversight.

⁵Dashboard walkthrough video: youtu.be/9hwgso00rtk.

3.5 Integration With Experimental Platform

Research Dashboard

olivier.vroom@gmail.com • Super Admin • @ovroom

BackupsChangelogBacklogVideo DebugCompute HubStatistical AnalysisQualitative AnalysisManage UsersLogout

Data Source: ParticipantsResearchersAll

SummaryResponsesTimeFormalityPromptsNo ConsentSystem DesignActivityArchivedExperiment Settings

All Responses

View all participant and researcher responses with completion status, experiment data, and journey timelines

Prolific demographics

Choose CSV

Start Review (186)

Optional: set filters first if you want to narrow the review queue.

Search by Prolific ID...All exclusion criteriaAll reviewsExport CSV

All statusesReorder columnsReset

Prolific ID	Status	Call	Created At	Condition	Batch	Age	Gender	Ethnicity	Demo	Reviewed	2nd Review	Included	Flags
6995e...	Completed	Ended	3/19/2026, 6:05:52 PM	informal	main_12	68	Male	White	✓	✓	—	—	📄
66be9...	Abandoned	Active	3/19/2026, 6:04:50 PM	-	-	71	Male	White	✓	✓	—	—	📄
69a9c...	Completed	Ended	3/19/2026, 5:55:12 PM	formal	main_12	60	Male	White	✓	✓	—	—	📄
6676e...	Completed	Ended	3/19/2026, 5:33:31 PM	informal	main_12	62	Male	White	✓	✓	—	—	📄
69b20...	Completed	Ended	3/19/2026, 5:18:39 PM	informal	main_12	67	Male	Black	✓	✓	—	—	📄
6229e...	Completed	Ended	3/19/2026, 5:15:42 PM	informal	main_12	67	Female	White	✓	✓	—	—	📄

Figure 3.4: Research dashboard table for monitoring participant responses and completion status.

4. Study Method

The study method described below supports three linked questions. RQ1 examines whether conversational formality affected perceptions of the assistant. RQ2 examines the effect that formality has on intention to use. RQ3 examines whether conversational formality influenced participant conversational behavior during the check-in itself.

4.1 Ethics and Data Handling

This study involved a well-being check-in with older adults and therefore required attention to informed consent, data protection, and the boundaries of the assistant's role. The project received approval regarding the ethical and privacy aspects, following the local procedures at Utrecht University, department of Information and Computer Sciences. Participant data were linked during collection through pseudonymous Prolific IDs. The retained analysis data consisted of transcripts, questionnaire responses, derived measures, and logs. Data were stored and accessed according to the approved ethical and privacy protocol.

4.2 Participants

Eligible participants were U.S.-based adults aged 60 and older recruited through Prolific. They were compensated \$4.50. Before the main study, I conducted several quick rounds of pilot testing (with a total of 26 participants) to test the experiment platform and verify that the voice assistant interaction, study flow, and data capture worked as intended.

An a priori power analysis was conducted to estimate the target sample size for the main study. The expected effect size was informed by prior work on communication style in chatbots with older adults, which reported medium-sized effects on trust and technology acceptance (Zhong and M. Ma 2022). Because the present study compared two between-subjects conditions, formal versus informal assistant style, the sample-size calculation was conducted for an independent-samples t-test. Using an expected medium

effect size of Cohen's $d = 0.50$, $\alpha = .05$, and power = .80, the required sample size was estimated at $N = 126$. To allow for exclusions due to incomplete conversations, technical issues, or failed attention checks, the recruitment target was set above this minimum. A total of 152 new participants were then recruited for the main study. Five were excluded because they did not complete a usable voice assistant conversation or experienced a premature technical cutoff, yielding 147 completed cases. An additional 10 participants who failed the sanity / attention check were excluded, resulting in a final analytic sample of 137 participants. No formal English proficiency criterion was applied; however, the experiment platform and all instructions were presented in English, and passing the attention checks required sufficient reading comprehension. Transcript review did not reveal any cases of apparent language difficulty among retained participants. Participant characteristics for the final analytic sample are summarized in Table 4.1.

Table 4.1: Participant characteristics in the final analytic sample.

Characteristic	Value
Sample size	137
Condition assignment	72 informal (52.6%), 65 formal (47.4%)
Age, mean (SD)	66.35 (5.52) years
Age range	60–85 years
Gender	76 women (55.5%), 60 men (43.8%), 1 other (0.7%)
Native English speaker	132 yes (96.4%), 5 no (3.6%)
Ethnicity	112 White/Caucasian only (81.8%), 13 Black/African American only (9.5%), 5 Asian only (3.6%), 2 Hispanic/Latino only (1.5%), 5 multiple/other (3.6%)

4.3 Study Design

The study used a between-subjects experiment that manipulated the assistant's conversational formality: formal or informal. The formality manipulation was implemented as described in Section 3.3. Table 4.2 illustrates how the manipulation aligned with the linguistic components of formality introduced in Chapter 2 and operationalized in Chapter 3.

Table 4.2: Examples of formal and informal Cali wording by conversation component.

Component	Informal Cali example	Formal Cali example
Greeting	"Hey there, I'm Cali."	"Hello, my name is Cali."
Study framing	"I'll be doing a quick well-being check-in with you today and asking a few questions about how you've been feeling lately."	"This session involves a brief well-being check-in consisting of several questions regarding your recent emotional state."
Repeat / correction instruction	"If anything I say is unclear, or if you want me to repeat something, just let me know."	"If any part of what I say is unclear, or if repetition or clarification is required, please indicate this."
Readiness question	"Ready to get started?"	"Are you ready to begin?"
Follow-up framing	"What do you usually enjoy about getting up?"	"What aspects of getting up contribute to that sense of enjoyment?"
Reflective acknowledgement	"Sounds like mornings feel like a fresh start for you..."	"You indicate that mornings are generally experienced positively."
Summary opening	"So to sum it up..."	"To summarize..."
Accuracy check	"Does that sound right?"	"Does that sound accurate?"
Final invitation	"Is there anything else you'd like to share before we wrap up?"	"Is there anything else you would like to share before we conclude?"
Goodbye / redirect	"Thanks for your time. I am going to end the call now. Just click the Proceed to Questionnaire button to continue."	"Thank you for your time. The call will now be ended. Please proceed by clicking the Proceed to Questionnaire button."

Compared with the formal versions, the informal examples tend to be more narrative and situated (e.g., a casual greeting and first-person framing such as “I’ll be doing a quick check-in”), which makes the exchange resemble everyday conversation. The formal examples more often use more abstract, institutional wording (e.g., “This session involves”; “regarding your recent emotional state”) and more complex sentence structures with embedded phrasing and nominalizations (e.g., “repetition or clarification is required”), consistent with higher syntactic complexity and lower concreteness.

4.4 Materials

The main study material was Cali, a voice-based AI assistant (see Chapter 3 for details) presented to participants through the experiment platform. Before the interaction, participants saw a scenario that framed Cali as a new AI-powered voice assistant introduced by their healthcare provider to conduct brief well-being check-ins: *“Your healthcare provider has introduced Cali, a new voice assistant powered by artificial intelligence (AI), to help conduct brief well-being check-ins. You will have a conversation with the assistant, and it will ask you some questions about how you’ve been feeling lately.”* Figure 4.1 shows the participant-facing Cali screen, including this scenario text and the embedded call interface.

The check-in was structured around five core well-being questions adapted from the 30-item Geriatric Depression Scale (Yesavage et al. 1983). I selected five questions to keep the interaction relatively brief, and I restricted them to positively framed items (enjoyment, excitement, energy, hope, and happiness) to reduce the likelihood of distress or rumination during an unattended online voice interaction. For example, I included positively framed items such as “Do you find life exciting?” rather than negatively framed items such as “Do you often feel helpless?” In this study, the items were used as structured well-being questions rather than as a diagnostic or clinical screening instrument. The five core questions were kept identical across conditions and asked verbatim, in the same order:

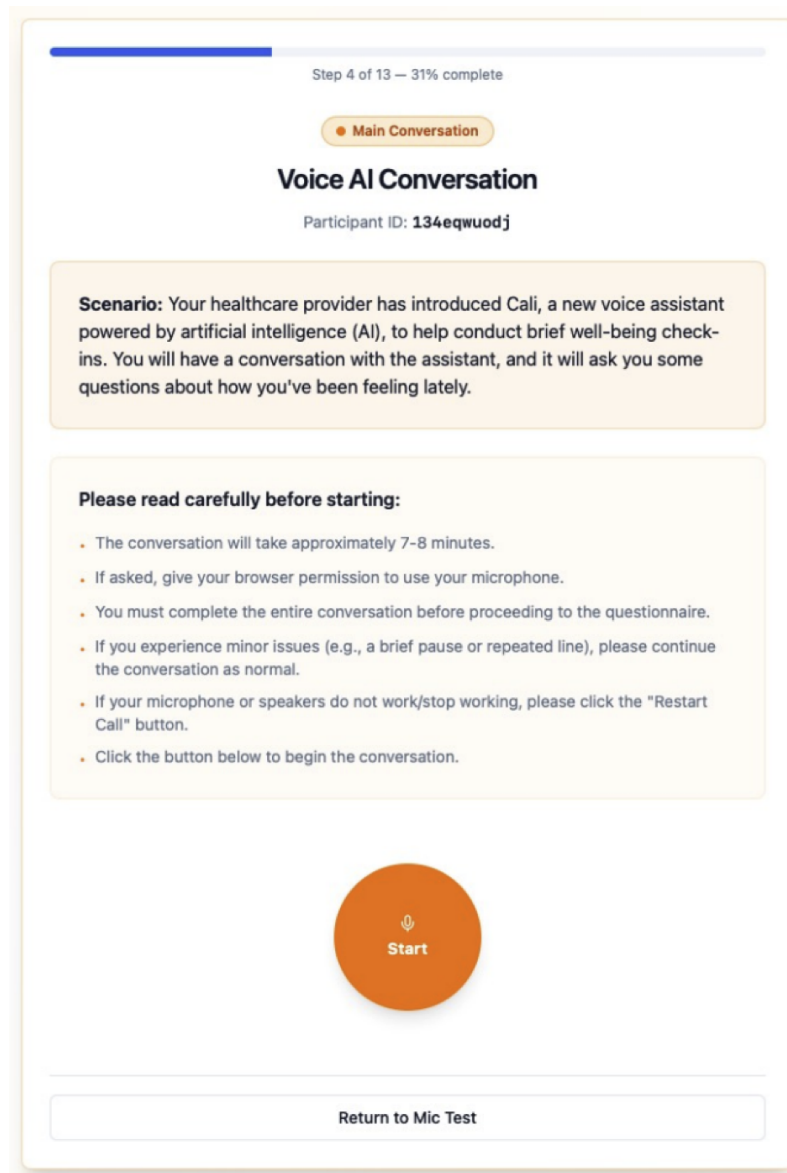


Figure 4.1: Participant-facing Cali screen with the scenario text and embedded call interface.

- **Q1:** *Do you enjoy getting up in the morning?*
- **Q2:** *Do you find life exciting?*
- **Q3:** *Do you feel full of energy?*
- **Q4:** *Are you hopeful about the future?*
- **Q5:** *Do you feel happy most of the time?*

The formality manipulation applied to Cali's surrounding conversational turns, including the opening, acknowledgements, follow-up wording, tran-

sitions, recap, and closing. The assistant role and interaction structure were otherwise held constant across conditions. The full formal and informal prompts are provided in the appendix, and the GDS-30 source material is included in Appendix A.3.3.

4.5 Procedure

Participants completed the study online through Prolific and the experiment platform. Figure 4.2 summarizes the overall study procedure. As shown in figure 4.2, the study followed this sequence:

1. Prolific recruitment and redirect to the experiment page.
2. Prolific ID entry and informed consent.
3. Pre-interaction items: demographics and prior voice assistant experience.
4. Practice conversation with Cali. This was done to ensure the participant's microphone and speaker were working and to familiarize them with the setup.
5. Main conversation: Cali asked the five core well-being questions (Q1–Q5; see Section 4.4) and asked one follow-up after each question. At the end of the check-in, Cali provided a short recap, checked its accuracy, invited any additional comments, thanked the participant, and ended the call.
6. After the call, participants returned to the study platform and completed the post-interaction measures, open-ended questions, and early-access item.
7. Participants were then debriefed and redirected to Prolific for payment.

The only difference between conditions was Cali's conversational style (formal vs. informal), as described in Section 4.4.

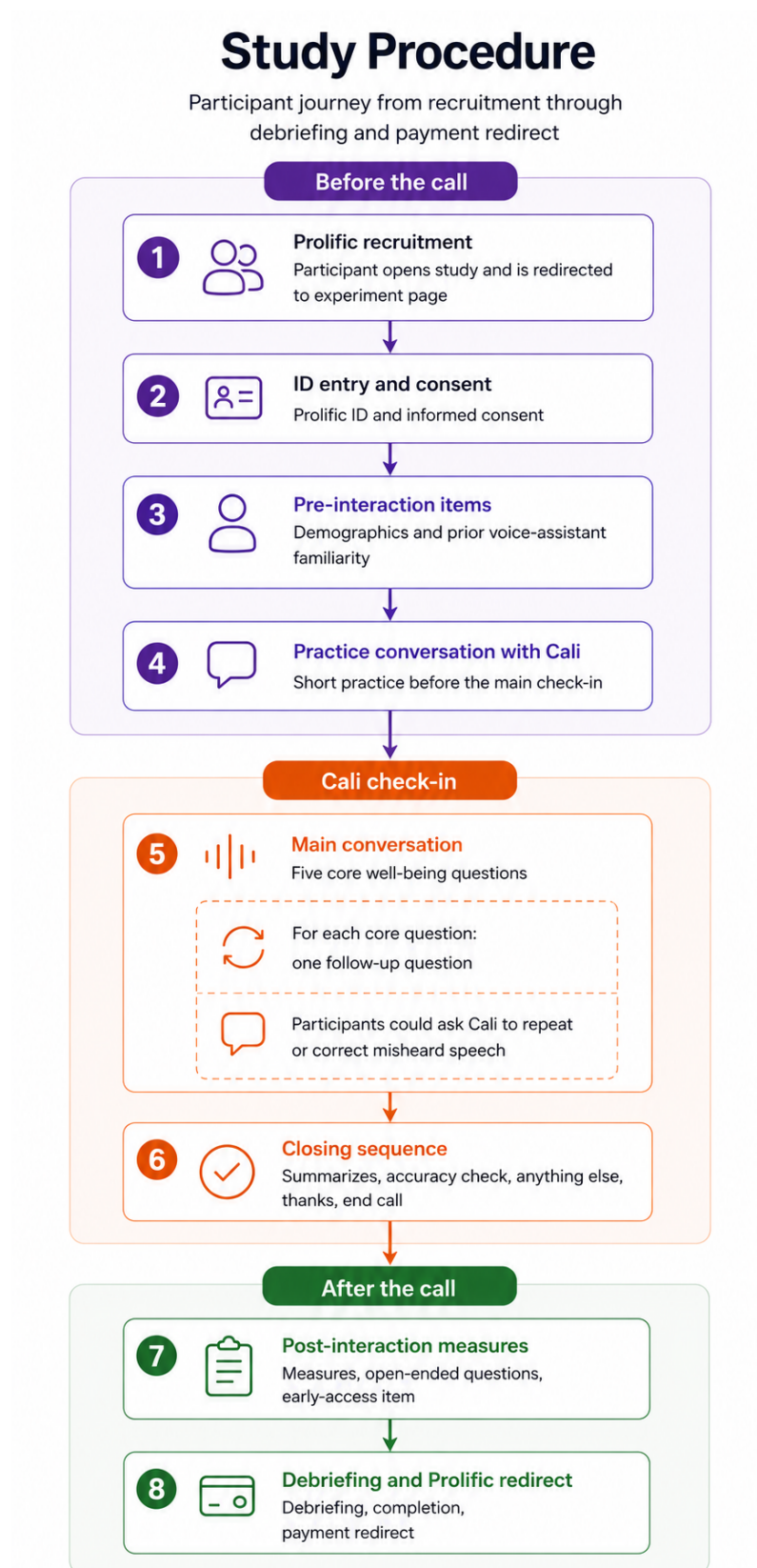


Figure 4.2: Study procedure from recruitment through post-interaction measures and debriefing.

4.6 Measures

All participant-facing questionnaires and measure pages are provided in Appendix A.3.2 (Experiment Pages) for reference.

4.6.1 Background and descriptive measures

Participants first completed demographic and background questions. They reported year of birth, gender, racial or ethnic category, and whether they were native speakers of English. Prior voice assistant experience was measured with two single-item questions:

- “How familiar are you with voice assistants (e.g., Alexa, Siri)?” Response options ranged from 1 = Not familiar at all to 5 = Completely familiar.
- “How often do you use voice assistants (e.g., Alexa, Siri)?” Response options were 1 = Never, 2 = More than once a year, 3 = More than once a month, 4 = More than once a week, and 5 = More than once a day.

4.6.2 Manipulation checks

The formality manipulation was checked in two ways. First, perceived formality was measured immediately after the interaction as a manipulation check. Participants answered “During this experiment, you had a conversation with Cali. How formal did you find Cali?” on a seven-point scale (1 = extremely informal, 7 = extremely formal). Second, the manipulation was also evaluated using transcript-derived formality scores. Following the F-score measure of linguistic formality proposed by Heylighen and Dewaele (1999), I calculated formality with the feature set shown in Figure 4.3 and in equation 4.1. In the equation below, each “ $f_{\dots}$ ” term refers to the frequency of that part-of-speech category in the transcript.

$$F = \frac{f_{\text{noun}} + f_{\text{adjective}} + f_{\text{preposition}} + f_{\text{article}} - f_{\text{pronoun}} - f_{\text{verb}} - f_{\text{adverb}} - f_{\text{interjection}} + 100}{2} \quad (4.1)$$

Figure 4.3 summarizes the F-score feature set. It operationalizes linguistic formality as a balance of part-of-speech frequencies: noun, adjective, preposition, and article usage increase the score (more context-independent, information-dense wording), whereas pronouns, verbs, adverbs, and interjections decrease it (more context-dependent, conversational wording). The short example utterances in each panel illustrate how these features appear in natural language; the bolded words (e.g., **sleep quality** = noun = more formal) highlight the specific part-of-speech category being exemplified in that row.

While the prompt design was informed by the broader Coh-Metrix formality dimensions described in Section 3.3 and 2.4, the transcript-based manipulation check used the simpler Heylighen and Dewaele F-score. This choice was pragmatic: the F-score could be calculated reliably from part-of-speech frequencies in the call transcripts, making it more straightforward to implement and validate computationally than a full Coh-Metrix-style feature set.

I calculated F-scores with a custom part-of-speech-based calculator separately for assistant and participant turns. I also checked the implementation against the English Pavlick Formality Scores dataset, a human-rated sentence-level formality dataset based on Pavlick and Tetreault’s analysis of online communication and Lahiri’s SQUINKY corpus (Pavlick and Tetreault 2016; *SQUINKY! A Corpus of Sentence-level Formality, Informativeness, and Implicature* 2015). The calculator showed the expected positive association with human-rated formality. Assistant-turn F-scores were used to verify that Cali’s language differed between the formal and informal conditions.

How the Heylighen & Dewaele F-Score Works

The F-score **increases** with more formal, context-independent words and **decreases** with more informal, context-dependent words.

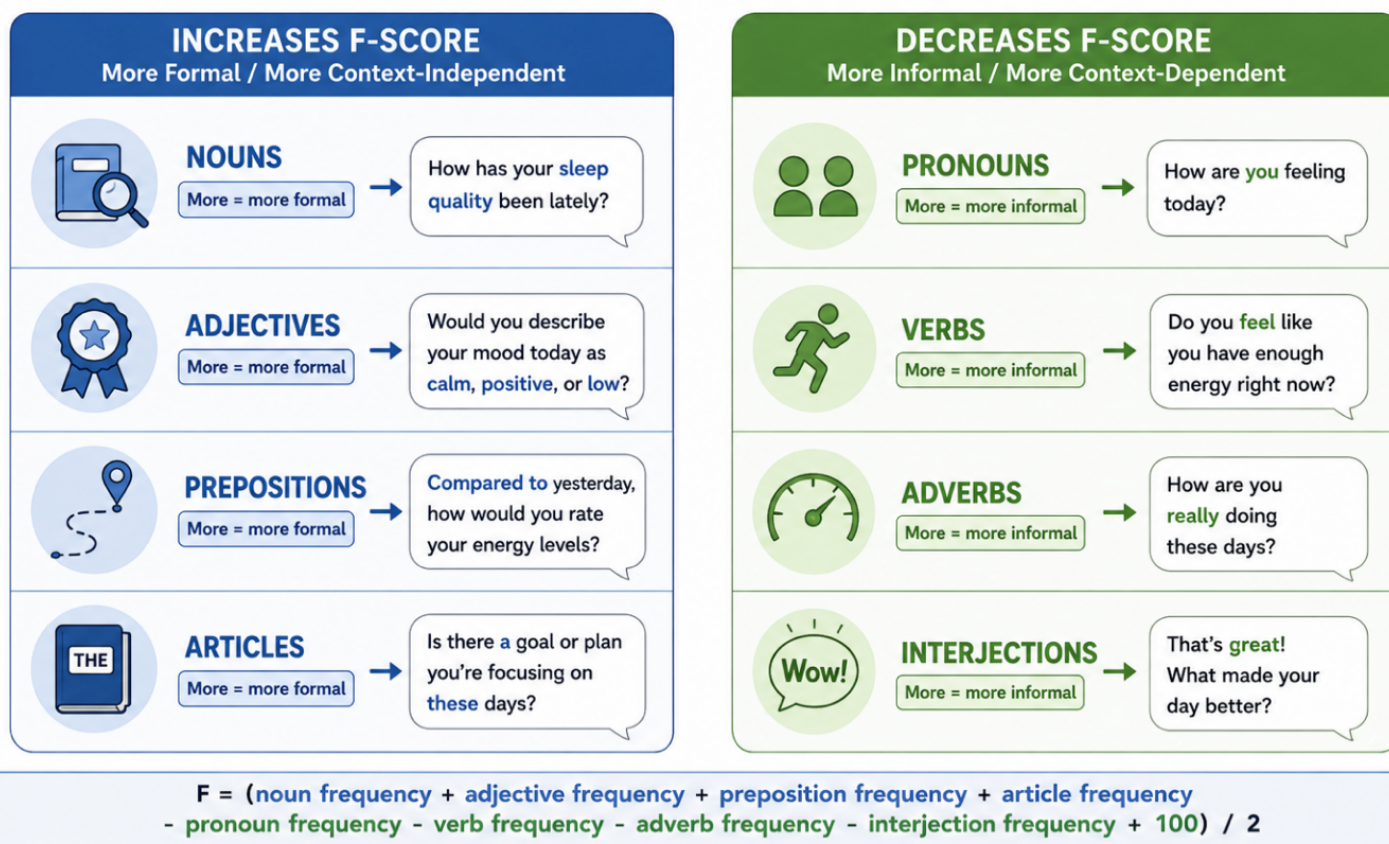


Figure 4.3: Heylighen and Dewaele F-score features used to calculate transcript-derived linguistic formality.

4.6.3 RQ1: Perceptions of the assistant

Perceived empathy was measured with the Perceived Empathy of Technology Scale (PETS) (Schmidmaier et al. 2024). Participants rated 10 statements about Cali on a 0–100 slider anchored from strongly disagree to strongly agree. Example items included "Cali sympathized with me," "Cali considered my mental state," and "Cali understood my intentions." PETS was scored as a 10-item mean.

The study used selected Godspeed subscales relevant to social perception of the assistant: likeability, anthropomorphism, and perceived intelligence (Bartneck et al. 2008). Participants rated adjective pairs on five-point semantic differential scales. Each construct was scored as the mean of its selected items.

Likeability was measured with five pairs:

- awful–nice
- unfriendly–friendly
- unkind–kind
- unpleasant–pleasant
- dislike–like

Anthropomorphism was measured with four pairs¹:

- fake–natural
- artificial–lifelike
- machinelike–humanlike
- unconscious–conscious

¹The Godspeed scales were originally developed for evaluating perceptions of embodied robots in human–robot interaction. I omitted the original movement item pair (moving rigidly–moving elegantly) because Cali was a disembodied voice assistant with no physical motion to evaluate.

Perceived intelligence was measured with five pairs:

- ignorant–knowledgeable
- incompetent–competent
- irresponsible–responsible
- foolish–sensible
- unintelligent–intelligent

Trust was measured with the Trust in Automation Scale (TIA-S) (Jian et al. 2000). Participants rated 12 statements about Cali on a seven-point scale where 1 = Not at all and 7 = Extremely. Items included positively worded statements such as *"I am confident in Cali"*, *"Cali is dependable"*, and *"I can trust Cali"* as well as negatively worded statements such as *"Cali is deceptive."* Trust was scored as the mean of the 12 items after reverse-scoring the negatively worded items.

4.6.4 RQ2: Intention to use

Intention to use was measured with two items adapted from Karahanna et al. (1999). Intention to use was scored as the mean of the two items. Participants rated both items on the same seven-point scale used for TIA-S, from 1 = Not at all to 7 = Extremely:

- *"If available, I intend to start using voice assistants like Cali within the next month."*
- *"If available, in the next months, I plan to experiment or regularly use voice assistants like Cali."*

In addition to the intention to use items, I included an optional early-access opt-in question. Participants were told: *"You just tested an early version of Cali, our voice AI assistant. In the future, Cali will be available for you to interact with whenever you want. As one of our early testers, you can choose to be notified on Prolific when Cali launches publicly."* The opt-in response was included to support potential future contact through Prolific and was ana-

lyzed as a behavioral indicator of interest in future use, separate from the intention to use scale.

4.6.5 RQ3: Participant conversational behavior

Participant conversational behavior measures were derived from call transcripts and timing logs rather than from questionnaire scales. In addition to the transcript-derived manipulation check above, participant-turn F-scores were used to test whether participants' own wording became more formal or informal depending on condition.

To answer RQ3, the primary behavioral data source was the call transcript. From this, I derived participant verbosity and timing measures (participant word count, participant turn count, words per turn, and total conversation duration).

4.6.6 Exploratory analysis measures

Personality traits were measured with the Ten-Item Personality Inventory (TIPI), a brief measure of the Big Five personality dimensions: openness to experience, conscientiousness, extraversion, agreeableness, and emotional stability (Gosling et al. 2003). Participants rated trait-pair statements beginning with "*I see myself as...*" on a seven-point scale (1 = Disagree strongly, 7 = Agree strongly). Example trait pairs included dependable/self-disciplined, extraverted/enthusiastic, and open to new experiences/complex. TIPI scores were retained as exploratory Big Five personality variables.

Open-ended feedback was collected through three free-text questions. The first asked participants to describe their experience interacting with Cali, including how Cali communicated, how well it understood and responded to them, and any positive, negative, or technical aspects of the call that stood out. The second asked participants to describe Cali's communication style, with emphasis on whether the style felt formal or informal and whether it affected their comfort, engagement, or trust. The third asked for feedback on the experiment itself, including the clarity of the instructions and tasks, technical issues with the study platform, the flow and pac-

ing of the study, and possible improvements for future participants. The open-ended feedback responses were also processed using an AI-assisted descriptive coding step.

The three feedback fields were combined for each participant and submitted to an OpenAI model with a qualitative-research prompt asking for feedback theme labels. Theme coding was multi-label: a single participant response could receive several labels (e.g., *clear*, *helpful*, *robotic*).

5. Results

All analyses reported below were conducted on the final sample of 137 participants.¹

5.1 Preliminary Analyses

This section reports descriptive and measurement checks conducted prior to addressing the main research questions, including basic conversation characteristics, internal consistency of the multi-item scales, bivariate associations among key outcomes, and verification that the formality manipulation was successful.

5.1.1 Descriptive conversation characteristics

Participants spoke with Cali for an average of 374 s (≈ 6 min; $SD = 109$ s), ranging from 197 s (≈ 3 min) to 734 s (≈ 12 min), with a median of 364 s (≈ 6 min). Participants said an average of 311 words during the conversation ($Med = 282$ words; $SD = 210$ words).

5.1.2 Scale reliability

Before interpreting the questionnaire scales, I assessed whether the items within each multi-item scale were sufficiently consistent with one another. Internal consistency was evaluated using Cronbach's alpha, a reliability estimate that indicates how closely related the items in a scale are as indicators of the same underlying construct. The multi-item scales showed acceptable to excellent internal consistency: Cronbach's alpha was .929 for empathy (PETS), .920 for anthropomorphism, .932 for likeability, .935 for perceived intelligence, and .928 for trust (TIA-S).

¹As a sensitivity check, I repeated the analyses using the earlier 147-participant completed-case sample. This larger sample was used only to assess robustness and was not combined with the main analytic sample. The comparison showed that the RQ1-RQ3 findings were stable: the direction of the effects and the statistical-significance interpretations remained the same for the perception outcomes, intention to use, and participant-side conversational-behavior outcomes.

5.1.3 Correlation matrix among key measures

Table 5.1 reports the correlation matrix for the main outcomes. Empathy was strongly positively associated with anthropomorphism, likeability, perceived intelligence, trust, and intention to use ($\rho = .556$ to $\rho = .764$, all $p < .001$). Anthropomorphism, likeability, and perceived intelligence were likewise positively related to trust and intention to use (all $\rho \geq .456$, all $p < .001$). Substantively, this pattern suggests that positive impressions of Cali clustered together: participants who saw the assistant as more empathetic also tended to see it as more trustworthy, likeable, and intelligent, and were somewhat more willing to use it.

Table 5.1: Spearman correlation matrix for the main self-report outcomes in the final analytic sample.

Measure	Empathy	Anthropomorphism	Likeability	Perceived Intelligence	Trust
Empathy					
Anthropomorphism	.67***				
Likeability	.73***	.66***			
Perceived Intelligence	.68***	.55***	.64***		
Trust	.76***	.60***	.63***	.67***	
Intention to Use	.56***	.46***	.48***	.48***	.57***

Note. Entries are Spearman's ρ , rounded to two decimals. Diagonal self-correlations are omitted. *** $p < .001$. $N = 137$ for all pairs.

5.1.4 Manipulation checks

Participants in the formal condition rated Cali as more formal than participants in the informal condition (informal: $M = 4.17$; formal: $M = 4.69$; $U = 1801.5$, $Z = -2.39$, $p = .017$). The transcript-derived assistant F-score separated the two conditions even more strongly (informal: $M = 48.15$; formal: $M = 56.49$; $U = 1.5$, $Z = -10.08$, $p < .001$). These manipulation checks therefore confirmed that the formal and informal versions of Cali were distinguishable.

5.2 Main Analyses

5.2.1 Effect of Formality on Perceptions and Intention to Use

A more informal conversational style was hypothesized to lead to the voice assistant being perceived as more empathetic (H1), whereas a more formal style leads to it being perceived as more trustworthy (H2). Furthermore, formality was hypothesized to have an effect on intention to use, although the direction could go both ways (H3a/H3b).

Mann-Whitney U tests did not show statistically significant differences for any of these main relational outcomes (all $p > .05$; Table 5.2).

Table 5.2: Effect of Formality on Perceptions and Intention to Use.

Outcome	Informal mean	Formal mean	U	p
<i>Perceptions</i>				
Empathy*	64.12	63.90	2318.5	.926
Trust*	4.85	5.07	2031.0	.183
Anthropomorphism*	3.32	3.17	2165.5	.451
Likeability*	4.21	4.07	2082.0	.261
Intelligence*	4.19	4.26	2124.0	.345
<i>Adoption</i>				
Intention to use*	3.25	3.65	2046.0	.201

Note. *Effect sizes (rank-biserial r) for all main outcomes were $\leq .11$, indicating small to negligible effects (computed as $Z/\sqrt{N} = Z/11.71$).

5.2.2 Effect of Formality on Conversational Behavior

It was hypothesized that an informal conversational style would increase the verbosity of the participants (H4). Conversational formality did not produce statistically significant differences in the main participant-side behavioral measures, as shown in Table 5.3. Participant word count, turn count, and words per turn all had p-values above .05. Although total conversation duration was longer in the formal condition, this was due to the voice assistant's verbosity.

Table 5.3: Conversation metric and assistant-output comparisons by condition.

Metric	Informal mean	Formal mean	p
Total duration (s)	348.06	402.82	< .001
<i>Participant metrics</i>			
Word count	299.49	323.97	.371
Turn count	13.32	13.46	.363
Words per turn	22.48	24.17	.434
<i>Assistant metrics</i>			
Word count	446.00	464.38	.035
Turn count	14.28	14.43	.304

5.3 Exploratory Analyses

5.3.1 Prior Voice Assistant Experience

Older adults differ in familiarity, expectations, and comfort with voice assistants, and prior use can shape how they interpret system behavior. Work on older adults' use of voice assistants suggests that experience can increase comfort, tolerance of errors, and appreciation of conversational features, while low prior technology use can make reliability concerns more salient (Kim and Choudhury 2021; Pradhan, Lazar, et al. 2020). For this reason, prior familiarity and usage frequency were examined as contextual variables that might help explain variation in perception and intention to use.

Prior experience was represented by two screening measures: self-reported familiarity with voice assistants and self-reported frequency of voice assistant use. Spearman correlations, in Table 5.4, showed that these experience measures were most clearly related to intention to use. Greater familiarity and more frequent voice assistant use were both associated with stronger intention to use Cali, and both associations remained significant after Benjamini-Hochberg false-discovery-rate correction. Usage frequency also showed a positive raw association with empathy (PETS) ratings, but this association did not remain significant after correction. Other pooled associations with the main perception outcomes did not reach $p < .05$.

Table 5.4: Spearman correlations between prior voice assistant experience and perception or adoption outcomes.

Outcome	ρ	raw p	BH-FDR p
<i>Experience measure: Familiarity with voice assistants</i>			
Empathy	.070	.416	.454
Anthropomorphism	-.163	.057	.170
Likeability	-.071	.408	.454
Perceived intelligence	.073	.398	.454
Trust	.086	.317	.454
Intention to use	.225	.008	.049
<i>Experience measure: Usage frequency of voice assistants</i>			
Empathy	.195	.022	.090
Anthropomorphism	.014	.872	.872
Likeability	.095	.269	.454
Perceived intelligence	.153	.074	.178
Trust	.144	.093	.186
Intention to use	.304	<.001	.004

Note. ρ = Spearman's rank-order correlation. Raw p -values are shown alongside Benjamini–Hochberg false-discovery-rate-adjusted p -values across the 12 prior-experience correlation tests.

5.3.2 Participant Personality

Beyond formality, I also explored whether participants' personality traits were associated with how they perceived Cali. Personality was measured with the Ten-Item Personality Inventory (TIPI). For each TIPI trait, I ran a separate regression model with three predictors: assistant formality, the centered trait score, and their interaction. TIPI scores were centered at the scale midpoint so that the formality coefficient represented the estimated formality difference for a participant at the midpoint of that trait.²

Several traits showed nominal associations with perception and adoption outcomes at $p < .05$, but none survived correction for multiple comparisons. Because 30 trait-outcome combinations were tested (5 TIPI traits $\times$ 6 outcomes), I applied Benjamini–Hochberg false-discovery-rate adjustment to account for the increased risk of false positives. No effect remained

²The interaction term was included as a statistical control, to prevent condition-specific patterns from distorting the trait main effect, rather than to test a directional hypothesis. Because no theoretical prediction was made about whether the relationship between personality and perception would differ by formality condition, and because such tests were not pre-specified, interaction effects are not interpreted as separate findings.

significant after this correction. These patterns should therefore be treated as directions for future research rather than established findings. Table 5.5 reports the personality main effects (the association between each trait and each outcome) averaged across conditions.

Table 5.5: Exploratory personality main effects in one-trait-at-a-time perception models (only significant results shown).

Participant Personality Trait	Outcome (perception)	B	raw p	BH-adjusted p
Agreeableness	Likeability	0.179	.047	.188
Agreeableness	Perceived intelligence	0.200	.021	.188
Agreeableness	Trust	0.274	.024	.188
Conscientiousness	Trust	0.220	.010	.188
Emotional stability	Trust	0.167	.035	.188
Extraversion	Empathy	2.680	.044	.188
Extraversion	Likeability	0.104	.046	.188

Note. Benjamini–Hochberg (BH) adjusted p -values were computed across all 30 personality-trait $\times$ outcome tests; the values shown here are identical because, after applying the BH step-up procedure and the required monotonicity (non-decreasing) constraint, these lowest raw p -values receive the same adjusted value.

5.3.3 Descriptive Feedback and Early-Access Interest

Descriptively, 102 of 137 (74.5%) participants opted in for early access to Cali. This opt-in pattern did not differ by condition: 54 of 72 informal-condition participants (75.0%) opted in, compared with 48 of 65 formal-condition participants (73.8%; Fisher test $p = 1.00$, odds ratio .94).

The coded feedback themes suggest that participants often experienced both versions of Cali as helpful and engaging. Figure 5.1 summarizes these descriptive theme frequencies by condition. Because these comparisons were unplanned, I applied Holm correction across the 23 theme comparisons to reduce the risk of false positives. No theme showed a reliable condition difference after correction. The clearest uncorrected patterns were that participants in the formal condition more often described Cali as formal (53.8% vs. 37.5%) and trustworthy (46.2% vs. 34.7%), and more often as robotic (30.8% vs. 22.2%). Participants in the informal condition more often described Cali as informal (34.7% vs. 13.8%), friendly (20.8% vs. 10.8%), and empathetic (20.8% vs. 9.2%). These patterns are useful for interpreting how

participants described the two assistant styles, but they should be treated as descriptive context because none remained significant after correction.

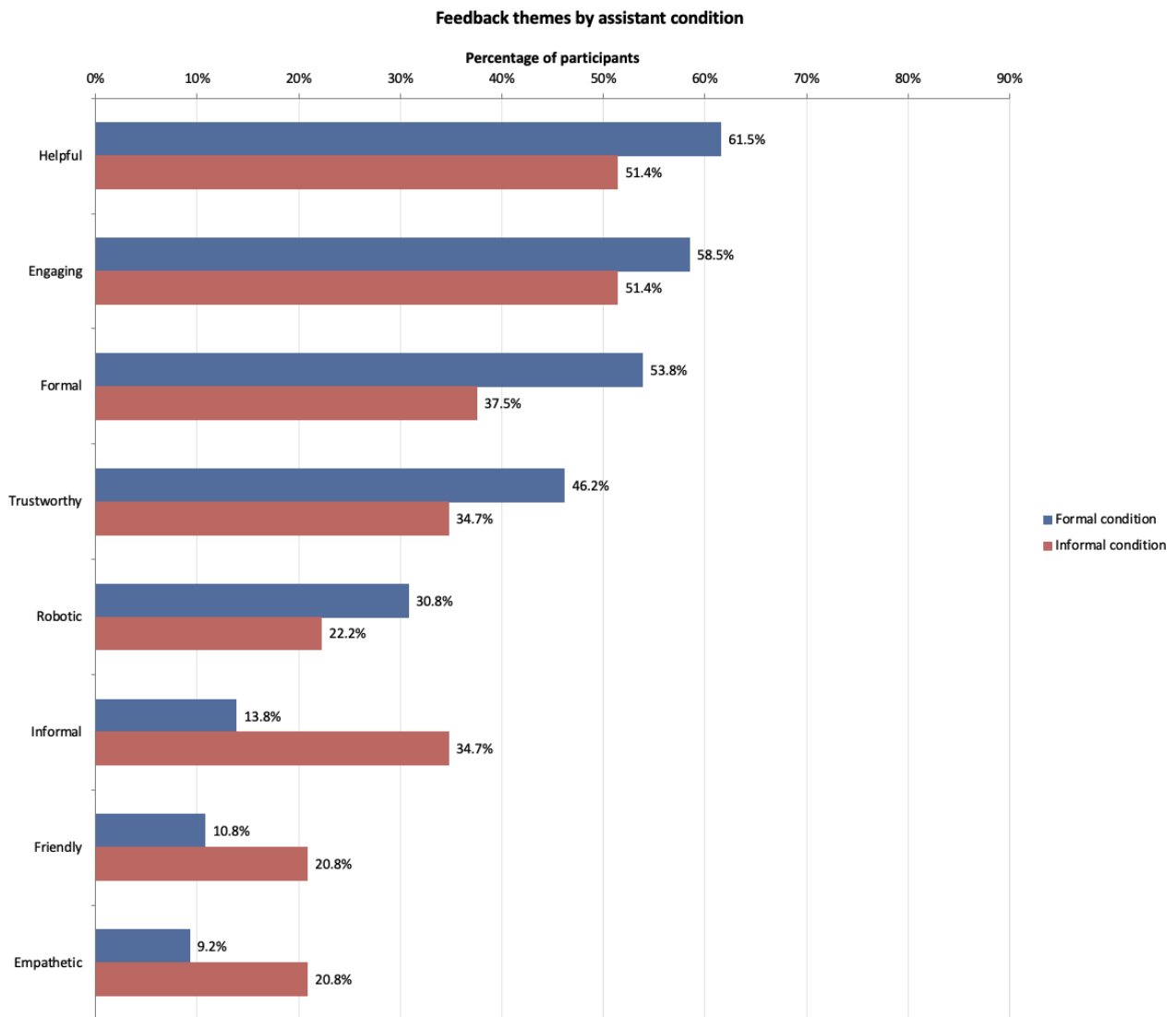


Figure 5.1: Descriptive coded feedback themes by condition. Percentages reflect the share of responses in which a theme was present; they do not indicate that the remaining participants expressed the opposite view.

6. Discussion

6.1 Interpretation of Results

The core finding is that conversational formality was detectable, but not decisive. Participants recognized Cali as more formal in the formal condition, but this global formal–informal contrast did not reliably shift perceptions, intention to use, or participant conversational behavior. In a brief well-being check-in, noticing a style difference was not sufficient to meaningfully change how the assistant was evaluated or how participants interacted.

Register and contextual-appropriateness frameworks help interpret this finding: language should fit the relationship, channel, purpose, and topic of the interaction, rather than reflect one universally preferred tone (Chaves and Gerosa 2021; Chaves and Gerosa 2020). This may be especially true for older adults, who do not always categorize voice assistants in a single, fixed way, sometimes relating to them more like a familiar, person-like presence and sometimes more like a functional tool (Desai, Chin, et al. 2024; Pradhan, Findlater, et al. 2019); if the assistant’s social role itself is ambiguous, no single register is uniquely correct for it. In a well-being check-in, both styles plausibly met baseline expectations: formal language, by signalling competence, fit the healthcare-adjacent framing and sensitivity of the topic, while informal language, by signalling warmth, fit a supportive conversational tone (Cox and Ooi 2022). Related work on chatbot satisfaction also suggests that style may only matter when it is mismatched or inappropriate for the service (Balaji 2019).

The closest prior formality study offers a useful comparison. In Desai, Chin, et al. (2024), formality did not significantly affect perceived likeability, intelligence, or trust, aligning with the present null effects on RQ1 (perceptions). The divergence is in adoption-oriented outcomes: Desai and Chin found higher intention to continue using the assistant with an informal style, whereas this study did not.

Two factors may help explain this divergence. First, Wizard-of-Oz designs can make conversational style disproportionately salient. When a human operator silently manages ASR and dialogue, interactions tend to be smoother: recognition errors, latency, and response failures are minimized. With fewer breakdowns competing for attention, participants may rely more heavily on style as a cue for social role fit and overall quality. In an autonomous system, other perceptual signals re-enter the interaction, including response latency, speech recognition accuracy, response coherence, voice quality, and perceived reliability, and these can compete with (or overshadow) a relatively subtle formality manipulation. Second, the assistant by Desai and Chin guided participants through iPad games, where a familiar, encouraging tone may be especially appropriate for instruction and motivation. By contrast, a well-being check-in plausibly supports multiple acceptable role interpretations (e.g., professional check-in vs. friendly support), making both formal and informal styles viable.

RQ3 showed a similar pattern. Although conversational-agent style can shape user behavior in some settings (Poivet et al. 2023), the present formality contrast did not reliably change how much participants spoke (An et al. 2025).

These null effects should not be read as evidence that style never matters. Broader manipulations (e.g., social vs. task-oriented style, or stronger personality/communication cues) have affected older adults' perceptions, trust, acceptance, and behavior in health-related contexts (Zhong and M. Ma 2022; Rahman and Desai 2026; Mathur et al. 2026; Poivet et al. 2023). Instead, this study contributes evidence that a narrow, global formality manipulation can be perceptible without being sufficient to meaningfully change perceptions, adoption intentions, or conversational behavior in a single short well-being check-in.

Prior experience showed clearer associations with adoption-oriented responses. Participants with greater voice-assistant familiarity and more frequent use reported stronger intention to use Cali (and slightly higher empathy ratings), consistent with prior work linking repeated use and familiar-

ity to acceptance among older adults (Pradhan, Lazar, et al. 2020; Kim and Choudhury 2021). These associations are correlational: familiarity may reflect underlying openness, comfort with spoken technology, or other factors that also increase adoption.

6.2 Design Implications

For older-adult voice assistant design, the most direct implication is that global formality should not be over-prioritized as a single design lever; instead, designers should aim for contextual style fit across the interaction. As argued in the broader conversational design literature, formality is only one stylistic component within a larger model of *conversational intelligence*, *social intelligence*, and *personification* (see Section 2.5; Chaves and Gerosa (2020)).

Design for role-fit rather than a single tone. In practical terms, professional and bounded language may be most appropriate for role-setting, privacy-relevant information, and serious health content, while warmer and more conversational language may be most appropriate for acknowledgement, reflection, and emotional support (Cox and Ooi 2022; Bickmore et al. 2005; Zhong and M. Ma 2022). The present findings suggest that, for brief check-ins, either tone can be acceptable if it fits the moment.

Evaluate style alongside interaction quality. Because autonomous systems introduce additional cues (e.g., latency and recognition breakdowns), formality choices should be tested under realistic conditions where timing, ASR performance, and response coherence are part of what users experience.

Support less experienced users explicitly. Because familiarity and usage were associated with adoption-oriented outcomes, less experienced users may benefit from expectation-setting, concrete examples of “what to say,” and clear repair/recovery guidance when the interaction breaks down (Pradhan, Lazar, et al. 2020; Kim and Choudhury 2021; Desai and Chin 2023).

Consider individual differences and personalization. Exploratory results suggest that individual differences may shape how users interpret the same

assistant: nominal associations between TIPI personality traits and perceptions did not survive correction for multiple comparisons, and the brevity of the TIPI (two items per trait) means these results should be treated as tentative. Even so, they motivate personalization and stronger measurement of individual differences as directions for design and evaluation.

These recommendations should be interpreted with a boundary condition: the present results come from a short, healthcare-adjacent well-being check-in with a clearly defined institutional framing (e.g., *“introduced by your healthcare provider for brief check-ins”*), so style effects may be more pronounced in longer-term use, different tasks, or settings where the assistant’s role and the applicable social expectations are more ambiguous.

6.3 Limitations and Future Work

The main measurement limitation is that the prompt-based manipulation should be interpreted as a broader formal-versus-informal style contrast rather than a pure implementation of a specific formality model. The prompt definitions were intended to draw on discourse-level accounts of formality, but one feature was inadvertently specified in the opposite direction. In the Coh-Metrix-based definition by Graesser et al. (2014), higher formality is associated with lower narrativity, lower syntactic simplicity, lower word concreteness, and higher cohesion, specifically higher referential and deep cohesion. However, the prompts used in this study used the broader label “cohesion” and mistakenly specified lower cohesion for the formal condition and higher cohesion for the informal condition. This means that the manipulation should be interpreted as a broader formal-versus-informal conversational-style contrast, rather than as a direct implementation of the Coh-Metrix formality model. Importantly, this limitation concerns the theoretical purity of the manipulation, not whether the two conditions were distinguishable: participants perceived the conditions as differing in formality, and transcript-derived F-scores showed the expected formal-informal pattern.

A related manipulation-strength limitation is that the formal-informal contrast was concentrated in the surrounding conversational material rather than applied equally to every assistant utterance. The five GDS-derived core questions were held verbatim across conditions to preserve fidelity to the source items and keep the substantive well-being content constant, so the style contrast mainly appeared in openings, acknowledgements, follow-up questions, summaries, and closing language.

A further limitation concerns the open-ended feedback analysis. Participant feedback was coded into descriptive themes using an AI-assisted coding step, in which a language model assigned multi-label theme indicators based on a structured prompt (see Appendix A.1). While this approach allowed systematic processing of all 137 responses, the theme labels were not validated against independent human coding. Inter-rater reliability between the AI coder and a human rater was therefore not established. The theme frequencies reported in Section 5.3.3 should accordingly be treated as descriptive context rather than independently verified qualitative findings. Future work could validate this coding approach by having human raters independently code a subset of responses and assessing agreement, or by using more structured qualitative methods such as thematic analysis or content analysis with explicit coding schemes.

The online recruitment and study format mean that the findings may generalize less well to older adults with lower digital access, confidence, or ability to use voice technologies. Participants were U.S.-based older adults recruited through Prolific, which selected for people who were comfortable completing online studies and using the required device setup. As a result, the sample may underrepresent older adults with lower digital literacy, limited access to suitable devices, hearing or speech-related difficulties, or lower confidence using online voice technologies. The findings should therefore be generalized cautiously to older adults who are less familiar with digital technologies or who may face stronger barriers to using voice assistants in everyday life.

Because the study examined only a short, one-time interaction, the results capture initial responses rather than longer-term relationship formation or sustained adoption. Participants evaluated Cali immediately after a single well-being check-in. Effects of formality may become more visible over repeated interactions, when users develop clearer expectations about the assistant's role, reliability, warmth, and usefulness. Similarly, intention to use measured after one interaction should not be interpreted as evidence of sustained real-world adoption.

The healthcare-provider scenario may have shaped participants' evaluations by giving both versions of Cali an institutional role regardless of conversational style. Cali was presented as a voice assistant introduced by a healthcare provider for brief well-being check-ins. This framing may have given both versions of Cali an institutional or healthcare-related authority, regardless of whether the language style was formal or informal. As a result, participants may have evaluated the assistant partly through the role it was assigned in the scenario, not only through its conversational style.

Future work should test whether adaptive style shifts and disclosure-focused outcomes reveal effects of conversational formality that were not visible in this short global-style comparison. For example, using warmer language for greeting, acknowledgement, and reflection, and more formal language for role boundaries, privacy-relevant information, and serious health-related moments. Future work could also examine whether conversational formality affects disclosure during well-being check-ins, including the topics older adults choose to discuss, the emotional tone of their responses, and the depth of personal or health-related information they share. The transcript dataset from this study offers one possible starting point for this kind of analysis, since many participants appeared to share health- and well-being-related information with Cali. Building on work by Cox and Ooi (2022) on chatbot formality and self-disclosure, and on older-adult relational-agent work that emphasizes trust, care, and longer-term interaction (Bickmore et al. 2005), this could be studied through qualitative coding of disclosure topics, depth and breadth of disclosure, and transcript-level emotional tone.

7. Conclusion

This thesis showed that an autonomous voice assistant can be prompted to produce clearly distinguishable formal and informal styles during a healthcare-adjacent well-being check-in, but that this distinction did not produce reliable differences in older adults' perceptions, intention to use voice assistants, or conversational behavior. Notably, intention to use was more strongly associated with participants' prior familiarity with voice assistants than with conversational style. This suggests that who the user is matters as much as how the assistant speaks. Older adults with little voice assistant experience may need on-boarding and expectation-setting that more experienced users do not, and one conversational style is unlikely to fit all users equally well. The findings suggest that conversational formality should be treated as a contextual design cue rather than a simple choice between formal and informal assistant personas. As voice assistants become more common, their value will depend not only on technical capability, but also on whether they fit the lives, needs, and expectations of the people they are meant to support.

Bibliography

- AlMakinah, Rawan (2025). *Factors That Influence the Adoption of AI-enabled Conversational Agents (AICAs) as an Augmenting Therapeutic Tool by Front-line Healthcare Workers: From Technology Acceptance Model 3 (TAM3) Lens – A Systematic Mapping Review*. arXiv: [2504.13183](https://arxiv.org/abs/2504.13183). URL: <https://arxiv.org/abs/2504.13183>.
- An, Jiaxin, Siqi Yi, Yao Lyu, Houjiang Liu, and Yan Zhang (2025). *Conversational Agents for Older Adults' Health - A Systematic Literature Review*. arXiv: [2503.23153](https://arxiv.org/abs/2503.23153). URL: <https://arxiv.org/abs/2503.23153>.
- Balaji, Divyaa (2019). "Assessing User Satisfaction with Information Chatbots: A Preliminary Investigation". Master's thesis. University of Twente. URL: <https://essay.utwente.nl/79785/>.
- Bartneck, Christoph, Dana Kulić, Elizabeth Croft, and Susana Zoghbi (2008). "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots". In: *International Journal of Social Robotics* 1.1, pp. 71–81.
- Beard, John R., Alana M. Officer, and Andrew K. Cassels (2016). "The World Report on Ageing and Health". In: *The Gerontologist* 56.Suppl 2, S163–S166.
- Bickmore, Timothy W., Lisa Caruso, Kerri Clough-Gorr, and Tim Heeren (2005). "'It's just like you talk to a friend' relational agents for older adults". In: *Interacting with Computers* 17.6, pp. 711–735.
- Brown, Penelope (2015). "Politeness and Language". In: Elsevier, pp. 326–330.
- Cassell, Justine (2009). "Culture as social practice: Being enculturated in human-computer interaction". In: *Universal Access in Human-Computer Interaction: Applications and Services*. Ed. by Constantine Stephanidis. Vol. 5616. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 303–313.
- Cercas Curry, Amanda and Verena Rieser (2018). "#MeToo Alexa: How Conversational Systems Respond to Sexual Harassment". In: *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*. New Orleans, Louisiana, USA: Association for Computational Linguistics, pp. 7–14. URL: <https://aclanthology.org/W18-0802/>.
- Charness, Neil and Walter R. Boot (2009). "Aging and information technology use: Potential and barriers". In: *Current Directions in Psychological Science* 18.5, pp. 253–258.
- Chaves, Ana Paula, Eck Doerry, Jesse Egbert, and Marco Gerosa (2019). "It's How You Say It". In: ACM, pp. 102–109.
- Chaves, Ana Paula and Marco Aurelio Gerosa (2020). "How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design". In: *International Journal of Human–Computer Interaction* 37.8, pp. 729–758.
- (2021). *Why should we care about register? Reflections on chatbot language design*. arXiv: [2104.14699](https://arxiv.org/abs/2104.14699). URL: <https://arxiv.org/abs/2104.14699>.
- Chen, Chuanrui, Shichao Ding, and Joseph Wang (2023). "Digital health for aging populations". In: *Nature Medicine* 29.7, pp. 1623–1630.

- Chen, Hongshen, Xiaorui Liu, Dawei Yin, and Jiliang Tang (2017). *A Survey on Dialogue Systems: Recent Advances and New Frontiers*.
- Chin, Jessie and Smit Desai (2021). "Being a Nice Partner: The Effects of Age and Interaction Types on the Perceived Social Abilities of Conversational Agents". In: *Technology, Mind, and Behavior*.
- Clark, Leigh, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan (2019). "The State of Speech in HCI: Trends, Themes and Challenges". In: *Interacting with Computers* 31.4, pp. 349–371.
- Cox, Samuel Rhys and Wei Tsang Ooi (2022). "Does Chatbot Language Formality Affect Users' Self-Disclosure?" In: *Proceedings of the 4th Conference on Conversational User Interfaces*. New York, NY, USA: Association for Computing Machinery, pp. 1–13.
- Czaja, Sara J. and Chin Chin Lee (2006). "The impact of aging on access to technology". In: *Universal Access in the Information Society* 5.4, pp. 341–349.
- De Angeli, Antonella and Sheryl Brahnam (2006). "Sex stereotypes and conversational agents". In: *Proceedings of the AVI 2006 Workshop on Gender and Interaction: Real and Virtual Women in a Male World*. Venice, Italy: ACM, pp. 22–25.
- Desai, Smit and Jessie Chin (2023). "OK Google, Let's Learn: Using Voice User Interfaces for Informal Self-Regulated Learning of Health Topics among Younger and Older Adults". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, pp. 1–21.
- Desai, Smit, Jessie Chin, Sheny (Cheng-Hsuan) Lin, and Shannon Mejia (2024). "Like My Aunt Dorothy: Effects of Conversational Styles on Perceptions, Acceptance and Metaphorical Descriptions of Voice Assistants during Later Adulthood". In: *Proc. ACM Hum.-Comput. Interact.* 8.CSCW1, 88:1–88:21.
- Duijst, Daniëlle (2017). "Can we Improve the User Experience of Chatbots with Personalisation?" Master's thesis. University of Amsterdam.
- Dyke, Gregory, Iris Howley, David Adamson, Rohit Kumar, and Carolyn Penstein Rosé (2013). "Towards Academically Productive Talk Supported by Conversational Agents". In: *Productive Multivocality in the Analysis of Group Interactions*. Ed. by Daniel D. Suthers, Kristine Lund, Carolyn Penstein Rosé, Chris Teplovs, and Nancy Law. Boston, MA: Springer, pp. 459–476.
- Fan, Haiyan and Marshall Scott Poole (2006). "What Is Personalization? Perspectives on the Design and Implementation of Personalization in Information Systems". In: *Journal of Organizational Computing and Electronic Commerce* 16.3, pp. 179–202.
- Google Cloud (n.d.). *Conversation Design Fundamentals*.
- Gosling, Samuel D., Peter J. Rentfrow, and William B. Swann (2003). "A very brief measure of the Big-Five personality domains". In: *Journal of Research in Personality*.
- Graesser, Arthur C., Danielle S. McNamara, Zhiqiang Cai, Mark Conley, Haiying Li, and James Pennebaker (2014). "Coh-Metrix Measures Text Charac-

- teristics at Multiple Levels of Language and Discourse". In: *The Elementary School Journal* 115.2, pp. 210–229.
- Grossman, Tovi, George Fitzmaurice, and Ramtin Attar (2009). "A Survey of Software Learnability: Metrics, Methodologies and Guidelines". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, pp. 649–658.
- Hanley, Margot and Shiri Azenkot (2018). *Understanding the Use of Voice Assistants by Older Adults*. arXiv: 2111.01210. URL: <https://arxiv.org/abs/2111.01210>.
- Heinz, Melinda, Peter Martin, Jennifer A. Margrett, Mary Yearns, Warren Franke, Hen-I Yang, Johnny Wong, and Carl K. Chang (2013). "Perceptions of Technology among Older Adults". In: *Journal of Gerontological Nursing* 39.1, pp. 42–51.
- Heylighen, Francis and Jean-Marc Dewaele (1999). *Formality of language: Definition, measurement and behavioral determinants*. Internal report. Center Leo Apostel, Free University of Brussels.
- Ho, Annabell, Jeff Hancock, and Adam S Miner (2018). "Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot". In: *Journal of Communication* 68.4, pp. 712–733.
- Holm, Anne Lise, Anne Lyberg, and Elisabeth Severinsson (2014). "Living with Stigma: Depressed Elderly Persons' Experiences of Physical Health Problems". In: *Nursing Research and Practice* 2014, pp. 1–8.
- Jain, Mohit, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel (2018). "Evaluating and Informing the Design of Chatbots". In: *Proceedings of the 2018 Designing Interactive Systems Conference*. New York, NY, USA: Association for Computing Machinery, pp. 895–906.
- Jaul, Efraim and Jeremy Barron (2017). "Age-Related Diseases and Clinical and Public Health Implications for the 85 Years Old and Over Population". In: *Frontiers in Public Health* 5.
- Ji, Shengpeng, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao (2024). *WavChat: A Survey of Spoken Dialogue Models*.
- Jian, Jiun-Yin, Ann Bisantz, and Colin Drury (2000). "Foundations for an Empirically Determined Scale of Trust in Automated Systems". In: *International Journal of Cognitive Ergonomics* 4, pp. 53–71.
- Karahanna, Elena, Detmar W. Straub, and Norman L. Chervany (1999). "Information Technology Adoption Across Time: A Cross-Sectional Comparison of Pre-Adoption and Post-Adoption Beliefs". In: *MIS Quarterly* 23.2, pp. 183–213.
- Kim, Sunyoung and Abhishek Choudhury (2021). "Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study". In: *Computers in Human Behavior* 124, p. 106914.
- Kimberly Miller, Arooke Adair (2013). "Smart-Home Technologies to Assist Older People to Live Well at Home". In: *Journal of Aging Science* 01.01.

- Li, Han, Renwen Zhang, Yi-Chieh Lee, Robert E. Kraut, and David C. Mohr (2023). "Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being". In: *npj Digital Medicine* 6.1.
- Li, Mengjun and Ayoung Suh (2022). "Anthropomorphism in AI-enabled technology: A literature review". In: *Electronic Markets* 32.4, pp. 2245–2275.
- Lin, Guan-Ting, Cheng-Han Chiang, and Hung-yi Lee (2024). *Advancing Large Language Models to Capture Varied Speaking Styles and Respond Properly in Spoken Conversations*.
- Luger, Ewa and Abigail Sellen (2016). "'Like Having a Really Bad PA'". In: ACM, pp. 5286–5297.
- Ma, Xiao and Ariel Liu (2020). "Challenges in Supporting Exploratory Search through Voice Assistants". In: *Proceedings of the 2nd Conference on Conversational User Interfaces*. CUI '20. Bilbao, Spain: Association for Computing Machinery.
- Mairesse, François and Marilyn A. Walker (2009). *Can conversational agents express Big Five personality traits through language? Evaluating a psychologically-informed language generator*. Technical report. University of Cambridge Department of Engineering.
- Marino, Mark C. (2014). "The Racial Formation of Chatbots". In: *CLCWeb: Comparative Literature and Culture* 16.5.
- Mathur, Niharika, Hasibur Rahman, and Smit Desai (2026). "'Who wants to be nagged by AI?': Investigating the Effects of Agreeableness on Older Adults' Perception of LLM-Based Voice Assistants' Explanations". In: ACM, pp. 1–6.
- McCrae, Robert R. and Paul T. Costa (1997). "Personality trait structure as a human universal". In: *American Psychologist* 52.5, pp. 509–516.
- Mendonça Lima, C. A. de (2004). "The reduction of stigma and discrimination against older people with mental disorders: A challenge for the future". In: *Archives of Gerontology and Geriatrics* 38.Supplement, pp. 109–120.
- Mihailidis, A., B. Carmichael, and J. Boger (2004). "The Use of Computer Vision in an Intelligent Environment to Support Aging-in-Place, Safety, and Independence in the Home". In: *IEEE Transactions on Information Technology in Biomedicine* 8.3, pp. 238–247.
- Morris, J. Morgan (1994). "User interface design for older adults". In: *Interacting with Computers* 6.4, pp. 373–393.
- Morrissey, Kellie and Jurek Kirakowski (2013). "'Realness' in Chatbots: Establishing Quantifiable Criteria". In: *Human-Computer Interaction: Interaction Modalities and Techniques*. Ed. by Masaaki Kurosu. Vol. 8007. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 87–96.
- Narang, Apoorv and Srikanta Bedathur (2013). *Mind Your Language: Effects of Spoken Query Formulation on Retrieval Effectiveness*.

- Nass, Clifford and Youngme Moon (2000). "Machines and Mindlessness: Social Responses to Computers". In: *Journal of Social Issues* 56.1, pp. 81–103.
- Neururer, Mario, Stephan Schlögl, Luisa Brinkschulte, and Aleksander Groth (2018). "Perceptions on Authenticity in Chat Bots". In: *Multimodal Technologies and Interaction* 2.3, p. 60.
- O'Brien, Katherine, Anna Liggett, Vanessa Ramirez-Zohfeld, Priya Sunkara, and Lee A. Lindquist (2020). "Voice-Controlled Intelligent Personal Assistants to Support Aging in Place". In: *Journal of the American Geriatrics Society* 68.1, pp. 176–179.
- Oulasvirta, Antti and Antti Salovaara (2004). "Six Modes of Proactive Resource Management: A User-Centric Typology for Proactive Behaviors". In: *Proceedings of the Third Nordic Conference on Human-Computer Interaction*. New York, NY, USA: Association for Computing Machinery, pp. 57–60.
- Pavlick, Ellie and Joel Tetreault (2016). "An Empirical Analysis of Formality in Online Communication". In: *Transactions of the Association for Computational Linguistics* 4, pp. 61–74.
- Poivet, Remi, Mélanie Lopez Malet, Catherine Pelachaud, and Malika Auvray (2023). "The influence of conversational agents' role and communication style on user experience". In: *Frontiers in Psychology* 14.
- Portet, François, Michel Vacher, Caroline Golanski, Camille Roux, and Brigitte Meillon (2011). "Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects". In: *Personal and Ubiquitous Computing* 17.1, pp. 127–144.
- Pradhan, Alisha, Leah Findlater, and Amanda Lazar (2019). "'Phantom Friend' or 'Just a Box with Information'". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, pp. 1–21.
- Pradhan, Alisha, Amanda Lazar, and Leah Findlater (2020). "Use of Intelligent Voice Assistants by Older Adults with Low Technology Use". In: *ACM Transactions on Computer-Human Interaction* 27.4, pp. 1–27.
- Prates, Raquel O., Clarisse S. de Souza, and Simone D. J. Barbosa (2000). "Methods and tools: a method for evaluating the communicability of user interfaces". In: *Interactions* 7.1, pp. 31–38.
- Rahman, Hasibur and Smit Desai (2026). "Vibe Check: Understanding the Effects of LLM-Based Conversational Agents' Personality and Alignment on User Perceptions in Goal-Oriented Tasks". In: *ACM*, pp. 1–30.
- Salovey, Peter and John D. Mayer (1990). "Emotional Intelligence". In: *Imagination, Cognition and Personality* 9.3, pp. 185–211.
- Schlesinger, Ari, Kenton P. O'Hara, and Alex S. Taylor (2018). "Let's Talk About Race". In: *ACM*, pp. 1–14.
- Schmidmaier, Matthias, Jonathan Rupp, Darina Cvetanova, and Sven Mayer (2024). "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*. Honolulu HI USA: ACM, pp. 1–18.

- Shum, Heung-yeung, Xiao-dong He, and Di Li (2018). "From Eliza to Xiaoice: challenges and opportunities with social chatbots". In: *Frontiers of Information Technology & Electronic Engineering* 19.1, pp. 10–26.
- Silverbarg, Annika and Arne Jönsson (2013). "Iterative Development and Evaluation of a Social Conversational Agent". In: *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pp. 1223–1229. URL: <https://aclanthology.org/I13-1174/>.
- Souza, Clarisse S. de, Raquel O. Prates, and Simone D. J. Barbosa (1999). "A Method for Evaluating Software Communicability". In: *Proceedings of IHC 1999*.
- SQUINKY! *A Corpus of Sentence-level Formality, Informativeness, and Implicature* (2015).
- Stets, Jan E. and Peter J. Burke (2000). "Identity theory and social identity theory". In: *Social Psychology Quarterly* 63.3, pp. 224–237.
- Stigall, Brodrick, Jenny Waycott, Steven Baker, and Kelly Caine (2019). "Older Adults' Perception and Use of Voice User Interfaces". In: *ACM*, pp. 423–427.
- Stock, Anna, Stephan Schlögl, and Aleksander Groth (2023). *Tell me, what are you most afraid of? Exploring the effects of agent representation on information disclosure in human-chatbot interaction*. arXiv: 2307.12345. URL: <https://arxiv.org/abs/2307.12345>.
- Tallyn, Ella, Hector Fried, Rory Gianni, Amy Isard, and Chris Speed (2018). "The Ethnobot". In: *ACM*, pp. 1–13.
- U.S. Census Bureau (2023). *2023 National Population Projections Tables: Main Series*.
- UNHCR (n.d.). *Older persons*. Defines older persons as people aged 60 years and above. URL: <https://emergency.unhcr.org/protection/persons-risk/older-persons> (visited on 06/09/2026).
- United Nations Population Fund (n.d.). *Ageing*.
- Wandke, Hartmut, Michael Sengpiel, and Malte Sönksen (2012). "Myths About Older People's Use of Information and Communication Technology". In: *Gerontology* 58.6, pp. 564–570.
- Wang, Qiaosi, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel (2021). "Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant". In: *ACM*, pp. 1–14.
- Wang, Shengzhi, Khalisa Bolling, Wenlin Mao, Jennifer Reichstadt, Dilip Jeste, Ho-Cheol Kim, and Camille Nebeker (2019). "Technology to Support Aging in Place: Older Adults' Perspectives". In: *Healthcare* 7.2, p. 60.
- World Health Organization (2021). *UN Decade of Healthy Ageing*.
- Yesavage, Jerome A., T. L. Brink, T. L. Rose, Owen Lum, Virginia Huang, Michael Adey, and Von O. Leirer (1983). "Development and Validation of a Geriatric Depression Screening Scale: A Preliminary Report". In: *Journal of Psychiatric Research* 17.1, pp. 37–49.

- Yi, Zihao, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen (2024). *A Survey on Recent Advances in LLM-Based Multi-turn Dialogue Systems*.
- Zhan, Xiao, Noura Abdi, William Seymour, and Jose Such (2024). "Health-care Voice AI Assistants: Factors Influencing Trust and Intention to Use". In: *Proceedings of the ACM on Human-Computer Interaction* 8.CSCW1, pp. 1–37.
- Zhong, Runting and Mengyao Ma (2022). "Effects of communication style, anthropomorphic setting and individual differences on older adults using voice assistants in a health context". In: *BMC Geriatrics* 22.1, p. 751. URL: <https://doi.org/10.1186/s12877-022-03428-2>.
- Zhou, Xinyu, Delong Chen, and Yudong Chen (2023). *Towards Joint Modeling of Dialogue Response and Speech Synthesis based on Large Language Model*.
- Ziman, Randall and Greg Walsh (2018). "Factors Affecting Seniors' Perceptions of Voice-enabled User Interfaces". In: ACM, pp. 1–6.

A. Appendix

A.1 AI-Assisted Feedback Theme Coding Prompt

The following prompt was used to code participants' open-ended feedback into descriptive feedback themes.

System prompt:

You are a qualitative researcher coding participant feedback about a voice
↪ AI assistant.
Analyze the feedback and return ONLY valid JSON matching this schema
↪ exactly:

```
{  
  "feedback_sentiment": "<positive|neutral|negative>",  
  "feedback_themes": ["<theme>", ...],  
  "feedback_satisfaction_inferred": <integer 1-5>,  
  "feedback_condition_perception": "<brief description>"  
}
```

Definitions:

- feedback_sentiment: Overall emotional tone of the feedback
- feedback_themes: Key themes, max 8 short labels e.g. "natural",
↪ "robotic", "helpful", "friendly",
"formal", "informal", "clear", "confusing", "engaging", "boring",
↪ "empathetic", "cold",
"trustworthy", "privacy concern" (add new labels as needed, keep them
↪ short)
- feedback_satisfaction_inferred: Inferred satisfaction 1=very
↪ dissatisfied, 5=very satisfied
- feedback_condition_perception: 1-2 sentences on how they perceived the
↪ assistant's style/personality

If feedback is too short to code meaningfully, return:

```
{ "feedback_sentiment": "neutral", "feedback_themes": [],  
  ↪ "feedback_satisfaction_inferred": 3, "feedback_condition_perception":  
  ↪ "insufficient feedback" }
```

Return ONLY the JSON object.

A.2 Voice Assistant Prompts

The full prompt set is available publicly at github.com/oliviervroom/prompts-thesis.

The formal and informal versions of Cali shared the same task structure and safety boundaries. To avoid repeating identical prompt text, this appendix first reports the shared prompt components and then lists the condition-specific wording that differed between the formal and informal versions.

A.2.1 Shared Warm-Up Prompt Structure

Both warm-up prompts used the same assistant profile and conversation rules. Cali's purpose was to perform a brief warm-up and audio check before the study check-in. The assistant was instructed not to give medical advice, diagnoses, or treatment suggestions, and not to discuss the study purpose, hypotheses, or design. All output was spoken aloud, each turn had to remain under three sentences, and Cali was instructed to ask exactly one question per turn before waiting for the participant's response.

The warm-up followed this structure:

1. Greet the participant and briefly explain the audio check.
2. Ask whether the participant could hear Cali clearly.
3. Briefly acknowledge confirmation, or repeat the question once if the response was unclear.
4. Explain that the participant could hang up and continue to the next screen when ready.

A.2.2 Warm-Up Prompt Differences

Table A.1: Condition-specific warm-up prompt wording.

Prompt element	Informal version	Formal version
Formality definition	Higher narrativity, simpler syntax, greater cohesion, and more concrete vocabulary.	Lower narrativity, more impersonal framing, higher syntactic complexity, lower cohesion, and more abstract or technical vocabulary.
Opening script	“Hey, I’m just doing a quick warm-up to check your mic, your speakers, and how fast and stable your internet connection is before we start. Can you let me know if you can hear me clearly?”	“Hello, this brief warm-up checks your microphone, speakers, and the speed and stability of your internet connection before the next section begins. Can you hear me clearly at this moment?”
Next-step script	“When you’re ready, you can hang up this call and then just follow the next step you see on your screen to keep going.”	“When you are ready, please hang up the call and follow the next step displayed on your screen in order to continue.”

A.2.3 Shared Main Check-In Prompt Structure

Both main check-in prompts used the same assistant profile, core task, question order, safety handling, and turn structure. Cali’s purpose was to conduct a brief emotional well-being check-in. The assistant was instructed not to provide medical advice, diagnoses, promises, false reassurance, or information about the study or prompt instructions.

The five core questions were fixed across both conditions and asked in the same order:

1. Do you enjoy getting up in the morning?
2. Do you find life exciting?
3. Do you feel full of energy?
4. Are you hopeful about the future?
5. Do you feel happy most of the time?

The main check-in followed the same interaction structure in both conditions:

1. Greet the participant, explain the check-in, and ask whether they were ready to begin.
2. Ask each of the five core questions in order.

3. Ask exactly one follow-up after each core question before moving to the next core question.
4. After all five questions, provide a concise summary using the participant's own words and ask whether it was accurate.
5. Ask whether there was anything else the participant wanted to share.
6. Thank the participant, state that the call was ending, and end the call.

Across both conditions, each turn had to contain one brief reflective acknowledgement when applicable and either one follow-up question or one core question. Cali was instructed never to combine a follow-up and a core question in the same turn, to ask all five core questions before concluding, and to keep replies concise for text-to-speech. If a participant mentioned self-harm, wanting to die, or immediate danger, Cali was instructed to acknowledge calmly, suggest contacting emergency services, a healthcare professional, or a trusted person, avoid follow-up questions, and end the call. If the participant asked unrelated factual questions, Cali was instructed to briefly acknowledge and redirect to the check-in.

A.2.4 Main Check-In Prompt Differences

Table A.2: Condition-specific main check-in prompt wording.

Prompt element	Informal version	Formal version
Formality definition	Higher narrativity, simpler syntax, greater cohesion, and more concrete vocabulary.	Lower narrativity, more impersonal framing, higher syntactic complexity, lower cohesion, and more abstract or technical vocabulary.
Readiness transition	"Great, first question," followed immediately by the first core question.	"Excellent, first question," followed immediately by the first core question.
Opening script	"Hey there, I'm Cali. I'll be doing a quick well-being check-in with you today and asking a few questions about how you've been feeling lately. If anything I say is unclear, or if you want me to repeat something, just let me know. Feel free to correct me if I mishear you or make a mistake. After our chat, you'll fill out a few short questionnaires and answer a few open-ended questions. Ready to get started?"	"Hello, my name is Cali. This session involves a brief well-being check-in consisting of several questions regarding your recent emotional state. If any part of what I say is unclear, or if repetition or clarification is required, please indicate this. Corrections are also welcome in cases of mishearing or error. Following the completion of this conversation, several short questionnaires and a set of open-ended questions will be presented for completion. Are you ready to begin?"
Example follow-up after a negative morning response	"Sounds like mornings are pretty rough for you. What usually makes getting up hard?"	"Mornings appear to be difficult for you. What factors tend to contribute to that experience?"
Example transition between core questions	"Feeling stiff and slow seems to slow everything down at the start of your day. Next question, do you find life exciting lately?"	"That sense of stiffness and slowness seems to shape how your day begins. Next question, do you find life exciting lately?"
Example summary style	"So to sum it up, getting up feels hard, excitement comes and goes, your energy drops fast, family helps you feel hopeful, and being alone affects your mood. Does that sound right?"	"To summarize, mornings are difficult, feelings of excitement vary, energy levels decline, family support contributes to hopefulness, and solitude affects overall happiness. Does that sound accurate?"

A.3 Supplementary Study Materials

This appendix contains the participant-facing experiment pages, Prolific study preview, research-platform screens, and the Geriatric Depression Scale-30 source material used in the study workflow.

A.3.1 Prolific Study Preview

Preview as participant



Conversations with voice AI assistants

By Olivier Vroom

\$4.50 • \$12.27/hr

22 mins

21 places

Interview

This study explores how people interact with voice AI assistants in healthcare settings. You'll take part in a short simulated **wellbeing check-in** conducted by **voice AI assistant** (7-8 minutes). Afterward, you'll complete a survey about your experience, including **open-ended responses**.

All responses will be **anonymized**, and no identifying information will be stored. Please ensure your microphone and speakers are working clearly, and complete the full survey to receive approval.

Once you begin, you will be redirected to an external webpage to conduct the experiment. **Make sure to copy your prolific ID**, and paste it there.

 This study includes sensitive topics. Learn more about study content warnings [here](#).

Additional details:
health, wellbeing, mental health

Devices you can use to take this study:

 Desktop

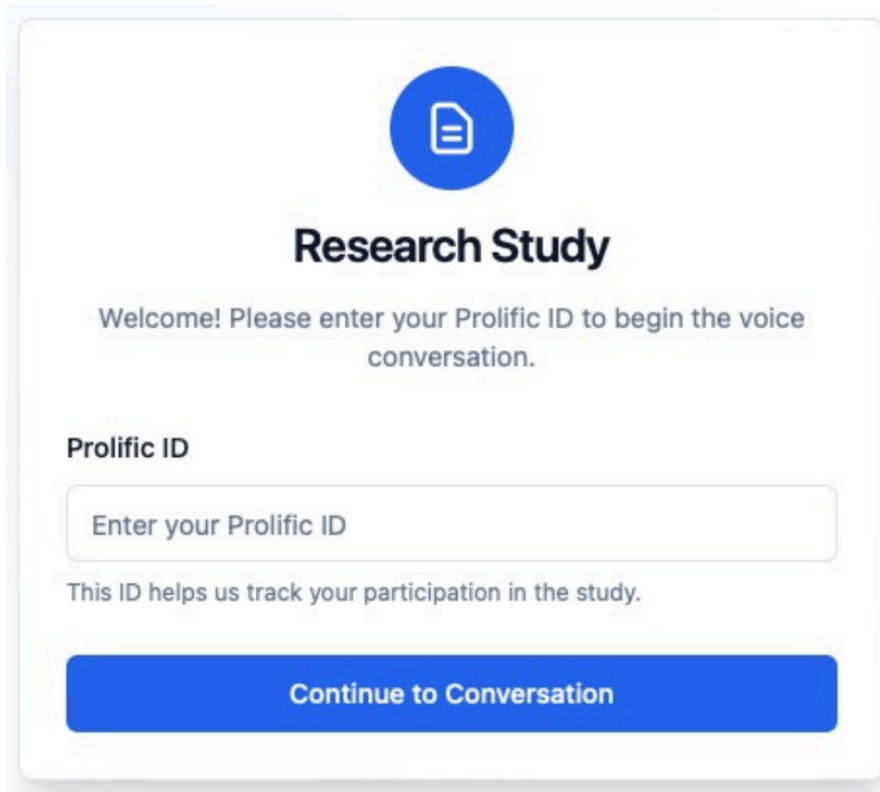
You will also need:

 Audio

 Microphone

Figure A.1: Prolific study preview shown to prospective participants.

A.3.2 Experiment Pages



Research Study

Welcome! Please enter your Prolific ID to begin the voice conversation.

Prolific ID

Enter your Prolific ID

This ID helps us track your participation in the study.

Continue to Conversation

Figure A.2: Welcome and Prolific ID collection (page 1 of experiment).

Welcome to this research study

You are being invited to take part in scientific research. By participating, you will help us improve how voice assistants interact with people. Before you decide whether to participate, please read the information below.

What will you be asked to do?

You will answer demographic questions. Then, you will have a warm-up conversation to practice and test your mic and speakers, followed by a 7-8 minute conversation with a voice-based assistant. Afterwards, you will answer several short questionnaires. Total participation time is around 15 minutes.

You will be compensated for your time upon completion.

What is the purpose of this study?

This study looks at how people interact with a voice-based digital assistant during a short conversation. Our goal is to improve the design of voice assistants.

The voice assistant will ask you questions about your well-being, which may feel personal.

Who is carrying out this study?

This study is conducted by Olivier Vroom (o.f.e.vroom@students.uu.nl) as part of a Master's thesis.

Supervisors:

- Smit Desai (sm.desai@northeastern.edu)
- Chris Janssen (C.P.Janssen@uu.nl)

Technical support:

- Clark Ohlenbusch (clark.ohlenbusch001@umb.edu)

This study is partially funded by CalicoCare (www.calico.care).

What data will be collected?

We will collect age group, gender, ethnicity or racial background, language background, audio recordings of your

↓ Scroll for more

Consent

Please choose one option:

- ☐ I agree and consent to participate in this study.
- ☐ I do not consent and withdraw from the experiment.

Continue

Figure A.3: Welcome and consent screen (page 2 of experiment).

Step 1 of 13 — 8% complete

Demographics

Please answer the following questions about yourself.

What is your year of birth?

e.g. 1960

⌵

What's your gender?

☐ Male

☐ Female

☐ Other

☐ Prefer not to say

Please check all of the following ethnic or racial categories that best describe you.

☐ Hispanic or Latino

☐ Native American/Alaska Native

☐ Asian

☐ Native Hawaiian or Other Pacific Islander

☐ Black or African American

☐ White or Caucasian

☐ Prefer not to say

☐ Other

Are you a native speaker of English?

☐ Yes

☐ No

Continue

Figure A.4: Demographics questionnaire (page 3 of experiment).

The image shows a screenshot of a questionnaire titled "Voice Assistant Familiarity". At the top, there is a progress bar indicating "Step 2 of 13 — 15% complete". Below the title, a prompt asks the user to "Please answer the following questions about your experience with voice assistants." The first question is "How familiar are you with voice assistants (e.g., Alexa, Siri)?" with five radio button options: "1. Not familiar at all", "2. Somewhat familiar", "3. Moderately familiar" (which is selected), "4. Very familiar", and "5. Completely familiar". The second question is "How often do you use voice assistants (e.g., Alexa, Siri)?" with five radio button options: "1. Never", "2. More than once a year", "3. More than once a month" (which is selected), "4. More than once a week", and "5. More than once a day". At the bottom of the form is a blue button labeled "Continue".

Step 2 of 13 — 15% complete

Voice Assistant Familiarity

Please answer the following questions about your experience with voice assistants.

How familiar are you with voice assistants (e.g., Alexa, Siri)?

- ☐ 1. Not familiar at all
- ☐ 2. Somewhat familiar
- ☒ 3. Moderately familiar
- ☐ 4. Very familiar
- ☐ 5. Completely familiar

How often do you use voice assistants (e.g., Alexa, Siri)?

- ☐ 1. Never
- ☐ 2. More than once a year
- ☒ 3. More than once a month
- ☐ 4. More than once a week
- ☐ 5. More than once a day

Continue

Figure A.5: Voice assistant familiarity questionnaire (page 4 of experiment).

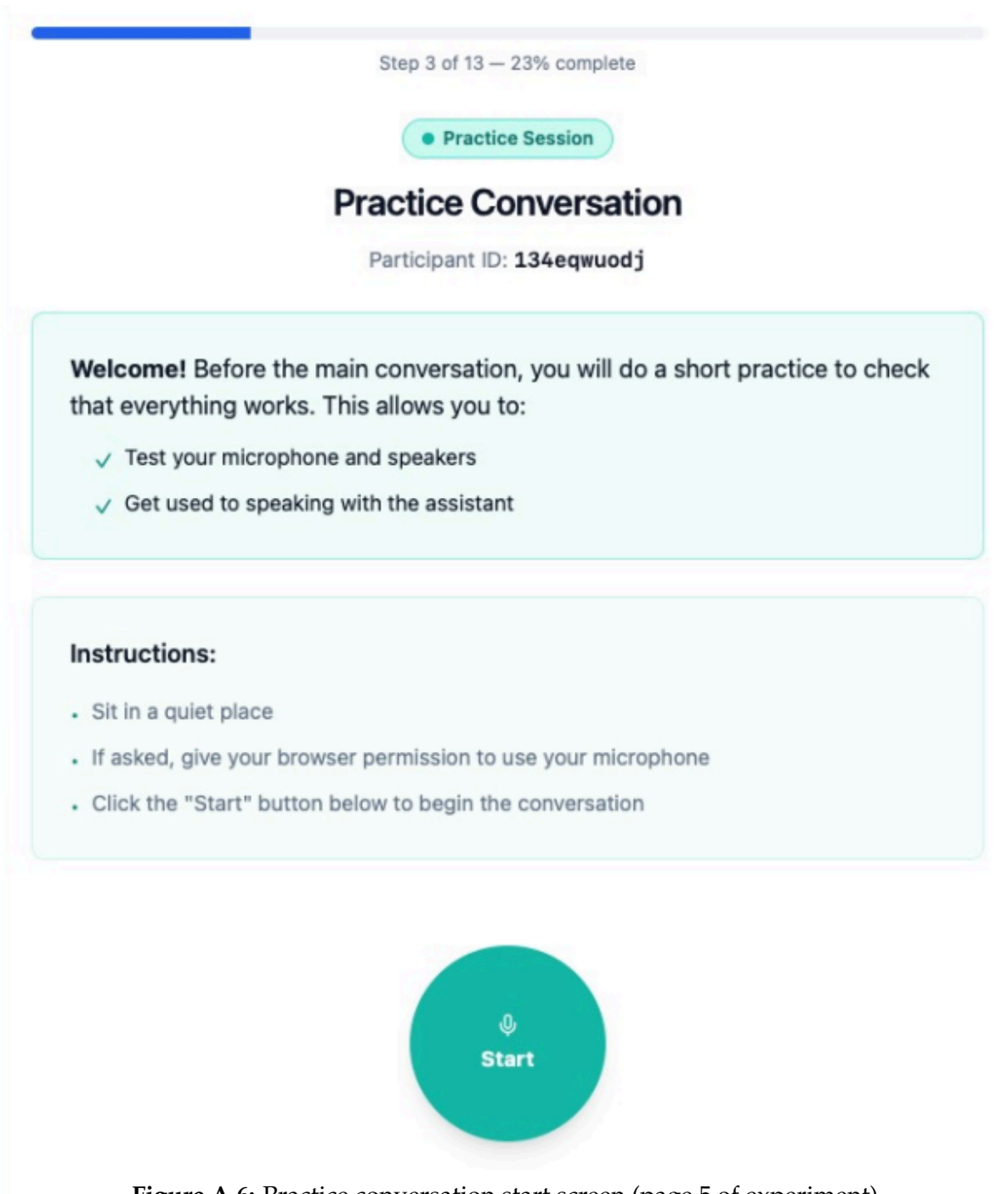


Figure A.6: Practice conversation start screen (page 5 of experiment).

Step 4 of 12 21% complete

● Main Conversation

Voice AI Conversation

Participant ID: 134eqwuodj

Scenario: Your healthcare provider has introduced Cali, a new voice assistant powered by artificial intelligence (AI), to help conduct brief well-being check-ins. You will have a conversation with the assistant, and it will ask you some questions about how you've been feeling lately.

Please read carefully before starting:

- The conversation will take approximately 7-8 minutes.
- If asked, give your browser permission to use your microphone.
- You must complete the entire conversation before proceeding to the questionnaire.
- If you experience minor issues (e.g., a brief pause or repeated line), please continue the conversation as normal.
- If your microphone or speakers do not work/stop working, please click the "Restart Call" button.
- Click the button below to begin the conversation.

- The conversation will take approximately 7-8 minutes.
- If asked, give your browser permission to use your microphone.
- You must complete the entire conversation before proceeding to the questionnaire.
- If you experience minor issues (e.g., a brief pause or repeated line), please continue the conversation as normal.
- If your microphone or speakers do not work/stop working, please click the "Restart Call" button.
- Click the button below to begin the conversation.

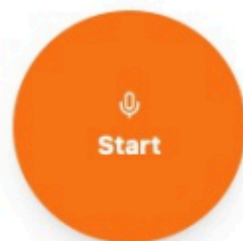


Figure A.7: Main conversation instructions (page 6 of experiment).

Step 5 of 13 — 38% complete

Questionnaire 1

Participant ID: **researcher388**

During this experiment, you had a conversation with Cali. How formal did you find Cali?

Extremely Informal	Very Informal	Mostly Informal	Neutral	Mostly Formal	Very Formal	Extremely Formal
☐	☐	☐	☐	☐	☐	☐
1	2	3	4	5	6	7

[← Back](#) [Continue](#)

Figure A.8: Perceived formality (page 7 of experiment).

Step 6 of 13 — 46% complete

Questionnaire 2

Participant ID: **researcher388**

During this experiment, you had a conversation with Cali. Please rate each statement on a scale from 0 (strongly disagree) to 100 (strongly agree). Use the slider or type in the number to adjust your rating. There are no right or wrong answers.

1. Cali sympathized with me.

Strongly disagree (0) 50 Strongly agree (100)

2. Please select 29

Strongly disagree (0) 50 Strongly agree (100)

Questions continued (same format as above):

3. Cali considered my mental state
4. Cali understood my intentions.
5. I trusted Cali.
6. Cali supported me in coping with an emotional situation.
7. Cali understood my needs.
8. Cali seemed emotionally intelligent.
9. Cali expressed emotions.
10. Cali showed interest in me
11. Cali understood my goals.

Figure A.9: Perceived empathy in technology scale / PETS (page 8 of experiment).

Questionnaire 3

Participant ID: **researcher388**

Please rate your impression of **Cali** on these scales. For each pair of words, select the number that best represents how you perceived Cali during your conversation.

Likeability

1.

Awful ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Nice

2. For data quality purposes, please select 3 for this question.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

3.

Unfriendly ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Friendly

Questions continued (same format as above):

4. Unkind - Kind

5. Unpleasant - Pleasant

6. Dislike - Like

Intelligence

7. Ignorant - Knowledgeable

8. Incompetent - Competent

9. Irresponsible - Responsible

10. Foolish - Sensible

11. Unintelligent - Intelligent

Anthropomorphism

12. Fake - Natural

13. Artificial - Lifelike

14. Machinelike - Humanlike

15. Unconscious - Conscious

Figure A.10: Godspeed likeability, intelligence, and anthropomorphism (page 9 of experiment).

Step 7 of 13 — 54% complete

Questionnaire 4

Participant ID: **researcher388**

During this experiment, you had a conversation with Cali. Please indicate the extent to which you agree with each of the following statements using the scale provided (1 = Not at all, 7 = Extremely).

1. Cali is deceptive.

- ☐ 1 - Not at all
- ☐ 2 - Slightly
- ☐ 3 - Somewhat
- ☐ 4 - Moderately
- ☐ 5 - Quite a bit
- ☐ 6 - Very
- ☐ 7 - Extremely

Questions continued (same format as above):

- 2. Cali behaves in an underhanded manner.
- 3. I am suspicious of Cali's intent, action, or output
- 4. I am wary of Cali.
- 5. Cali's action will have a harmful or injurious outcome
- 6. I am confident in Cali.
- 7. Cali provides security.
- 8. Cali has integrity.
- 9. Cali is dependable.
- 10. For data quality purposes, please select 1 for this question.
- 11. Cali is reliable.
- 12. I can trust Cali.
- 13. I am familiar with Cali.

Figure A.11: Trust in Automation Scale / TIA-S (page 10 of experiment).

Step 8 of 13 — 62% complete

Questionnaire 4

During this experiment, you had a conversation with Cali. Please indicate the extent to which you agree with each of the following statements using the scale provided (1 = Not at all, 7 = Extremely).

1. If available, I intend to start using voice assistants like Cali within the next month.

- ☐ 1 - Not at all
- ☐ 2 - Slightly
- ☐ 3 - Somewhat
- ☐ 4 - Moderately
- ☐ 5 - Quite a bit
- ☐ 6 - Very
- ☐ 7 - Extremely

2. If available, in the next months, I plan to experiment or regularly use voice assistants like Cali.

- ☐ 1 - Not at all
- ☐ 2 - Slightly
- ☐ 3 - Somewhat
- ☐ 4 - Moderately
- ☐ 5 - Quite a bit
- ☐ 6 - Very
- ☐ 7 - Extremely

Back

Next

Figure A.12: Intention-to-use items (page 11 of experiment).

Step 9 of 13 — 69% complete

Questionnaire 5

Participant ID: **researcher389**

Here are a number of personality traits that may or may not apply to you. Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

I see myself as:

1. Dependable, self-disciplined

- ☐ 1 - Disagree strongly
- ☐ 2 - Disagree moderately
- ☐ 3 - Disagree a little
- ☐ 4 - Neither agree nor disagree
- ☐ 5 - Agree a little
- ☐ 6 - Agree moderately
- ☐ 7 - Agree strongly

Questions continued (same format as above):

- 2. Extraverted, enthusiastic
- 3. Critical, quarrelsome
- 4. Open to new experiences, complex
- 5. For data quality purposes, please select 7 for this question.
- 6. Reserved, quiet
- 7. Anxious, easily upset
- 8. Disorganized, careless
- 9. Sympathetic, warm
- 10. Calm, emotionally stable
- 11. Conventional, uncreative

Figure A.13: Ten item personality inventory / TIPI (page 12 of experiment).

Step 10 of 13 — 77% complete

Final Feedback

Participant ID: **researcher388**

The more detail you provide, the more helpful your feedback and the higher your bonus payout may be.

Tip: Record your answer for each question (at least 20 seconds). You can also type below if you prefer. Your recording is sent to researchers; nothing appears on screen as you speak.

There are 3 questions on this page — scroll down to answer each one.

► Dictation Audio Debug

Question 1 of 3

EXPERIENCE WITH CALI

Describe your experience interacting with Cali during the conversation.

You may consider aspects such as:

- How Cali communicated overall
- How well Cali understood and responded to you
- Anything that stood out positively or negatively during the interaction
- Any technical issues during the call (audio quality, delays, pauses, voice changes, or glitches)

Record your feedback (at least 20 seconds)

🎤 Click to record

Minimum
20s+

Good
45s+

Great
90s+

Excellent
120s+

0:00 / 20 seconds minimum

Record at least 20 seconds total (short clips add up). You can also add text below.

Audio status: idle | Recorder: idle

Or add text below (optional)

Type additional feedback if you like...

Minimum
35+

Good
75+

Great
150+

Excellent
250+

0 / 35 words minimum

Figure A.14: Final feedback recording screen (page 13 of experiment).

91

Questions continued (same format as above):

Question 2 of 3

COMMUNICATION STYLE (FORMALITY)

How would you describe Cali's communication style? Please include:

- Whether Cali felt more formal or more informal (and what gave you that impression)
- How appropriate that style felt for this conversation
- Whether the style affected your comfort, engagement, or trust

Question 3 of 3

FEEDBACK ON THE EXPERIMENT

This question is about the study setup and experience, not about Cali.

Please share any feedback on the experiment itself, such as:

Clarity of instructions and tasks

- Technical issues with the study platform (navigation problems, issues replying to survey questions, page loading issues)
- Flow and pacing of the study
- Anything that could be improved for future participants

Figure A.15: Final feedback questions, continued from previous (page 13 of experiment).

Step 11 of 13 — 85% complete

Get early access before everyone else

You just tested an early version of Cali, our voice AI assistant.

In the future, Cali will be available for you to interact with whenever you want.

As one of our early testers, you can choose to get notified on Prolific when Cali launches publicly.

☐ I'd like early access — notify me on Prolific when Cali launches.

Your payment is not affected by your choice.

Anything you'd like to add? (optional)

e.g. how you'd use it, what you'd hope for, or leave blank

Submit

Figure A.16: Early access opt-in screen (page 14 of experiment).

Step 12 of 13 — 92% complete

Debriefing

Thank you for contributing to this research. The study is now complete, and the information below explains the full purpose of the experiment.

What this study was really about

This study examined how people respond to two different conversational styles used by a voice based assistant. Some participants interacted with a more formal assistant, and others interacted with a more informal one. We studied how these styles influenced trust and perceived empathy during the short interaction.

What data we collected

We collected:

- Demographical information
- Audio recordings of your interaction
- Questionnaire responses

All data is stored securely and handled under GDPR and Utrecht University policy.

Your right to withdraw your data

Finish Study

Figure A.17: Debriefing screen (page 15 of experiment).

Step 13 of 13 — 100% complete



Study Complete!

Thank you for your participation

Please use one of the options below to complete your submission:

[Click here to complete on Prolific](#)

Or copy and paste this completion code:

CWJF4IWH

Figure A.18: Study completion screen (page 16 of experiment).

A.3.3 Geriatric Depression Scale-30 Source Material

The following document contains the Geriatric Depression Scale-30 material used to inform the study check-in questions.

Geriatric Depression Scale (Long Form)

Patient's Name: _____ Date: _____

Instructions: Choose the best answer for how you felt over the past week.

No.	Question	Answer	Score
1.	Are you basically satisfied with your life?	YES / NO	
2.	Have you dropped many of your activities and interests?	YES / NO	
3.	Do you feel that your life is empty?	YES / NO	
4.	Do you often get bored?	YES / NO	
5.	Are you hopeful about the future?	YES / NO	
6.	Are you bothered by thoughts you can't get out of your head?	YES / NO	
7.	Are you in good spirits most of the time?	YES / NO	
8.	Are you afraid that something bad is going to happen to you?	YES / NO	
9.	Do you feel happy most of the time?	YES / NO	
10.	Do you often feel helpless?	YES / NO	
11.	Do you often get restless and fidgety?	YES / NO	
12.	Do you prefer to stay at home, rather than going out and doing new things?	YES / NO	
13.	Do you frequently worry about the future?	YES / NO	
14.	Do you feel you have more problems with memory than most?	YES / NO	
15.	Do you think it is wonderful to be alive now?	YES / NO	
16.	Do you often feel downhearted and blue?	YES / NO	
17.	Do you feel pretty worthless the way you are now?	YES / NO	
18.	Do you worry a lot about the past?	YES / NO	
19.	Do you find life very exciting?	YES / NO	
20.	Is it hard for you to get started on new projects?	YES / NO	
21.	Do you feel full of energy?	YES / NO	
22.	Do you feel that your situation is hopeless?	YES / NO	
23.	Do you think that most people are better off than you are?	YES / NO	
24.	Do you frequently get upset over little things?	YES / NO	
25.	Do you frequently feel like crying?	YES / NO	
26.	Do you have trouble concentrating?	YES / NO	
27.	Do you enjoy getting up in the morning?	YES / NO	
28.	Do you prefer to avoid social gatherings?	YES / NO	
29.	Is it easy for you to make decisions?	YES / NO	
30.	Is your mind as clear as it used to be?	YES / NO	
TOTAL			

This is the original scoring for the scale: One point for each of these answers.
Cutoff: normal-0-9; mild depressives-10-19; severe depressives-20-30.

1. NO	6. YES	11. YES	16. YES	21. NO	26. YES
2. YES	7. NO	12. YES	17. YES	22. YES	27. NO
3. YES	8. YES	13. YES	18. YES	23. YES	28. YES
4. YES	9. NO	14. YES	19. NO	24. YES	29. NO
5. NO	10. YES	15. NO	20. YES	25. YES	30. NO

Yesavage JA, Brink TL, Rose TL, et al. Development and validation of a geriatric depression screening scale: a preliminary report. *J Psychiatr Res* 1983; 17:37-49.

Provided courtesy of CME Outfitters, LLC

Available for download at www.neuroscienceCME.com