

Situated Understanding of Errors in Older Adults’ Interactions with Voice Assistants: A Month-Long, In-Home Study

AMAMA MAHMOOD, The Johns Hopkins University, USA
JUNXIANG WANG, Carnegie Mellon University, USA
CHIEN-MING HUANG, The Johns Hopkins University, USA

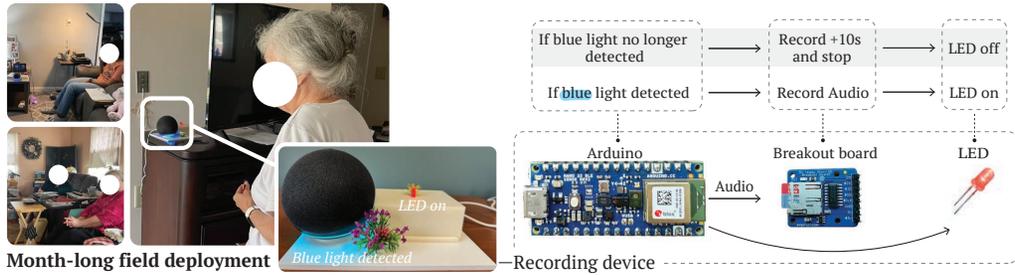


Fig. 1. We investigated the interactions of older adults with commercial voice assistants (VAs) in their homes over a period of four weeks with a focus on the emergence of interaction errors and participants’ repair strategies. To augment our collected interaction data, we recorded audio using the device depicted above. Each participant’s interaction with the VA was recorded in its entirety—plus an additional 10 seconds after the interaction concluded—to capture their spontaneous reactions and responses to the VA.

Our work addresses the challenges older adults face with commercial Voice Assistants (VAs), notably in conversation breakdowns and error handling. Traditional methods of collecting user experiences—usage logs and post-hoc interviews—do not fully capture the intricacies of older adults’ interactions with VAs, particularly regarding their reactions to errors. To bridge this gap, we equipped 15 older adults’ homes with smart speakers integrated with custom audio recorders to collect “in-the-wild” audio interaction data for detailed error analysis. Recognizing the growing use of Large Language Models (LLMs) to enhance capabilities of voice assistants, our study also explored how this integration of LLMs changes older adults’ interaction dynamics, specifically during errors. Midway through our study, we deployed ChatGPT-powered VA to investigate its efficacy for older adults. Our research suggests that while technical improvements—such as leveraging vocal and verbal responses combined with LLMs’ contextual capabilities—can enhance error prevention and management in VAs, interaction-level challenges still remain, particularly those unique to older adults. We propose design considerations to better align future VAs with older adults’ expectations and lived experiences.

Authors’ Contact Information: Amama Mahmood, amama.mahmood@jhu.edu, The Johns Hopkins University, Baltimore, Maryland, USA; Junxiang Wang, jwang334@jhu.edu, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Chien-Ming Huang, chienming.huang@jhu.edu, The Johns Hopkins University, Baltimore, Maryland, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym ’XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

<https://doi.org/XXXXXXXX.XXXXXXX>

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing Methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: voice assistant, older adults, human-agent interaction, personal assistant, errors, conversational breakdowns, conversational AI, large language models

ACM Reference Format:

Amama Mahmood, Junxiang Wang, and Chien-Ming Huang. 2018. Situated Understanding of Errors in Older Adults' Interactions with Voice Assistants: A Month-Long, In-Home Study. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 37 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Conversations are the most natural form of human communication. Speech interfaces can make AI technology more accessible to special populations, such as older adults, who often struggle with text-based interfaces [Kowalski et al., 2019, Liu et al., 2020, Wulf et al., 2014]. Thus, voice assistants (VAs) hold the potential to significantly enhance older adults' quality of life and autonomy [Blair and Abdullah, 2019, Smith et al., 2023]. Despite their potential and growing popularity, commercial voice assistants are not designed for special populations such as older adults [Stigall et al., 2019]. Data-driven, intelligent VAs are not immune to errors [Pearl, 2016], and older adults' adoption and sustained use of VAs are often impeded by common AI failures [Kim and Choudhury, 2021, Pradhan et al., 2020]. Errors are particularly prevalent in their interactions due to older adults' verbose and disfluent speech patterns [Mortensen et al., 2006]. Consequently, it is crucial to investigate and analyze the conversational breakdowns specifically experienced by older adults when interacting with these systems.

While researchers have investigated older adults' long-term interactions with VAs at home—focusing on their adoption of the technology, its benefits and challenges, and their perceptions of the experience [Kim and Choudhury, 2021, Pradhan et al., 2019, 2020, Purao et al., 2021]—a significant research gap remains in the area of comprehensive error analysis. A major barrier to studying errors in detail is insufficient error logging in VA usage data—resulting in entries like “Text not available” or “Unknown” [Pradhan et al., 2020]—which often leaves conversational context and real-time user reactions unrecorded. Furthermore, speech recognition errors are often not identifiable solely from transcribed usage logs [Kim and Choudhury, 2021]; supplementary audio recordings are required to capture more nuanced aspects of how smart speakers are embedded into conversational settings [Porcheron et al., 2018]. Given that commercial VAs do not provide a mechanism to record complete audio interactions, they fall short in capturing users' immediate thoughts and reactions; such audio data can also be particularly crucial in instances where errors disrupt conversational flow. Documenting how users respond or adapt to continue the conversation can offer insights into not only the nature, frequency, and occurrence of errors, but also users' perceptions of VAs and the strategies they employ to advance the conversation. Therefore, we pose: **What are the interaction dynamics between older adults and VAs during errors and breakdowns, and how do their interactions evolve over time?**

To overcome the lack of complete interaction data, we designed and developed a custom audio recorder to capture and log user interactions with a commercial VA—Amazon's “Alexa”—as audio files. We recorded each interaction from start to finish, plus an additional 10 seconds post-interaction to document users' verbal and vocal responses during and immediately after their interactions. We designed the audio recorder to be compact, packageable and easily deployable, also evidenced by our recent work [Mahmood et al., 2025b]. We deployed Amazon Echo Dot smart speakers augmented with our recording device at 15 older adults' homes for four weeks to capture rich, “in-the-wild” interaction data. While prior work has explored error types and recovery strategies

for both older adults [Angelini et al., 2021, Cuadra et al., 2022, Kim and Choudhury, 2021, Pradhan et al., 2019, 2020, Stigall et al., 2019] and general users [Ashktorab et al., 2019, Beneteau et al., 2019, Myers et al., 2018], our study presents new insights into older adults' interactions with VAs and highlights design considerations for making VAs more accessible to this special population: 1) most errors arise from VAs' limited understanding capabilities and inaccuracies in speech capture, highlighting the need for improved comprehension of diverse speech styles; 2) temporal error trends reveal that query reformulation and learning by experience alone are insufficient for effective error management for older adults and there is need to account for their age-related challenges such as forgetfulness; 3) leveraging users' immediate reactions as implicit cues and increasing the use of VA-initiated recovery can effectively identify and manage interaction breakdowns.

Recent advancements in natural language processing have enabled the integration of large language models (LLMs) into voice assistants [OpenAI, 2023, Rausch, 2023], significantly improving their ability to understand and generate human-like speech, thereby reducing erroneous interactions, as shown in our prior work [Mahmood et al., 2025c]. Prior research has examined the design of LLM-powered VAs to meet the needs of older adults from various perspectives, such as supporting health and daily well-being [Guo et al., 2025, Mahmood et al., 2025a, Rudnik et al., 2024, Yang et al., 2024], as well as general voice-based communication needs [Liu et al., 2025, Shahid et al., 2025]. However, most of this work has been conducted through one-off sessions and does not account for how interactions—particularly errors—may evolve in real home environments over extended periods when using LLM-powered VAs. This study provides an opportunity to further examine the potential benefits and challenges associated with integrating LLMs into voice assistants, specifically for information retrieval tasks undertaken by older adults within their homes. The set up allows us to explore how errors may differ for older adults over extended periods of use, as opposed to young adults in controlled lab settings, investigated in our previous work [Mahmood et al., 2025c]. Therefore, we integrated ChatGPT into Alexa skill, similar to our prior lab study (referred to as ChatGPT+Alexa skill [Mahmood et al., 2025c]), and used it as a technology probe [Hutchinson et al., 2003] during Weeks 3 and 4 of the study to observe how a VA designed and evaluated for young adults works for older adults in home settings. Our preliminary findings while emphasize the resilience of LLMs in conversation and illustrate the learning curve necessary for older adults to engage effectively with more sophisticated VAs, highlight that the challenge of LLM integration or any advancement of VAs is not only a technical problem but it an interaction one and LLM-powered VAs still face usability challenges unique to this special population. LLM-integration alone does not fix the usability challenges for older adults as well as for young adults and creates new ones. Thus our findings provide insights for the design of future, advanced VAs where LLMs is just one of the technologies enabling that. Our work's contributions are threefold:

- (1) **Analysis of real-world erroneous interactions with prevalent VAs:** We qualitatively analyzed “in-the-wild” audio interaction data to understand how older adults incorporate commercial VAs into their lives, focusing especially on the nuances of erroneous interactions. Our findings reveal both opportunities and challenges, offering valuable design considerations for creating more robust and fluid VA interactions to improve accessibility.
- (2) **Exploratory analysis of in-home longitudinal interaction with LLM-powered VA:** We deployed an LLM-powered VA as a technology probe to explore how older adults engage with conversationally adept, futuristic VAs to identify the opportunities and limitations of LLM integration into VA technologies for this specific population.
- (3) **Open-source lightweight data collection tool:** We open-source a compact, packagable and easily deployable tool for collecting in-the-wild audio data, enabling a deeper understanding of how users interact with and perceive VAs within their personal environments (Fig. 1 and 2).

2 Related Work

Our study explored the integration of voice assistants¹ into older adults' daily activities, focusing on errors and the impact of LLMs on their experience. We review relevant prior work below.

2.1 VAs in Older Adults' Lives

VAs have become a staple in many households [Lis, 2022], with notably high adoption amongst older adults [Kakulla, 2021, Lifestyle, 2022]; VAs not only exhibit significant potential in enhancing their quality of life [Blair and Abdullah, 2019, Smith et al., 2023], but also have been shown to uplift their spirits, provide more opportunities for mental engagement, and aid in fostering social connections [Oewel et al., 2023, Upadhyay et al., 2023]. Common uses of VAs among older adults include seeking information, getting weather updates, listening to music, and setting reminders [Arnold et al., 2022, O'Brien et al., 2020]. Voice-based agents have also been explored in assisting older adults with specific tasks—for example, in interacting with complex user interfaces, where VAs help them locate interface features [Yu et al., 2023], or controlling home appliances [Choi et al., 2020, Kowalski et al., 2019]—highlighting their preference for voice interfaces over text-based ones [Kowalski et al., 2019, Liu et al., 2020, Wulf et al., 2014].

While initial VA interactions among older adults are often entertainment-oriented, there is a notable shift toward practical applications as they become more familiar with the technology, particularly with regard to health-related queries [Brewer et al., 2022, Pradhan et al., 2020]. In terms of such task-specific applications, VAs have shown significant promise as health aids; prior research focusing on specific needs such as pain management found VAs to be particularly beneficial for older adults [Shade et al., 2020], with their abilities to set reminders and promote accountability in health routines especially valuable [Bolaños et al., 2020, Shade et al., 2020]. Other works have explored health information-seeking [Brewer et al., 2022, Harrington et al., 2022, Sanders and Martin-Hammond, 2019], preventive care [Yamada et al., 2018], and the potential use of reminders for medication administration [Bolaños et al., 2020, Chen et al., 2021, Jesús-Azabal et al., 2020].

Apart from VAs' practical use, older adults' social perceptions of these tools are influenced by various factors, such as their familiarity with technology, prior experiences with VAs, and personal preferences. For instance, a user's emotional state and even a smart speaker's physical location can affect whether the user views the assistant as just a machine or as a more humanlike companion [Pradhan et al., 2019]. As users grow more accustomed to VA technology, they begin to value not just the systems' operational ease, but also the sense of digital companionship they offer [Kim and Choudhury, 2021].

2.2 Conversational Challenges Faced by Older Adults While Interacting with VAs

Although older adults recognize the benefits of using VAs, some potential barriers to the adoption of this technology remain for the aging population. For instance, older adults' inability to find utility in smart speakers and their habitual reliance on other ways of completing tasks are mentioned as major factors contributing to their disuse of such devices [Trajkova and Martin-Hammond, 2020]; they also point out potential data protection issues and fear of being patronized and manipulated by VAs as concerns [Horstmann et al., 2023]. While VAs have the potential to support older adults' daily activities and well-being, frustrating conversational breakdowns that occur [Kim and Choudhury, 2021] may impact their adoption and continued use of the technology. Older adults experience higher error rates and often require more time to complete tasks when interacting with voice-based

¹In this manuscript, "voice assistant" refers to assistants in smart speakers, "voice-based" applies to agents on other platforms such as smartphones, watches, or web-based systems, and "voice interface" refers to the user interaction modality rather than the assistant.

virtual assistants [Barros and Seabra, 2020, Chiaradia et al., 2019], underscoring the increased conversational challenges they face due to factors such as loss of cognitive and motor skills and limited prior technology use [Barros and Seabra, 2020].

While research shows that older adults perceive that they have a good mental model of VAs' capabilities (such as the necessity of asking questions in a certain way) [Pradhan et al., 2020], other research points out that they are not fully sure of the extent of those capabilities [Kim and Choudhury, 2021]. Such a mismatch between their perceptions and reality can cause conversational breakdowns, negatively affecting user experience and leading to increased frustration. Even though older adults may realize that they need to say things a certain way, they often find it challenging to remember commands associated with VA features and skills [Pradhan et al., 2020, Stigall et al., 2019]—even those as simple as a wake word [Angelini et al., 2021, Cuadra et al., 2022, Pradhan et al., 2020]. Instead, they frequently resort to trial and error following conversational breakdowns [Kim and Choudhury, 2021] and may eventually give up trying to reformulate their queries after multiple failed attempts, leading to incomplete error recovery and thus limiting their use of VAs [Pradhan et al., 2020].

Although they are somewhat cognizant of VAs' limitations—such as their lack of follow-up and contextual understanding—older adults prefer and initially attempt a more conversational interaction style [Pradhan et al., 2019]. However, frustration arising from VAs' inability to remember context [Pradhan et al., 2019] often forces older adults to modify their approach. This leads to a breakdown, especially in information retrieval (*e.g.*, health-related queries), causing older adults to reluctantly shift from their preferred conversational style to more scripted, self-contained queries [Brewer et al., 2022]. Consequently, many interactions with VAs end up becoming command-based—indicating a transactional, rather than conversational relationship [Purao et al., 2021, Purao and Meng, 2019] and usually lack proactive interchanges [Yamada et al., 2018]. These shifts to transactional dynamics have led researchers to debate whether such interactions should truly be considered “conversations” in the first place [Porcheron et al., 2018].

The lack of holistic analysis of daily user-VA interactions (due to the absence of real-world audio interaction data) makes it difficult to concretely interpret the extent of older adults' awareness of VAs' capabilities and their efforts to recover from conversational breakdowns [Kim and Choudhury, 2021]. Therefore, an in-depth analysis of VA failures in interactions with older adults is necessary to better understand the complexities of their daily interactions. While previous research has identified error types, estimated error rates, and examined recovery strategies in older adults' VA use [Kim and Choudhury, 2021, Pradhan et al., 2020], these studies do not address the sequential nature of how errors compound when users attempt to correct them. Our study closes this gap by utilizing recorded interaction data, which captures participants' immediate reactions and actions (*e.g.*, recovery attempts) after conversational breakdowns, allowing us to analyze the error “snowball effect” and gain a deeper understanding of the dynamics involved in older adults' erroneous interactions with VAs.

2.3 Research Tools and Methodologies for Longitudinal Field Studies on Older Adults and VAs

Several rigorous methodologies have been adopted to gather data about older adults' use and perception of VAs in longitudinal at-home studies, which offer insights into their changing perceptions and interaction dynamics over time. One primary approach is the combination of semi-structured interviews [Kim and Choudhury, 2021, Pradhan et al., 2019, 2020] and daily diary entries [Pradhan et al., 2020], enabling researchers to gather subjective user perceptions of their experiences with VAs. A second approach relies on complementing understanding of user perceptions with a detailed analysis of usage logs to reveal usage frequency and evolving patterns in VA utilization [Kim and

Choudhury, 2021, Oewel et al., 2023, Pradhan et al., 2020, Purao et al., 2021]. However, this method is constrained by transcription inaccuracies, which complicate error comprehension based solely on usage logs [Kim and Choudhury, 2021]—further suggesting that the usage logs are not always objective, either (*i.e.*, they are not always truly reflective of users’ interactions with VAs). Moreover, usage logs are unable to capture organic vocal reactions, interruptions, and overlapping speech during VA interactions. Longitudinal studies on VAs for older adults lack real audio interaction data, which has shown to be beneficial for understanding user behavior [Porcheron et al., 2018]; such data can be particularly useful in understanding conversational breakdowns. Prior work has also highlighted the need for further research on incorporating objective audio data to deepen our understanding of such interactions [Purao et al., 2021]. However, the existing tool [Porcheron et al., 2018] is not open-sourced, bulky, and captures one minute of data before and on onset of interaction which can be intrusive, specifically for vulnerable populations. Therefore, in this study, we developed a compact, packageable, and easily deployable tool to augment interaction data with real-world audio recordings, thus capturing older adults’ verbal and vocal reactions and responses to Amazon’s Alexa; this feedback also includes interruptions and overlapping speech, which are not logged in traditional usage data.

2.4 Integration of LLMs into Voice Assistance

Despite older adults’ desires to have conversational interactions with VAs, current commercial VA capabilities are far more limited [Pradhan et al., 2019]. LLMs, on the other hand, have shown enhanced conversational capabilities, yielding robust and fluid interactions with users [Chan et al., 2023, Mahmood et al., 2025c]; our prior work shows LLMs have also shown potential in absorbing various speech recognition errors for young adults [Mahmood et al., 2025c].

Recent efforts to incorporate LLMs into voice interactions for older adults have shown improvement in conversational aspects [Jo et al., 2023, Mahmood et al., 2025a, Mathur et al., 2022, Yang et al., 2024]. Design of LLM-powered voice assistants for older adults has been explored for various purposes in recent work. Most of this work is centered around providing support for health needs and daily tasks and well-being [Guo et al., 2025, Mahmood et al., 2025a, Rudnik et al., 2024, Yang et al., 2024] and general voice-based communication needs [Li et al., 2024, Liu et al., 2025, Shahid et al., 2025]. Most of this work focuses on acceptance of technology, general experiences and perceptions, and task specific needs. Additionally most of this work [Guo et al., 2025, Liu et al., 2025, Mahmood et al., 2025a, Rudnik et al., 2024, Shahid et al., 2025, Yang et al., 2024] studies experiences or validates their systems in one off settings and more importantly studying breakdowns is not the focus, especially in longitudinal study and through analysis of interaction data and immediate user perceptions. In this work, we explore if incorporating LLMs into VAs can reduce errors for older adults as seen in our prior work for young adults [Mahmood et al., 2025c] and what new error and recovery patterns might emerge.

3 Methods

To understand older adults’ perceptions and interactions with VAs, we conducted a four-week field deployment of a smart speaker, the 5th generation Amazon Echo Dot. We collected data via usage logs, audio interactions from a novel recording device, and semi-structured interviews at the beginning and end of deployment. This section outlines the implementation of our recording device and integration of ChatGPT into an Alexa skill, followed by a description of our longitudinal deployment, participant details, and data analysis approach.

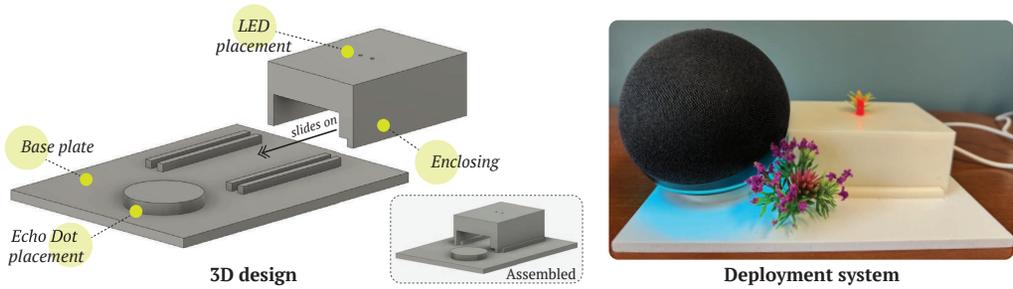


Fig. 2. Design of the enclosure and base plate for placing the Echo Dot smart speaker and recording device. CAD files are available online at <https://bit.ly/3UIgAri> for 3D printing.

3.1 Data Collection: Recording Device

In our study, we utilized a specially developed recording device to capture users' interactions with Alexa, including their immediate responses and reactions following each voice interaction. The recording device comprises of three modules:

- (1) **Detector:** The recording device detects the activation of Alexa—indicated by the blue light ring on the bottom of Amazon Echo Dot—via the light color detector on an Arduino NANO BLE SENSE 33.
- (2) **Recorder:** Triggered by the detection of the Echo Dot's blue light, the Arduino begins recording audio onto a microSD card inserted into a connected Adafruit breakout board. Recording ceases 10 seconds (tunable parameter) after the blue light turns off. However, if the user reactivates Alexa within this 10-second window, the device continues recording until the end of the new interaction, plus an additional 10-second period.
- (3) **Privacy indicator:** To address privacy concerns and enhance user comfort, our device does not continuously buffer audio for keyword detection, as utilized in prior research [Porcheron et al., 2018]. Instead, it only buffers audio when the blue light is detected at the start of an interaction. To ensure transparency, a red LED light positioned on top of the recording device illuminates to indicate active audio recording, allowing the user to closely monitor recording status. At any point, the device can be unplugged from the power outlet to stop recording and simply plugged back in to resume the recording function.

The recording device is adaptable for use with other smart speakers, as integration with wake word detection is also feasible. We also designed and 3D-printed a compact enclosure for the recording device, which can be mounted on a base plate with a designated spot for the Echo Dot, as shown in Fig. 2. The code for the recording device, 3D printing sketch, and assembly instructions will be publicly available².

To validate the functionality of our recording device, we conducted extensive in-lab testing during development, followed by two pilot tests—one with a young adult and one with an older adult. In each pilot, a smart speaker and our recording device were deployed in participants' homes for one week. The first author listened to all the recordings and cross-checked them against Alexa usage logs to identify any unnecessary lag. The in-lab and pilot tests confirmed that the device operated as intended: blue light detection was instantaneous, the first word spoken after “Alexa” was accurately captured, and the 10-second duration was sufficient to record participants' immediate thoughts and feedback without being overly intrusive, according to the pilot participants' perceptions.

²Link to the code and design for our recording device and additional study materials: <https://bit.ly/3UIgAri>.

3.2 Technology Probe: Integrating ChatGPT into an Alexa Skill

To investigate older adults' interactions with more conversational VAs and to examine the associated benefits and challenges thereof, we deployed an LLM-powered VA by incorporating ChatGPT-3.5 into an Amazon Alexa skill, similar to our prior work with young adults [Mahmood et al., 2025c] that implemented a dual-API mechanism to handle Alexa skill timeout issues. To enable intuitive activation by older adults, we used the same natural activation phrases for the ChatGPT skill, such as "Alexa, let's chat!" or "Alexa, question" (complete list is provided in Appendix A.1). Upon user activation, the ChatGPT-powered VA introduced itself; users could then have follow-up conversations without repeatedly using the wake word "Alexa." We refer to this skill as the "ChatGPT+Alexa skill" or simply "ChatGPT skill." This implementation, designed for general information retrieval and deployed as a technology probe, aims to reveal the potential advantages and limitations of LLM integration into VAs. Adopting the technology probe framework in [Hutchinson et al., 2003], we focused on investigating the usability of an LLM-powered VA in real-world settings.

3.3 Longitudinal Study: Procedure

The study comprised a four-week field deployment that included an initial interview, device setup, a demo session, a check-in at Week 1, the deployment of our LLM-powered Alexa skill at Week 3, and a final session at Week 4 involving a semi-structured exit interview.

3.3.1 Week 1: Initial interview and smart speaker setup. Taking place at the beginning of Week 1, the first session consisted of following activities in given order:

- (1) *Introduction and consent.* The experimenter explained the study and obtained consent.
- (2) *Demographics survey.* Participants completed a demographics survey (Table 1) and a short quality of life survey [Bowling et al., 2013].
- (3) *Pre-study interview.* The experimenter conducted a brief semi-structured interview to gauge participants' awareness and perception of smart technologies in general and of VAs (e.g., Amazon's Alexa and Apple's Siri).
- (4) *Smart speaker setup.* The experimenter set up a smart speaker at each participant's desired location. The experimenter informed participants that they could move the speaker and demonstrated how. An instruction page for debugging internet issues was provided. Participants could additionally contact the experimenter at any time via email, text, or phone call. At this time, the Amazon Alexa app was installed in their device of choice (e.g., smart phone or tablet); the experimenter set up participants' name pronunciations, voice profiles, and device location, and turned on the built-in adaptive listening feature and follow-up mode in the smart speaker's settings. (The adaptive listening feature provides users an extended duration with which to complete their query before Alexa initiates its response; follow-up mode allows them to ask Alexa follow-up questions without repeating the wake word.)
- (5) *Demonstration of Alexa and recording device.* The experimenter gave participants an introductory tutorial that covered fundamental VA capabilities including tasks like setting alarms, reminders, and timers; creating shopping and to-do lists; playing music; asking to hear jokes; posing questions; and engaging in casual, unstructured question-and-answer sessions. Furthermore, the experimenter demonstrated the process of accessing and utilizing various Alexa skills. A printed copy of this introductory tutorial was left with each participant². The experimenter explained that the audio recorder only records when the red LED light is on, removed the microSD card, and played sample recordings to ensure participants understood what would be recorded during their interactions. Additionally, participants were informed that they could stop the recording simply by unplugging the device and resume it by plugging it back in. They

Table 1. Demographics for our field study. All participants were fluent in English. All were retired except P10a (employed full-time) and P11a (employed part-time). Overall quality of life (QOL) is self-reported; their response to one question.

P	Gender	Age	Ethnicity	Disabilities/Aids	Overall QOL	Prior VA Use	VA Placement
Community Center – Assisted Living (by self)							
2	M	66	Prefer not to say	Wheelchair	Alright	Never	Studio
3	F	82	Caucasian	Hearing aid	Alright	Never	Studio
4	M	79	Caucasian	Wheelchair	Good	In the past	Studio
Community Center – Independent Living (by self)							
5	F	77	Caucasian		Good	Never	Living room
6	F	81	Caucasian	Cane/Rollator	Alright	Never	Living room
7	F	74	African American			Never	Living room
14	F	73	African American		Good	Siri	Living room
15	F	73	African American		Good	Never	Living room
Homeowner (by self)							
8	F	84	Caucasian	Walker	Alright	Never	Bedroom
9	M	75	Caucasian		Very good	Alexa	Dining room
13	M	80	Caucasian	Vision issues	Alright	Never	Dining room
Homeowner (Couple, with spouse)							
1a	M	76	Caucasian		Good	Alexa, multiple	Living room
1b	F	75	Caucasian			Alexa; multiple	Living room
10a	M	68	Caucasian		Good	Never	Living room
10b	F	94	Caucasian	Walker/scooter	Bad	Never	Living room
11a	M	72	Caucasian		Very good	Siri	Kitchen
11b	F	71	Asian		Very good	Siri	Kitchen
12a	M	75	Caucasian		Very good	Alexa; multiple	Living room
12b	F	72	Caucasian		Good	Alexa; multiple	Living room

were also shown that pressing the microphone-off button on Alexa or unplugging the device would deactivate it, thereby preventing any recording, as indicated by absence of the blue light.

At the end of Week 1, the experimenter contacted participants by phone to briefly inquire about their smart speaker usage and check for any technical difficulties.

3.3.2 Week 3: ChatGPT deployment. At the beginning of Week 3, we remotely deployed the implemented ChatGPT skill. Comprehensive instructions, along with illustrative examples and suggested conversation prompts, were provided to participants²; these instructions were communicated either via email or in print through our liaison at the community center. The experimenter then contacted all participants over the phone to introduce and demonstrate the skill as a means to hold more natural conversations with a VA that is capable of responding to follow-up questions and statements. This skill was available for participants throughout week 3 and 4 of the study in addition to traditional Alexa functionality. As this part of study was exploratory in nature, the LLM-powered skill was introduced only during the final two weeks of the study to avoid disproportionately influencing participants' interactions with the more commonly available VAs at the time.

3.3.3 Week 4: Collection of smart speaker. At the end of Week 4, the study concluded with smart speaker collection and a semi-structured exit interview aimed at gathering insights from the participants regarding their four-week experience specifically targeting errors and challenges they faced. Participants could request to review their recorded interaction data, the experimenter presented the recordings stored on the SD card. One participant chose to review their recordings.

Table 2. Errors and their definitions, grouped by category, along with their occurrences in the interaction data and resolution rate in the next immediate user retry. *Percentage of time participants reacted to each error type.

Error	Definition	Count	Resolved	Reactions*
Human errors	User made a mistake	72	38.9%	6.9%
Wrong wake word	User used the wrong wake word	49	30.6%	8.2%
Partial query	User did not complete their query	23	56.5%	4.3%
Speech errors	Errors caused due to speech recognition inaccuracies	154	31.8%	15.6%
Not listened	User query was not listened by Alexa	69	36.2%	4.3%
Mis-trigger	Alexa was triggered when user did not intend as such	15	6.7%	40%
Partially listened	Alexa captured user query partially	29	27.6%	20.7%
Interruption	Alexa interrupted the user query	1	100%	100%
Transcription	User query was transcribed inaccurately	40	35%	20%
VA errors	Errors attributed to the processing of accurately captured speech	406	20.44%	14.3%
System	User query was captured accurately but Alexa failed for an unknown reason and responded with a default phrase such as <i>"I'm having trouble getting that right now. Please try again."</i>	15	26.7%	13.3%
Not handled	User query was captured accurately but Alexa did not respond at all	57	17.5%	0.0%
Limitation	User query was accurately captured but Alexa failed to fulfill the request, responding with a default phrase, <i>"Sorry, I don't know about that."</i>	130	10.8%	14.6%
Intent recognition	User query was captured accurately but Alexa failed to recognize user intent and responded inaccurately	204	27%	18.1%

3.4 Participants

We deployed 15 smart speakers in three waves for four weeks each. Recruitment was based from a community center attached to both independent and assisted living communities; we also recruited community-dwelling adults from previous connections in the Baltimore, Maryland area. Table 1 summarizes their demographics and awareness of VAs. Participants were compensated with either the option to keep their smart speaker or an amount equivalent to its market price (\$50). All participants except three—P5 and couples P10 and P11—kept the speaker. This study was approved by our institutional review board.

3.5 Data and Analysis

Audio recordings from each device were transcribed and matched to usage logs from the Amazon Alexa dashboard. We also noted non-verbal and vocal cues such as laughter or filler noises (e.g., "huh," "hmm") in our transcripts. We took note of any interruptions, overlapping conversations, and comments or remarks about interactions—especially those occurring around interaction errors. We collected a total of 20 hours and 40 minutes of usage logs and audio recordings from the devices; audio data for participant P6 was missing due to a recorder malfunction. For the audio data and Alexa usage logs, we identified 2552 user query-VA response pairs (also referred to as "turns") across all participants.

Initially, the first author examined the interaction data, using an inductive approach to identify preliminary codes to gain a broader understanding of conversational breakdowns. Following this, two researchers independently coded the data for two participants (approximately 10% of the total) using these initial codes. They then convened to discuss and resolve any conflicts, resulting in the formulation of the final codebook given in Appendix B. All transcripts were subsequently

analyzed with the established codes deductively applied to each user query-VA response pair. Each query-response pair was coded for its purpose (type of interaction as informed by prior work [Kim and Choudhury, 2021, Pradhan et al., 2020]), the presence and type of error (as defined in Table 2), whether the error was resolved, and the user's recovery method (e.g., moving on, repetition, or clarification). In our analysis, we define errors as clear mistakes identifiable by human coders, and conversational breakdowns as instances where such errors lead to a discernible disruption in the interaction flow with the participant. We noted whether an error led to a conversational breakdown and if it was evident that the participant identified the error from their reaction(s) or actions (by initiating recovery). Participant interviews were transcribed, and inductive thematic analysis was conducted to identify their perceptions of the underlying causes of errors, attribution of blame, and recovery strategies. We also coded barriers to usage and their expectations from VAs. The inter-rater reliability using Cohen's Kappa on 10% data was 0.92 which signals strong agreement. After coding, we employed several quantitative measures: Error rate is defined as the total number of errors divided by the total number of one-turn user-query pairs. Categorical error rate refers to the proportion of errors within a specific category relative to all one-turn user-query pairs in that category. Resolution rate captures the percentage of errors that were successfully resolved in the immediate subsequent user-query turn.

4 Findings

Our dataset has a total of 2,552 one-turn (user query-VA response pair) interactions; the interaction data highlights the varying nature of participant engagement with the VA, ranging from brief, singular commands to more extended dialogues³. Overall, we observed a variety of interactions—from setting reminders to seeking information about various topics—as shown in Tables 3 and 4 and observed decline in interaction frequency over time, consistent with prior work [Lopatovska et al., 2019, Pradhan et al., 2020]. Since the focus of this paper is on conversational errors, analyses on participants' usage patterns and social nature of interactions are provided in supplementary materials³. In this section, we present a detailed analysis of participants' erroneous interactions with the VA—including our LLM-powered VA (64 queries), though the focus is on default Alexa interactions—followed by additional analysis of their interactions with our LLM-powered VA.

4.1 Analysis of Erroneous Interactions

We conducted a detailed analysis of errors, and users' experience and behavior surrounding them—an area less explored in prior research. By examining users' immediate responses and efforts to rectify errors as captured via our recording device, we were able to explore the intricate dynamics of user interactions during conversational breakdowns.

True error rate in VA interactions. Our study revealed that a considerable 24.76% of user-VA one-turn queries ($n = 632$ out of 2552) had errors, equating to almost one in every four queries failing. Of these, 98.10% manifested as conversational breakdowns, while the remainder were speech errors (e.g., transcription errors) that did not cause breakdowns. This number represents the true error rate, contrasting with prior work [Kim and Choudhury, 2021] where activation errors (among others) were unaccounted for, as they typically are not captured in user logs. However, our methodology, which included the use of a recording device, enabled us to capture these activation errors (e.g., wrong wake word, mis-trigger, not listened, partially listened, etc.), providing a more comprehensive understanding of the true error rate in VA interactions. We acknowledge, however, that some wake-word errors may have gone unrecorded if participants used entirely unrelated words that did not activate Alexa's blue light. While coding the data, we noted whether an error was

³Details in [Supplementary materials](#).

Table 3. Interaction categories and types with their respective counts and success rates.

Category	Type	Count	Success	Category	Type	Count	Success
Functional	Reminder	251	48.60%	Entertainment	Music	324	79.01%
	Weather	205	83.41%		Radio	232	73.70%
	Timer	174	90.23%		Alexa skills	89	58.43%
	Alarm	98	80.61%		Joke	22	86.36%
	Command	94	94.68%		TV	10	20.00%
	List	90	77.78%		Story	7	71.43%
	Communication	71	98.59%		Poem	3	33.33%
	Alexa capabilities	69	52.17%		Total	753	71.84%
	Time	46	100.00%		Information-seeking	Question	414
	Repeat	16	62.50%	Social	Greetings & gratitude	174	93.68%
	Calendar	12	58.33%	ChatGPT-Alexa Skill (ChatGPT)		64	68.75%
	Notes	1	100.00%	Other	Misc.	17	0.00%
	Total	1130	75.93%				

Table 4. Type of information-seeking questions asked. n = count; ER = error rate (percentage of queries of a particular question type that resulted in an error). For instance, 59.62% of total “local information” questions resulted in at least one error ($n = 31$ out of 52).

Type of question	n	ER	Type of question	n	ER	Type of question	n	ER
Entertainment	98	28.57%	Miscellaneous	38	44.74%	STEM-related	16	31.25%
<i>What was the score of [baseball team]?</i>			<i>Who makes the hydro razor?</i>			<i>What's the square root of three?</i>		
Local information	52	59.62%	Food & drink	37	21.62%	Specific places	14	64.29%
<i>Where can I get the best price for eggs?</i>			<i>How do I make macaroni salad?</i>			<i>Is [community center] a nonprofit facility?</i>		
Meaning & definitions	45	42.68%	Famous personalities	31	58.06%	History	10	60.00%
<i>What does the term pernicious mean?</i>			<i>What are Joe Biden's roots?</i>			<i>How many US men were killed in WWII?</i>		
Health	43	55.81%	Alexa-related	21	52.38%	Travel	9	11.11%
<i>What does low sodium level indicate?</i>			<i>Do you ever get angry?</i>			<i>Is there a backup on [X] bridge today?</i>		

resolved in an immediate retry; only 25.47% ($n = 160$) of all errors were resolved in the immediate next attempt by participants, indicating that many errors either remained unresolved or required multiple retries.

Consistent errors rates over time. We observed that error rate remained consistent, between 20–30%, across four weeks despite a varying number of interactions (Fig. 3, top). However, error rates in Weeks 1 and 3 were slightly higher than in Weeks 2 and 4, possibly due to the novelty effect from the introduction of the VA at the beginning of Week 1 and the novelty and Hawthorne effects from the introduction of a new feature (*i.e.*, the ChatGPT+Alexa skill) at the beginning of Week 3, which resulted in exploratory user behavior aimed at testing the boundaries of the system.

Below, we report our findings on types of errors and their resolution, the distribution of errors across interaction types, the indication of error recognition by participants through their actions and reactions, observed recovery strategies, and the compounding nature of errors. We also demonstrate how various aspects of interaction errors evolved over time.

4.1.1 Types of errors and their resolution rates. Our interaction data encompassed a diverse range of errors (Table 2), which we broadly categorized into VA, speech recognition, and human errors: **VA errors.** We found that the most frequent errors encountered by participants while interacting with Alexa were intent recognition errors, accounting for 32.3% ($n = 204$) of all 632 errors (Table 2). The next most common error category involved Alexa’s limitations in handling user requests

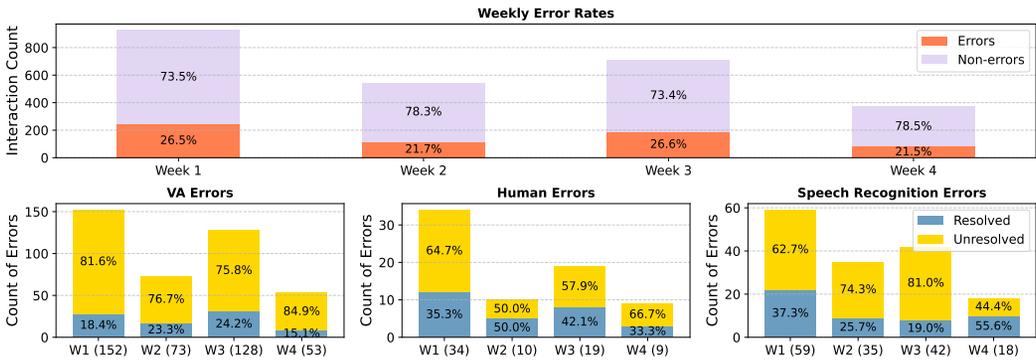


Fig. 3. Error categories and their resolution upon immediate participant retry attempt across four weeks.

($n = 130, 20.6\%$). VA errors, while being most frequent, had the lowest resolution rates (20.44%) as opposed to speech recognition and human errors (Table 2), further highlighting Alexa’s limited capabilities in correcting errors despite its attempts at recovery.

Human error. On 23 occasions (3.6%), an error occurred because a participant did not complete their query. Participants also incorrectly used “Alexis” instead of “Alexa” ($n = 49, 7.8\%$), a common confusion noted previously [Angelini et al., 2021, Cuadra et al., 2022, Pradhan et al., 2020]; the system was able to pick up the wrong wake word as a potential trigger (thus, picked up by our recording device), but marked it as “Not intended for Alexa.” Human errors were the least frequent and had the highest resolution rate (38.9%).

Speech recognition errors. Speech recognition errors such as not listened ($n = 69, 10.9\%$), transcription errors ($n = 40, 6.3\%$), and partially listened queries ($n = 29, 4.6\%$) were less frequent. Moreover, we observed that resolution success was higher for such errors (31.8%) as opposed to VA errors, suggesting that the primary challenges of smart speaker use revolve around understanding user intent and taking requisite action rather than fixing listening inaccuracies.

Consistent over time. We noticed that there was no clear pattern of improvement in resolution rates over the span of four weeks across all error categories (Fig. 3, bottom); the lack of apparent improvement in error resolution indicates that participants struggled with errors regardless of gaining familiarity with voice assistance through direct experience.

4.1.2 Occurrence of errors by interaction type. Error frequency varied across interaction types as discussed below.

Information retrieval. Most errors occurred during information retrieval interactions (42.51%, 176 of 414); only 8.7% (36 of 414) of these were resolved (see Fig. 4, top and example conversations C1 and C2 in Table 5). Within information-seeking interactions, certain query types had higher error rates (Table 4), such as those dealing with specific places (64.29%), history (60.00%), local information (59.62%), famous personalities (58.06%), and health-related queries (52.38%), as compared to queries about travel (11.11%), food and drink (21.62%), and entertainment (28.57%). This distinction is important because errors in high-stakes queries (e.g., local or health information) cause more frustration—participants may wish to act on the information provided in these critical settings—compared to lower-stakes entertainment queries. Other types of questions that were less frequently asked by participants—such as those about specific places (64.29% error rate) and historical information (60.00%) or those that did not fit into any category (44.74%)—had higher error rates, highlighting the VA’s limitations in handling less common or unique edge queries; this bears

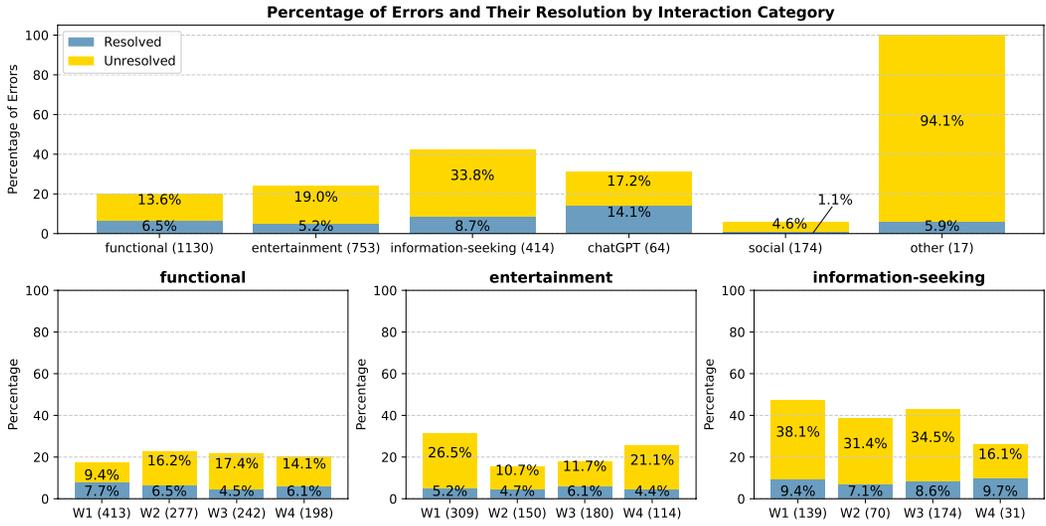


Fig. 4. Error resolution across interaction types over time. The top plot shows for each interaction category the percentage of interactions that resulted in errors and the percentage of interactions in which the errors were resolved; for instance, 414 interactions were categorized as “information-seeking” (100%), ~ 44% (~ 34%+ ~ 9%) of those interactions resulted in an error, and only ~ 9% of the original 414 were resolved after the first retry. The remaining plots for the top three individual categories show how their error resolution rates changed over the full four weeks. Note: Not all errors had retries and some had multiple retries, as discussed later.

semblance to the “long tail problem” in information retrieval [Bernstein et al., 2012]. In statistics, a “long-tailed” distribution describes a scenario in which a large number of occurrences are far from the norm—referring to situations in computation where a small set of common cases are easy to solve, but a vast range of uncommon edge cases render the problem intractable. This is evident in user queries for information retrieval: frequent entertainment-related queries such as “*Who won the baseball game?*”, ($n = 98$ of 414) form the “head” of the distribution, while the remaining less frequent or unique queries ($n = 316$ of 414) constitute the “tail.” Despite their individual infrequency, these long-tail queries dominate overall [Bernstein et al., 2012]. Notably, the error rate for entertainment-related queries was lower than those for “long-tail” queries, underscoring the challenges of handling diverse and infrequent requests and rendering them more error-prone.

Functional. Frequent errors were also noted in functional interactions; about 20% of such interactions resulted in breakdowns and only 6.5% were resolved after a retry. For simple functional tasks such as reminders, a surprisingly significant number of queries failed—88 (35.06%) of the 251 total reminder interactions—despite being relatively straightforward. Further examining errors related to Alexa’s reminder functionality, we noted that the majority ($n = 69$, 78.41%) of erroneous reminder interactions occurred when participants tried to stop a ringing reminder. Interaction data revealed that they often used naturalistic phrases such as “*Thank you*” or “*I got it*” to indicate their intent to stop reminders and timers. However, the VA did not recognize or interpret these phrases as stop commands, as it typically only responds to more scripted commands like “*Stop*,” “*Cancel*,” or “*I’m done*” (C4 in Table 5).

Entertainment. Significant errors were also found in popular entertainment tasks, such as accessing music (20.99% error rate out of 324 interactions) and playing radio (26.30% out of 232). Additionally, edge-case queries—accessing TV channels through the smart speaker (80% error

Table 5. Example conversations related to errors. “...” represents the truncation of a conversation for brevity.

Tag Participant	Alexa
C1 What are Joe Biden’s roots?	According to an Alexa Answers contributor: Ireland.
P2 Alexa, that’s his lineage. I wanna know about his mother and his father.	[Not listened by Alexa]
Alexa, tell me about virus family, his parents. [Transcription error; he actually said: “Alexa, tell me about Joe Biden’s family, his parents.”]	From cdc.gov: Mayaro virus (MAYV), the cause of Mayaro fever, is a member of the genus Alphavirus, family Togaviridae ...
Alright Alexa, you did a really good job at answering a question I did not ask you.	[Not listened by Alexa]
Alexa, tell me about Joe Biden’s family, his mother, his father.	Joe Biden’s children are Ashley Biden, Beau Biden, Hunter Biden, and Naomi Biden. His parent is Jean ...
Alexa, show his mother was a single mother. [Transcription error; he actually said: “Alexa, so his mother is a single mother?”]	From educalingo.com: The definition of “single mother” in the dictionary: A mother who has a dependent child or dependent children and who is widowed ...
Alexa, Joe Biden’s family was—did he have a father?	Joe Biden’s father is Joseph R. Biden, Sr.
What did he do?	[Not listened by Alexa]
Alexa, what was Joe Biden, Sr.’s educational background and what did he do?	This might answer your question: Joe Biden was educated at Syracuse University and the University of Delaware.
Alexa, what was Joe Biden’s father’s background?	Hmm, I don’t know that one.
C2 Alexa, do you have a memory of what I say to you?	I didn’t find any stickies. If you’d like to create one, just ask me to make a new sticky note.
P2 Alexa, I don’t want a sticky note, I want you to remember the conversation. Do you have any kind of memory about what I asked you?	Okay, I’ve saved your sticky. [Intent error]
Alexa, never mind.	[No response from Alexa]
Alexa, in the most recent literature produced by the FDA, what kind of cancer does Aspartame cause?	According to Mayo Clinic: Cancer is caused by changes, or mutations, to the DNA ... [Intent error]
Alexa, you’re a circular idiot.	[No response from Alexa]
C3 Alexa, tell me something wonderful.	I wasn’t able to understand the question I heard.
P11a It’s okay. [laughter] It’s amazing.	[Limitation error] Great! [“It’s amazing” not listened by Alexa]
C4 Alexa, I took my medicine.	Sorry, I don’t know that. [Intent error]
P15 Yes, I took my medicine.	Hmm, I don’t know that one. [Intent error]
Cancel medicine.	[P15’s name], I can’t find medicine. What’s the date or time of the event you want to delete? [Intent error]
Cancel the alert.	[Reminder stops ringing]

rate), asking for poems (66.7% error rate), and other miscellaneous queries (100% error rate)—predominantly resulted in errors, indicating VAs’ limitations in handling less common queries. **Error occurrence and resolution over time.** There was no distinct pattern that indicated any reduction of error rate or increased rate of error resolution across various tasks over four-week period (Fig. 4, bottom).

4.1.3 *Error identification and reactions to errors.* Participants recognized 79.84% of conversational breakdowns through verbal or vocal reactions, or immediate corrective action, with individual error

Table 6. Verbal reactions to errors, not including attempts to correct the error. 545 errors had no reactions. Participants reacted to the remaining 87 errors. 6 errors had compound reactions that were separated and counted under each type of reaction.

Verbal Reaction to Error	Count	Verbal Reaction to Error	Count	Verbal Reaction to Error	Count
No reaction	545	Laughter	14	Gratitude	6
Negation	27	Negative comment	10	Positive comment	5
Interruption	21	Acknowledgement	7	Vocal (e.g., sigh, shout)	3

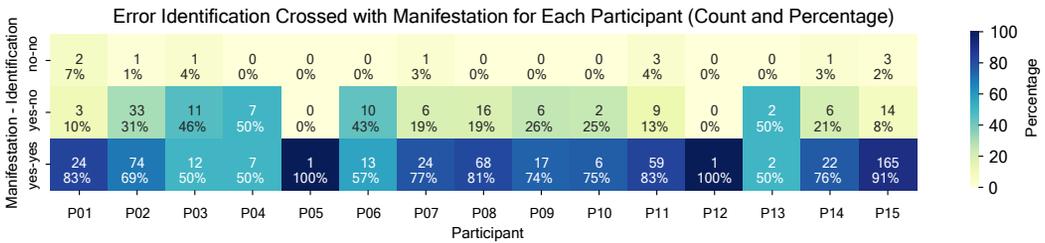


Fig. 5. Participants’ responses and explicit actions indicate their positive identification of most errors. For a y-axis error manifestation-identification pair, “no-no” indicates that the error did not manifest and was not identified by the participant, “yes-no” indicates that the error manifested but there was no indication of its identification by the participant, and “yes-yes” indicates that the error manifested and was identified by the participant as evidenced by either their immediate response or their action to rectify the error.

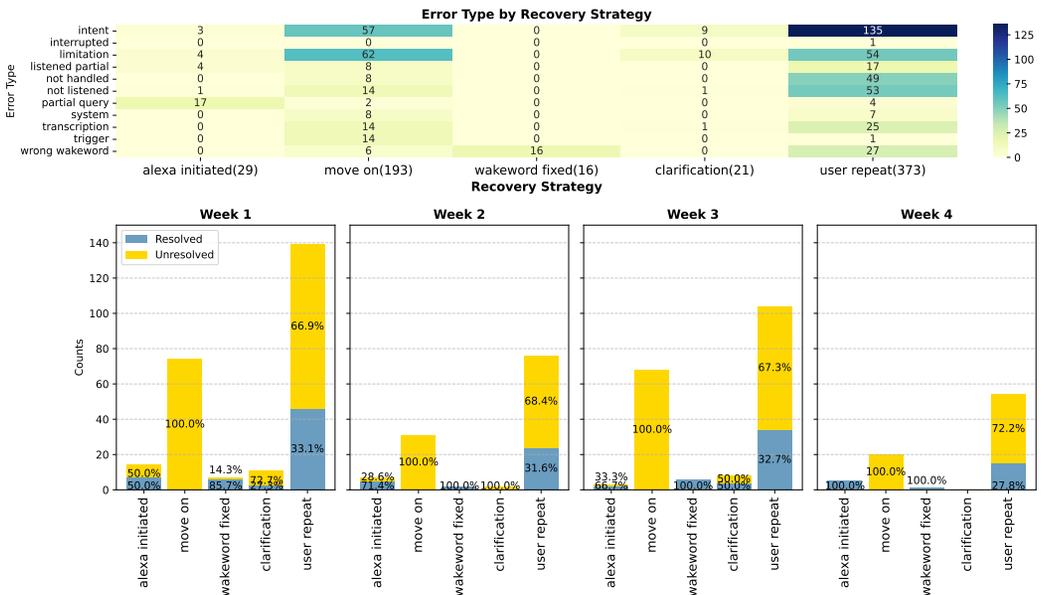


Fig. 6. Recovery strategies used by participants to address various types of errors. Cumulative totals are shown below each strategy on the x-axis in the top plot. The bottom plots illustrate trends in error recovery strategies over the four-week study period.

identification rates of at least 50%, often higher. (Fig. 5). We captured participants' immediate verbal and vocal responses to Alexa's errors using our recording device (Table 6); overall, participants reacted verbally or vocally to 13.77% of errors ($n = 87/632$ errors).

Impact of error type on user reactions. Speech and VA errors prompted more than double the participant reactions (15.6% and 14.3%, respectively) compared to human errors (6.7%), as shown in Table 2. Alexa's unintentional activation's triggered reactions 40% of the time, suggesting an element of surprise. Errors such as partial listening, transcription, and intent had higher rates ($> 18\%$) due to leading to incorrect actions or information, while system, not handled, not listened, and limitation errors ($< 15\%$) typically involved Alexa not responding or issuing disclaimers, indicating that incorrect information elicits more reactions due to its perceived higher impact.

Articulating dissatisfaction through negative responses. The most frequent type of reaction observed was negation, occurring 27 times; this involved participants saying "no" once or repeatedly in response to an error. Participants also took immediate action, such as interrupting Alexa in 21 instances. Beyond negation and interruption, participants expressed negative comments toward Alexa ($n = 10$), reflecting frustration with its performance, as seen in C2 (Table 5), where P2's frustration grew with repeated interaction errors.

Forgiving attitude reflected by positive or neutral reactions. Instances of laughter were recorded 14 times and there were seven acknowledgments of incorrect responses from or actions taken by Alexa—for example, saying "okay" or "hmm" as neutral or positive expressions. However, some participants responded more positively to errors; expressions of gratitude (e.g., "Thank you") and positive comments, reflecting a forgiving attitude toward the VA (see C3 in Table 5).

4.1.4 Recovery strategies. User-initiated recovery. Participants employed strategies such as repeating or reformulating queries, providing clarification, and correcting the wake word to recover from errors (Fig. 6). Repetition was the most prevalent strategy ($n = 373$) that participants used for VA (e.g., intent recognition and not-handled) and *speech* (e.g., listening and transcription) errors. These findings align with prior research [Mavrina et al., 2022, Myers et al., 2018].

VA-initiated recovery. In instances when the VA itself initiated the recovery sequence—particularly in cases where the participant's query was incomplete and the system prompted them to clarify their request—the resolution rate was notably higher (Fig. 6, bottom) as opposed to when the user initiated recovery themselves. Alexa initiated recovery when there was missing information (partial-listening), misunderstanding (intent error), or action not available (limitation). For example, Alexa responded to an error by providing alternative action, "I am only able to clear your Shopping list, not delete it. Would you like me to clear the 9 items on it?" However, Alexa only initiated recovery in less than 5% of errors (29/632), whereas user-initiated recovery took place in 65% of errors.

Recovery strategies over time. We observed no distinct changes in recovery strategy patterns or error resolution rates over the span of four weeks, further confirming that user interaction patterns did not change over time.

4.1.5 Compound errors and user retries to fix conversational breakdowns. While looking at the sequential nature of errors, we observed that interaction errors often compounded, with participants typically attempting at least one recovery effort (Fig. 7). While for 108 errors participants did not attempt a retry, in 238 instances they made at least one attempt to correct the error; of these, 107 errors (44.96%) were resolved after the first retry, 43 instances (15.19%) led to participants giving up, and in 88 cases (36.97%) users attempted additional retries (Fig. 7). In contrast to our findings in Sec. 4.1.1 that only 25.32% (160/632) of errors were eventually resolved, we now shift our focus to compound errors: errors emerging from participants' reattempts to rectify the original error. By categorizing these compound errors as a single error instance, we discovered that the actual count of distinct errors stands at 346; this figure is the aggregate of both resolved ($n = 160$) and

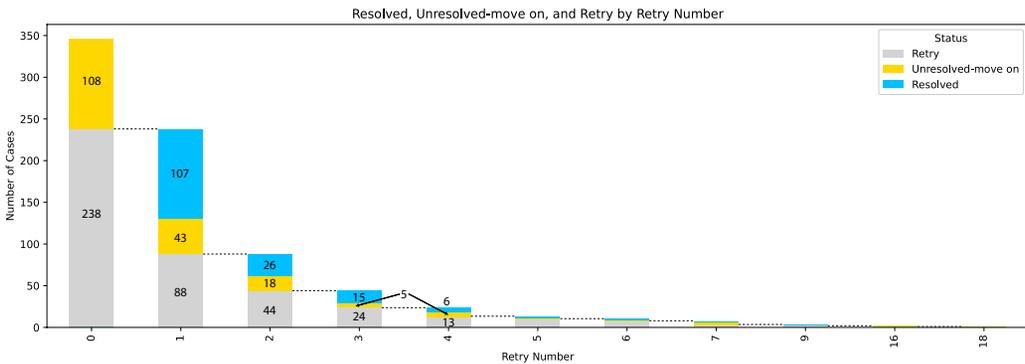


Fig. 7. Error recovery retries and their resolution. Each bar represents the number of errors present at a given retry level (x-axis), where 0 indicates the initial occurrence of the error and 1 indicates one user retry. The bar then shows whether those errors were resolved, remained unresolved, or were retried again. Queries that were retried could either be resolved or carried forward as unresolved errors in the next turn and so on. Across all retries, there were 160 resolved (blue: $107 + 26 + 15 + \dots$) and 186 unresolved errors (yellow: $108 + 43 + 18 + \dots$), totaling 346 compound errors out of 632 erroneous one-turn interactions—the remaining 446 instances represent retries (gray: $238 + 88 + 44 + \dots$). *The number of errors and successful resolutions decreased with each additional retry, indicating that most recoveries occurred within the first few attempts.*

unresolved errors ($n = 186$). This adjustment in our calculations indicates a greater level of user resilience and persistence in attempting error resolution, as we recalculated the actual resolution rate to be higher, at 46.24% ($160/346$) in contrast to the previously noted 25.32% ($160/632$). Finally, Our analysis showed that while 30.92% of errors ($107/346$) were resolved with one retry (Fig. 7), many persisted despite multiple attempts.

4.2 Interactions with an LLM-powered VA

At the start of third week of the study, we introduced our ChatGPT+Alexa skill to explore older adults' adoption and initial perceptions of an LLM-powered conversational VA.

4.2.1 Barriers to adoption. We observed that only a few participants (P1, P3, P11, P15) interacted with the ChatGPT skill, primarily on or around the day of deployment. Participants reported challenges in accessing the skill as the reason they did not try it out; for example, P9 pointed out, “I don’t like the ‘skill’ part of it. It’s too complicated, too cumbersome,” and further found the idea of chatting with a machine unusual: “I don’t want to try to have an artificial conversation. It’s just repugnant to me.” highlighting their reluctance to interact with advanced conversational VAs.

4.2.2 Nature of questions asked. Participants asked a variety of questions about health ($n = 7$, example C6 in Table 7), history ($n = 6$), the VA ($n = 3$, e.g., “I wanna know your age”), specific places ($n = 2$), stories ($n = 2$), and recipes ($n = 1$, C5 in Table 7), as well as two other questions, shown in C7 in Table 7.

4.2.3 Conversational breakdowns. Compared to traditional information-seeking interactions with Alexa alone, the ChatGPT skill resulted in a lower percentage of errors (Fig. 4) with a higher resolution rate of 14.1% versus 8.7%. Most errors stemmed from the integration of ChatGPT into the Alexa skill and involved system issues with listening, transcription, and intent recognition for activating and exiting the skill. Table 8 details the distribution and resolution of errors in the ChatGPT-skill interactions. All intent recognition errors were related to difficulties in exiting the

Table 7. Example conversations with the ChatGPT+Alexa skill. "... " represents the truncation of a conversation for brevity.

Tag Participant	ChatGPT+Alexa skill
C5 What's the best way for—to make hard boiled eggs? P1b: I take a nap. [P1b thanked Alexa afterwards.]	Searching . . . [filler + small talk] How do you unwind? Here it is. To make hard-boiled eggs, follow these steps: 1. Place your desired number of eggs . . .
C6 I'm having trouble falling asleep. P15 I take medicine but it doesn't help. What should I use to help me fall asleep? Thank you—a lot of good information. Why should I keep the room dark [cut off by Alexa] to help with sleep? Alexa, why should I keep the room dark to help with sleep? [Not listened by Alexa] Okay, thank you. Ummm, nothing else, thank you.	Here it is. I'm sorry to hear that you're having trouble falling asleep. Here are a few techniques that might help: 1. . . . Make sure your bedroom is quiet, dark, and at a comfortable temperature . . . What else can I help you with? Searching . . . [P15 did not wait for the VA's response and repeated herself.] Here it is. Keeping the room dark can promote better sleep because darkness signals to your body that it's nighttime and time to rest. . . What else can I help you with? [Intent error: cannot detect intent to end] Sorry, I don't know about that. Is there something else you would like to know?
No, nothing. [P15 interrupted Alexa] Alexa, stop.	[Intent error: cannot detect intent to end] Here it is. If you feel that none of the suggestions are helpful or applicable to . . . Goodbye! [Scripted end intent worked]
C7 Is the extreme heat, ummm, natural to weather? P15 Okay, thank you. Hmmm, can you tell me a story? Can you tell me a story? [repeat] Umm, thank you. Good story. I have a question: What's good music to help me relax?	Here you go. Extreme heat is a natural occurrence in weather patterns. It can be caused by various natural factors such as high-pressure. . . What else can I help you with? [Not handled error] Any other questions for me? Here it is. Certainly! Here's a short story for you: Once upon a time, in a small village nestled among rolling hills, there lived a . . . You're welcome! I'm glad you enjoyed the story. If you have any more requests or questions, feel free to ask. Here you go. The choice of relaxing music varies from person to person, as everyone has different preferences. However, here are a few . . .

skill (e.g., C6 in Table 5) and were resolved successfully. Partial listening errors ($n = 3$), often caused by a participant taking longer to formulate their queries, were not resolved as participants did not make any reattempts to fix them; P15 also mentioned a perceived speed requirement to interact with the ChatGPT skill: *"I think I wasn't thinking fast enough."*

4.2.4 Learning curve. Successful interactions with the ChatGPT skill (C5–C7 in Table 7) illustrated a learning curve, especially in managing multiple interaction layers, such as initiating questions, asking follow-ups, and closing the conversation. For instance, in C6, P15 initially faced challenges navigating a conversation with the ChatGPT-powered Alexa skill; however, with practice, P15 was able to engage in a much more fluid interaction, demonstrating adaptability and learning over time. Despite participants' initial successes, repeated interactions with the ChatGPT-powered skill were limited during the study which can be attributed to needing an additional step to access the skill.

Table 8. Errors in ChatGPT+Alexa skill interactions. **Resolved** represents number of errors resolved on first retry.

Individual	Resolved	Observations on how errors compounded
Intent	3/5	All intent errors stemmed from difficulties exiting the Alexa skill, which we attribute to the skill rather than ChatGPT. Participants attempted a few retries—two retries for two of the errors and one retry for the last error—eventually resolving all intent errors.
System	3/4	System errors compounded but were generally resolved on the second try. One system error remained unresolved as the participant chose to move on without retrying.
Partially listened	0/3	None of the errors resulting from partial listening were resolved.
Not handled	1/3	Only one of the errors was resolved in one retry.
Transcription	1/3	Two transcription errors compounded, but were resolved on the second retry; however, another of these errors remained unresolved as the participant did not attempt a retry.
Interrupted	1/1	Upon retry the error was fixed.
Limitation	0/1	The participant did not retry to fix the error.

4.3 Older Adults' Perceptions of Erroneous Interactions

Thematic analysis of the coded interview data revealed the following themes.

4.3.1 Underlying causes of errors. Participants identified the VA's failure to understand their requests as the most common cause of errors ($n = 28$). This was followed by speech-related challenges, such as the VA not allowing sufficient time for users to respond and interrupting them mid-sentence ($n = 13$). Other reported issues included users being unsure of what actions to take ($n = 6$), confusion about the VA's intentions or functionality ($n = 3$), and forgetting what to say or how to phrase their commands ($n = 3$).

4.3.2 Blame assignment. While participants mentioned being frustrated ($n = 6$) and confused ($n = 3$) when errors occurred, they internalized the blame for errors ($n = 8$; P14: "I could have been stumbling because I do that sometimes."). This tendency largely stemmed from their awareness of having different speech patterns. In contrast, only a few participants externalized the blame ($n = 5$; P9: "I think my thought would be, this is a stupid device.").

4.3.3 How they fixed errors. The most common error-recovery method was reformulating queries ($n = 18$), often multiple times, though these attempts were not always successful, as corroborated by the interaction data. For example, P15 mentioned: "She doesn't understand and you have to be more creative and ask the question in a different way." In some cases, participants relied on family or friends for assistance with errors and debugging ($n = 2$).

4.3.4 Hindrance to use. Participants mentioned the inability to find useful functionalities for the VA ($n = 7$) and the need to put in deliberate effort to integrate it into their daily lives ($n = 7$) as significant hindrances to use. They also emphasized the challenges of needing assistance and time to adapt to new technology ($n = 5$), along with not knowing how to perform tasks independently ($n = 7$), as contributing factors.

4.3.5 Expectations from the VA. Participants expressed the need for assistance in understanding the VA's functionality and capabilities to use it effectively ($n = 9$). They desired ongoing support, both in learning what and how VAs can do tasks and within interactions, such as guidance on reformulating queries or handling errors. For instance, P11 highlighted expectations for the VA to

Table 9. Design considerations (DCs) mapped out across various themes of user-VA interaction.

Main Theme	Sub-Theme	Corresponding Design Considerations
Challenges faced by older adults	General VA challenges	DC1, DC8
	Forgetfulness	DC2
	Speech difficulties	DC3, DC6
	Varying familiarity and expectations	DC4, DC7
	Varying conversation styles	DC5
Error categories	Automatic speech recognition errors	DC3, DC6
	VA limitations and human errors	DC4, DC7
	Intent recognition errors	DC5
Error management steps	Avoiding errors	DC2, DC3, DC4, DC5, DC6, DC7
	Error detection	DC1
	Recovery and repair	DC1, DC9
Implementation-level	Voice user interface	DC3
	Model	DC5 (Modeling vocal and verbal cues) DC6 (Using LLMs)
	Interaction	DC1, DC9 (Mitigation), DC2 (Proactive), DC2, DC4, DC7, DC9 (Educational)

proactively seek clarification: “*And I think, I think what... could be better maybe is... if Alexa wasn't clear about what the question is, it could come back and say, can you clarify, you know, are you asking for what was the question just now about?*”

5 Discussion

Our study delves into the nuanced dynamics of older adults' interactions with smart speaker-based VAs using real-world audio data. We investigate the nature of conversational breakdowns, older adults' difficulty in recovering from errors, and the need for multiple attempts and rigid query formulation. In this section, we pinpoint missed opportunities in developing resilient VAs, highlight age-specific challenges leading to speech recognition errors, and discuss the mismatch between older adults' expectations and VAs' limited capabilities. We further discuss how LLM integration into VAs would change interaction dynamics for older adults. We offer design considerations for VAs tailored to older adults (Table 9).

5.1 Designing Context-Aware, Resilient VAs for Improved Error Management

Current VAs heavily rely on users to initiate recovery upon errors (65%), yet our findings indicate that most errors remain unresolved. More importantly, learning by experience over time does not decrease error rates or improve error resolution, even for straightforward tasks (Fig. 4), nor do users effectively learn over time to recover from errors using methods such as query reformulation (Fig. 6). This temporal trend underscores the need to design robust error management mechanisms that shift the burden of recovery from the user to the VA.

Our research indicates that 80% of errors in VA interactions can be identified from participants' immediate verbal reactions (e.g., negation, remarks, and comments) or by their attempts at recovery (e.g., reformulating a query), as illustrated in Fig. 5. However, such valuable feedback is not effectively leveraged by current commercial VAs—occurring in less than 5% of cases despite users' expectations that the system should ask for clarification when needed (Sec. 4.3.5)—even though they may capture immediate user reactions and follow-up queries. Alexa's follow-up mode captures a portion of users' immediate responses but often fails to categorize them accurately—labeling many as “*Audio*

not intended for Alexa” or simply responding, “Sorry, I don’t know about that.” The system does not utilize these cues for error detection, a crucial step for initiating self-repair (VA-initiated recovery), which has been shown to improve users’ assessment of VAs [Cuadra et al., 2021]. However, excessive correction can degrade user perceptions [Cuadra et al., 2021]. Therefore, effectively incorporating users’ immediate reactions could improve error detection to avoid over-correction. Once an error is identified and corrected, the VA can engage in verbal trust repair (e.g., by apologizing for the mistake [Mahmood et al., 2022]) to further enhance the recovery process.

Opportunity 1: *Users react, respond, or attempt to fix most errors made by VAs.*

Design Consideration 1: *Refine VAs to better utilize users’ immediate vocal and verbal reactions and responses for improved error detection, error recovery, and trust repair.*

5.2 Understanding and Addressing Suboptimal Interactions Due to Age-Related Factors

VAs are often not designed for older adults [Stigall et al., 2019], leading to challenges and errors unique to this population. Limited technological familiarity, cognitive decline, and speech difficulties contribute to speech recognition failures. Below, we discuss how these challenges are reflected in various speech recognition failures (partial listening, transcription, and wrong wake word errors) and propose design considerations to enhance the voice assistance experience for older adults.

5.2.1 Forgetfulness. In exploring older adults’ continued use of VAs, a significant limitation we observed was their forgetfulness, which manifested in three principal ways. First, participants often **forgot the VA’s utility—or even its presence in the home**. This could potentially be due to age-related cognitive decline; however, more importantly, it reflects a mismatch between the design of current VAs and the everyday routines and expectations of older adults (Sec. 4.3.4). Because these systems are not well integrated into their existing habits or designed to support ongoing, meaningful use, they are more easily forgotten. Similar findings have been reported in prior work on older adults’ interactions with VAs [Trajkova and Martin-Hammond, 2020]. While the current reactive nature of VAs necessitates user initiation, our findings show increased engagement when users are reminded of the VA’s presence (Fig. 3, supplementary materials³). Such brief usage spikes suggest a potential benefit of transitioning toward more proactive interactions based on user routines—for example, proactively reminding users of daily tasks or medications they might otherwise forget. For instance, since P15 habitually set medication reminders almost every day for 6 p.m., the VA could proactively suggest or set daily reminders, addressing forgetfulness common among older adults [Pradhan et al., 2020]. However, proactive interactions should carefully consider social and environmental contexts to preserve user agency, clearly stating intent and seeking permission when appropriate [Zargham et al., 2022]. Future research should further investigate balancing reactive and proactive interactions through mixed-initiative approaches.

Secondly, we observed that participants mistakenly **used the incorrect wake word** “Alexis” instead of “Alexa,” leading to conversational breakdowns as Alexa did not respond despite detecting the similar-sounding trigger word (Sec. 4.1.1). Such human error—often a result of forgetfulness—could potentially be addressed by system-generated messages and reminders such as “*Did you mean to say ‘Alexa?’ I’m not certain if you were addressing me.*” upon error detection, rather than the VA simply remaining unresponsive. However, the practice of the system assuming the correct wake word when it has not been clearly articulated may be perceived as intrusive, raising concerns about user privacy.

The third observed aspect of forgetfulness pertains to older adults’ **inability to recall the purpose of interactions**, such as setting reminders. Participants often found themselves forgetting and questioning the intent behind an alert, asking, “What am I supposed to do?” when a reminder rang without an accompanying message, indicating a mismatch between the VA’s functionalities

and their needs. To address this, VAs should be designed to solicit more detailed information when setting reminders or alarms, thus making them more effective and tailored to their users' needs. Our findings indicate that verbal and vocal reactions can signal potential breakdowns or confusion (Section 4.1.3). Prior work similarly shows that older adults' speech during such moments often includes remarks, negations, question words, or negative sentiment, accompanied by higher loudness, pitch, and speech rate [Fan et al., 2021]. These cues can serve as triggers for proactive system engagement to resolve issues.

Limitation 1: *The reactive nature of voice assistance fails to accommodate the forgetfulness of older adults.*

Design Consideration 2: *Design VAs to proactively suggest actions and prompt older adults for further details to help them formulate their queries, rather than merely reacting to their requests.*

5.2.2 *Speech difficulties and challenges.* While most errors faced by the older adults in our interaction data stemmed from the VA's failure to recognize user intent or fulfill requests, a considerable portion of errors ($n = 154$) were related to automatic speech recognition technology, such as the partial listening and transcription errors. Moreover, these errors exhibited low resolution rates (31.8%) without noticeable improvement over time. Such speech recognition inaccuracies are particularly prevalent among older adults [Kobayashi et al., 2019], who may take longer to formulate their queries due to unfamiliarity with technology and speech difficulties associated with aging—pointed out by our participants in Sec. 4.3.1—, suggesting the need for better accommodations for users with slower speech patterns and speech difficulties. Since future VAs are envisioned to be more conversational, it is crucial for these systems to allow older users sufficient time to process new information and articulate their queries; such adaptations would make voice user interfaces (VUIs) more responsive to the needs and communication styles of older adults, enhancing their overall experience.

Limitation 2: *VUIs have limited adaptability to the slower speech patterns of and extended speaking time often required by older adults, even as VAs are becoming more conversational.*

Design Consideration 3: *Tailor VUIs to better suit the pace of older adults, ensuring fewer speech recognition errors.*

5.3 Aligning Older Adults' Expectations with VAs' Functional and Comprehension Abilities

Our findings show that participants often misestimated VA capabilities, leading to limitation errors. The VA also struggled with participants' natural speech patterns, causing intent recognition failures (Sec. 4.3.1). Continuous advancements in VA capabilities require recalibrating older adults' mental models for effective use. We discuss and present design considerations to adjust older adults' mental model and improve VAs' understanding of user intentions.

5.3.1 *Inaccurate mental models of VA capabilities.* Our study reveals notable discrepancies in participants' understanding of VA capabilities, leading to inaccurate mental models. The high incidence of limitation errors ($n = 130, 20.6\%$)—where the VA failed to understand the user's intent and responded, “*I don't know about that*” (Table 2)—illustrates that Alexa often failed to fulfill older participants' requests. The increased occurrence of limitation errors thus suggests that participants may not have been fully aware of the VA's capabilities. Additionally, certain VA interaction types exhibited notably higher error rates, particularly in less frequent “edge” categories, such as miscellaneous queries, TV-related commands, or asking for stories and poems. This trend extends to high error rates in less common queries about health, famous personalities, and other topics, exemplifying the “long-tail problem”—a significant machine learning challenge where infrequent events hinder response accuracy. Users expect consistent responsiveness across all

Table 10. Additional illustrative conversations. “...” represents the truncation of a conversation for brevity.

Tag	Participant	Alexa
C8	Alexa, play the history of Tony Bennett.	[P11b to P11a:] What’s the history of Tony Bennett?
P11a	[To P11b:] Oh, she’ll tell you. [To P11b:] She didn’t tell you.	[P11b chuckles mockingly] [P11b to P11a:] Because it makes no sense.
C9	Alexa, give me a 10-minute warning.	10 minutes, starting now. [successful initiation]
P11a	[P11b to P11a, overlapping with Alexa’s speech:] What? Wait, wait, what, [P11b’s name?] Alexa, set timer for 10 minutes.	[P11b to P11a:] No, you need to say “timer for 10 minutes.” [P11b to P11a:] Set a time for 10 minutes. Second timer, 10 minutes, starting now.
C10	Alexa, can you keep a recipe for me?	Sure, I can search by dish name and ingredients. What would you like?
P1b	No, I wanna keep one of my own recipes. Never mind.	Sure, I can search by dish name and ingredients. What would you like? [not listened by Alexa]

query types, akin to human conversations [Cowan et al., 2017], but due to limited training on these “long-tail” cases, VAs struggle with less frequent queries. This mismatch between user expectations and VA capabilities highlights the need for VAs to either enhance their adaptability to a wider range of queries or to more clearly communicate their limitations, thereby aligning users’ mental models more closely with actual system performance.

The disconnect between users’ mental models and a VA’s capabilities becomes more apparent upon examining VA interactions between couples with divergent mental models. For instance, in the case of P11, the couple demonstrated differing perceptions: P11a overestimated the capabilities of Alexa (C8, Table 10 and C3, Table 5), whereas P11b underestimated them (C9, Table 10). Instances of overestimation were also observed in individual participants, such as when P2 asked the VA to remember their conversations (C2, Table 5). Even experienced users still misunderstood VA capabilities; for example, participant P1b tried to control non-smart lights via Alexa and habitually commanded it to “stop” talking unnecessarily. Such overly cautious behaviors—also observed in new users—suggest a common underestimation of VAs, likely shaped by past erroneous interactions.

Previous research indicates that users often rely on trial and error to develop their mental model of a VA [Kim and Choudhury, 2021]—however, the mental model mismatch we found in “experienced” users, coupled with no noticeable improvement in error rate or recovery success over the span of this study (even for straight-forward functional tasks), suggests that this method might not always correct users’ mental models; instead, it tends to reinforce behaviors that “work” rather than those that are more effective and accurate. Providing accurate, clear information is thus crucial, especially for older adults who may struggle with technology (Sec. 4.3.5).

VAs should also be designed to detect user misconceptions and respond with context-specific explanations rather than generic messages such as, “Sorry, I don’t know that” [Kim and Choudhury, 2021, Pradhan et al., 2020]. For example, if a user tries to control a non-smart appliance, the VA should clarify prerequisites for smart-home integration. Tailored responses help educate users about VAs’ capabilities and limitations, accommodating diverse technological familiarity, particularly among older adults.

Limitation 3: *A VA’s inability to properly convey its capabilities results in continual inaccuracies in older adults’ mental models—even amongst long-time users.*

Design Consideration 4: *VAs should employ context-aware responses with dynamic tutorials and real-time guidance in daily interactions to better align older adults' mental models with evolving voice assistance technology.*

5.3.2 *Lack of VA comprehension of users' natural conversation methods.* Our findings reveal significant gaps in VAs' comprehension of natural speech, evident from high intent recognition errors (32.3%, $n = 204$ of 632; Sec. 4.1.1), aligning with participants' perceptions (Sec. 4.3.1). Participants often preferred intuitive and polite conversational phrases (e.g., "Thank you") over direct commands (e.g., "Stop"), aligning with older adults' general preference for politeness in interactions with VAs [Hu et al., 2022, Pradhan et al., 2019]. This preference for intuitive interactions was illustrated when participants repeatedly used natural language (e.g., "Thank you, I have taken my medicine") instead of specific commands (Table 5, C4). Such intent recognition failures become critical in high-stakes scenarios such as medication reminders; for instance, participant P15 struggled to inform Alexa upon medication intake, resulting in frustration ("I don't know what I am supposed to say to it..."), even though the VA could track reminder completion. Older adults find rigid keyword requirements frustrating [Pradhan et al., 2020, 2018].

Our data demonstrate that older adults adopt human dialogue as a metaphor for interaction with VAs [Cowan et al., 2017, Doyle et al., 2019], which can be attributed to two factors: 1) speech interfaces are increasingly designed and marketed to emulate the nuances of human speech [Gilmartin et al., 2017, Harris, 2004] and 2) speech alone as interaction modality may inherently provide a false affordance for possible interactions—essentially tricking people into believing that natural interactions are possible, while in reality, most VAs are only able to carry out structured human-machine interactions. These factors often lead to an overestimation of VAs' abilities resulting in communication breakdowns [Leahu et al., 2013, Luger and Sellen, 2016, Moore et al., 2016]. Furthermore, we observe that such misattribution of humanness to VA interactions is challenging to correct through trial and error, particularly among the aging population. This underscores the need for a more intentional design of voice-based interaction conventions—*i.e.*, *natural human interactions vs. structured human-machine interactions*—and for more appropriate introduction and marketing of VAs.

Current commercial VAs require a structured human-machine interaction style (*i.e.*, scripted commands and clear queries), reinforcing the mismatch between user mental models and VA capabilities (Sec. 5.3.1). Older adults frequently revert to a natural, conversational style, reflecting their preference for more social and informal interactions compared to younger [Oh et al., 2020] and middle-aged adults who adapt more easily to structured commands [Chin et al., 2024]. Adapting VAs to accommodate natural conversational patterns—especially considering older adults' age-related challenges (Sec. 5.2) and slower technology adoption [Czaja et al., 2006, Vaportzis et al., 2017]—is essential to improving their user experience and communication effectiveness.

Therefore, intent recognition should be *human-* and *context-aware*, considering older adults' expectations for socially capable VAs, which vary within their demographic [Harrington et al., 2022]. Leveraging verbal and vocal social cues, along with other contextual information, can improve implicit intent recognition, while ignoring such context leads to ineffective responses [Cuadra et al., 2022].

Limitation 5: *VAs' lack of comprehension of the naturalistic and diverse conversational styles practiced by older adults leads to intent recognition failures and subpar user experience.*

Design Consideration 5: *Design VAs to be human- and context-aware by detecting implicit intent through verbal and vocal cues.*

5.4 Integration of LLMs into VAs for Older Adults: An Interaction Challenge

In this section, we discuss how the shift from traditional voice assistants to more advanced, LLM-powered ones changes the interaction dynamics for older adults, particularly in how they experience and manage errors. While LLMs substantially reduce traditional technical failures, they also introduce new interactional challenges. These challenges are especially salient when compared to younger adults, who often hold more adaptive mental models of emerging VAs.

5.4.1 Persistent and emergent interaction challenges in LLM-powered VAs. LLM integration improves natural language understanding, contextual awareness, and multi-turn dialogue [Mahmood et al., 2025c]. In our study, the ChatGPT+Alexa skill produced smoother conversational flow and more accurate intent recognition than standard Alexa; for instance, P15 issued a follow-up (“*Why should I keep the room dark?*”) after sleep advice (C6, Table 7), indicating stronger contextual continuity. Errors were often resolved within one or two retries (Sec. 4.1.5). Moreover, we observed cases where LLMs absorbed imperfect inputs and still advanced the task (e.g., C6’s truncated utterance due to VA interruption), consistent with prior findings that LLMs can mitigate recognition issues or trigger recovery prompts [Mahmood et al., 2025c].

However, for older adults these gains were constrained by interactional factors. For instance, partial listening more frequently resulted in breakdowns than in our prior study with younger adults (about 9% vs. nearly all in this study), potentially due to older adults’ slower or fragmented speech [Kobayashi et al., 2019], which hindered capture of enough context. This finding should be viewed conservatively since the number of interactions in our study was small.

Activation and exit also posed friction: initiating or leaving the ChatGPT skill (Sec. 4.2.3) often failed. In contrast, younger adults encountered activation confusion less frequently (about 6% of interactions compared to 20% or 4/20 for older adults) and typically recovered quickly [Mahmood et al., 2025c]. Overall, what appear to be technical problems are interactional—rooted in speech pacing and style, and repair dynamics that differ by age group.

Opportunity 2: *LLMs show potential in absorbing speech recognition errors—transcription inaccuracies—to further reduce conversational breakdowns.*

Limitation 6: *LLM-powered VAs still struggle to adapt to older adults’ slower or fragmented speech, leading to frequent loss of context and conversational breakdowns.*

Design Consideration 6: *Leverage LLMs’ contextual awareness to reduce speech recognition errors and accommodate older adults’ unique speech patterns and characteristics.*

5.4.2 Mental models and transparency needs. Our findings reveal that older adults’ difficulties with LLM-powered VAs extend beyond speech or activation issues to include gaps in understanding how the system operates and communicates. Many participants treated the VA as a fixed-function device, unable to distinguish between standard and LLM-enabled capabilities (e.g., requiring “*Alexa, let’s chat*” before asking open-ended questions). This mental model contrasts sharply with that of younger adults in our previous work [Chan et al., 2023, Mahmood et al., 2025b,c], who more readily understood the integration of LLMs and adapted their interactions accordingly. For older adults, these new conversational affordances often appeared unpredictable or “too human,” creating uncertainty about what the system could or should do, thus leading to mental model mismatches similar to the ones discussed in Section 5.3.1.

The phrasing “*Let’s chat*” and describing the VA to be able to sustain more natural dialogue also contributed to misperceptions such as perceiving VAs as conversational partners that are more competent, reliable, human-like, and flexible on a partnership scale [Doyle et al., 2025]. Consequently, older adults, potentially perceiving the VA higher on the partnership scale, reduced their willingness to intervene or attempt recovery. In contrast, younger adults, who tend to perceive

such systems lower on the participation scale, compensated more actively by initiating recovery behaviors [Doyle et al., 2025, Mahmood et al., 2025c]. While younger adults' familiarity with technology helped them form more accurate expectations, older adults often relied on surface-level cues and framing rather than underlying technical knowledge. This may also explain their psychological resistance to "chatting" with a machine (P9: "I don't want to try to have an artificial conversation. It's just repugnant to me"), as they tend to perceive it as a social rather than a need-based, functional task. Because older adults depend more on explicit signals than tacit understanding, transparency and timely feedback become essential for calibrating expectations (Section 5.3.1).

Limitation 7: *Mental model mismatch can be caused by how the VA is presented or portrayed to the user.*

Design Consideration 7: *Provide transparent information and real-time cues about VA capabilities and limits; favor concise, task-oriented messaging over anthropomorphic framing.*

Another consideration is matching VA capabilities to older adults' expectations and needs (Section 5.3.2), such as being natural and intuitive. However, VAs—even those powered by LLMs—often remain on the lower end of the partnership scale due to unintentional or superficial integration. Recent work with purposefully designed LLM-powered VAs for older adults has shown that intentional design for both general assistance and specific task support is more effective and well-received [Liu et al., 2025, Mahmood et al., 2025b, Rudnik et al., 2024, Yang et al., 2024]. For instance, designing systems that support interruptions and backchannels can create more natural, engaging, and fluent conversations [Liu et al., 2025]. Therefore, future work should focus on designing VAs to fit into older adults' contexts meaningfully.

Limitation 8: *Unintentional design of VAs creates interaction challenges for older adults.*

Design Consideration 8: *Actively involve older adults and other special populations in the design of LLM-powered VAs to ensure that they align with their expectations and needs.*

5.4.3 Supporting gradual learning and proactive recovery. Our preliminary analysis revealed a gradual learning curve: participants' later interactions showed more coherent question phrasing and improved follow-up behavior with the LLM-powered VA (e.g., P15's C5 vs. C7, Table 7). Yet for older adults, learning emerged primarily through trial and error. Older adults were also less motivated to retry or troubleshoot errors, often waiting for the VA to recover on its own. This behavior was amplified by a lack of transparency and feedback—participants reported confusion about what went wrong, whether the VA was still processing, or how to correct an error. Learning through experience alone proved insufficient (Section 5.3.1); while participants showed some improvement over time, progress often stalled when the VA offered no explicit recovery cues. In contrast, younger adults more frequently self-corrected, likely due to more appropriate understanding of LLM-powered VAs [Mahmood et al., 2025c]. However, even younger adults sometimes faced "uncommunicated affordances," (e.g., a user wondering: "Oh no, do I have to start again?") [Chan et al., 2023, Hwang et al., 2023]). These findings highlight that even as LLMs make VAs more capable, conversational success depends on mutual intelligibility: users must understand the VA's limits, and the system must anticipate users' confusion. Proactive recovery behaviors—such as briefly explaining missteps, suggesting rephrasings, or soliciting clarifications—can bridge this gap.

Limitation 9: *Lack of clear interaction cues during error recovery can heighten conversational breakdowns and contribute to user frustration.*

Design Consideration 9: *Implement proactive, context-sensitive feedback and repair mechanisms to support gradual learning and prevent frustration.*

In summary, while LLM integration reduces technical understanding errors, it surfaces interaction-level challenges that are particularly salient for older adults. Designing for *effective interactions*—not only using better models—will be essential to ensure LLM-powered VAs are usable, trustworthy,

and inclusive across age groups. Thus, our findings offer implications for future generations of voice assistants.

5.5 Limitations and Future Work

While our four-week study provided insights into the behavior of older adults toward VAs, especially during interaction breakdowns, extended studies spanning several months could provide a deeper understanding in this area. Longer-duration interaction trends could additionally serve as a rich data source for training VAs to better recognize and recover from errors. Another limitation of this study is that the presence of a recording device may have influenced participant behavior. Although participants may have become less aware of the device over time, its presence could still have affected their usage, reactions, or responses, consistent with a potential Hawthorne effect, and therefore the study's findings. As such, the results should be interpreted with this consideration in mind. Currently, our analysis focuses qualitatively on queries and responses to assess their potential utility; however, a more detailed query analysis [Brewer et al., 2022] would offer further insights into the nuances of older adults' interactions with VAs.

In this study, we investigated a generic application of LLM-powered VAs; future research should consider specific use cases, such as health care information retrieval or support tailored for older adults with disabilities and impairments. Moreover, our findings should be interpreted in light of the specific demographic context of our participants—predominantly healthy older adults without speech impairments, living in the Baltimore, Maryland area. Older adults from different cultural contexts, educational levels, or with varying degrees of technological familiarity may exhibit distinct patterns of engagement, trust, and error recovery. While our participants showed a range of familiarity with voice assistants, a more systematic analysis is needed to better understand how such factors influence usage behaviors. Serious health conditions can also shape technology use; for instance, speech impediments may exacerbate conversational breakdowns, while cognitive decline could necessitate additional education, guidance, and ongoing reminders. Future studies should therefore adopt a cross-cultural and cross-experience perspective to examine how these dimensions impact interaction challenges and inform design considerations for diverse older populations.

6 Conclusion

Traditional VAs often fail to effectively utilize immediate user reactions, users' explicit corrective actions, and conversation history in assisting users or recovering from errors. In our longitudinal study, we collected immediate user responses and reactions via a supplemental audio recorder, demonstrating that such social cues can often lead to increased error identification. Additionally, we found that VAs' current requirements for scripted commands and well-formed queries exacerbate conversational breakdowns with older adults, which may be caused by their inaccurate mental models of VAs' capabilities, their tendency to forget more effective query styles, and a preference for more naturalistic communication. We also explored integrating an LLM into an Alexa skill to enhance VAs' conversational capabilities with older adults; our findings provide insights and considerations for the design and research of nuanced interaction dynamics between older adults and futuristic conversational VAs, emphasizing the value of "in-the-wild" data collection to better comprehend such relationships.

Acknowledgments

This work was supported by the National Science Foundation award #1840088, the Johns Hopkins University Institute for Assured Autonomy, and Malone Center for Engineering in Healthcare. We thank Jaimie Patterson for her feedback and assistance in this work.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to decrease repetition and improve readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of this publication.

CRedit author Statement

Amama Mahmood: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - Original draft, Writing - Review & editing, Visualization, Project Administration.

Junxiang Wang: Formal analysis, Writing - Original draft.

Chien-Ming Huang: Conceptualization, Methodology, Resources, Writing - Original draft, Writing - Review & editing, Visualization, Supervision, Funding acquisition.

References

- Leonardo Angelini, Maurizio Caon, Emmanuel Michielan, Omar Abou Khaled, and Elena Mugellini. 2021. Seniors' perception of smart speakers: challenges and opportunities elicited in the Silver&Home Living Lab. In *Congress of the International Ergonomics Association*. Springer, 137–144.
- Anneliese Arnold, Stephanie Kolody, Aidan Comeau, and Antonio Miguel Cruz. 2022. What does the literature say about the use of personal voice assistants in older adults? A scoping review. *Disability and Rehabilitation: Assistive Technology* (2022), 1–12.
- Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.
- Tiago Carneiro Gorgulho Mendes Barros and Rodrigo Duarte Seabra. 2020. Usability assessment of google assistant and Siri virtual assistants focusing on elderly users. In *17th International Conference on Information Technology–New Generations (ITNG 2020)*. Springer, 653–657.
- Erin Beneteau, Olivia K Richards, Mingrui Zhang, Julie A Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication breakdowns between families and Alexa. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.
- Michael S Bernstein, Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. 2012. Direct answers for search queries in the long tail. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 237–246.
- Johnna Blair and Saeed Abdullah. 2019. Understanding the needs and challenges of using conversational agents for deaf older adults. In *Conference Companion Publication of the 2019 On computer supported cooperative work and social computing*. 161–165.
- Manuel Bolaños, César A Collazos, and Francisco L Gutiérrez. 2020. Adapting a Virtual Assistant Device to Support the Interaction with Elderly People. In *ICT4AWE*. 291–298.
- Ann Bowling, Matthew Hankins, Gill Windle, Claudio Bilotta, and Robert Grant. 2013. A short measure of quality of life in older age: The performance of the brief Older People's Quality of Life questionnaire (OPQOL-brief). *Archives of gerontology and geriatrics* 56, 1 (2013), 181–187.
- Robin Brewer, Casey Pierce, Pooja Upadhyay, and Leeseul Park. 2022. An empirical study of older adult's voice assistant use for health information seeking. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 12, 2 (2022), 1–32.
- Szeyi Chan, Jiachen Li, Bingsheng Yao, Amama Mahmood, Chien-Ming Huang, Holly Jimison, Elizabeth D Mynatt, and Dakuo Wang. 2023. "Mango Mango, How to Let The Lettuce Dry Without A Spinner?": Exploring User Perceptions of Using An LLM-Based Conversational Assistant Toward Cooking Partner. *arXiv preprint arXiv:2310.05853* (2023).
- Chen Chen, Janet G Johnson, Kemeberly Charles, Alice Lee, Ella T Lifset, Michael Hogarth, Alison A Moore, Emilia Farcas, and Nadir Weibel. 2021. Understanding barriers and design opportunities to improve healthcare and QOL for older adults through voice assistants. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–16.
- Thiago Silva Chiaradia, Rodrigo Duarte Seabra, and Adriana Prest Mattedi. 2019. Evaluating the usability of the Siri virtual assistant on mobile devices with emphasis on Brazilian elderly users. In *16th International Conference on Information Technology–New Generations (ITNG 2019)*. Springer, 437–441.
- Jessie Chin, Smit Desai, Shen 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. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–21.
- Yong K Choi, Hilaire J Thompson, and George Demiris. 2020. Use of an internet-of-things smart home system for healthy aging in older adults in residential settings: Pilot feasibility study. *JMIR aging* 3, 2 (2020), e21964.
- Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What can i help you with?" infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th international conference on human-computer interaction with mobile devices and services*. 1–12.
- Andrea Cuadra, Hyein Baek, Deborah Estrin, Malte Jung, and Nicola Dell. 2022. On Inclusion: Video Analysis of Older Adult Interactions with a Multi-Modal Voice Assistant in a Public Setting. In *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development*. 1–17.
- Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My bad! repairing intelligent voice assistant errors improves interaction. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–24.
- Sara J Czaja, Neil Charness, Arthur D Fisk, Christopher Hertzog, Sankaran N Nair, Wendy A Rogers, and Joseph Sharit. 2006. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). *Psychology and aging* 21, 2 (2006), 333.
- Philip R Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R Cowan. 2019. Mapping perceptions of humanness in intelligent personal assistant interaction. In *Proceedings of the 21st international conference on human-computer interaction with mobile devices and services*. 1–12.

- Philip R Doyle, Iona Gessinger, Justin Edwards, Leigh Clark, Odile Dumbleton, Diego Garaialde, Daniel Rough, Anna Bleakley, Holly P Branigan, and Benjamin R Cowan. 2025. The Partner Modelling Questionnaire: A validated self-report measure of perceptions toward machines as dialogue partners. *ACM Transactions on Computer-Human Interaction* 32, 4 (2025), 1–33.
- Mingming Fan, Qiwen Zhao, and Vinita Tibdewal. 2021. Older adults' think-aloud verbalizations and speech features for identifying user experience problems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
- Emer Gilmartin, Marine Collery, Ketong Su, Yuyun Huang, Christy Elias, Benjamin R Cowan, and Nick Campbell. 2017. Social talk: making conversation with people and machine. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents*. 31–32.
- Yijie Guo, Ruhan Wang, Zhenhan Huang, Tongtong Jin, Xiwen Yao, Yuan-Ling Feng, Weiwei Zhang, Yuan Yao, and Haipeng Mi. 2025. Exploring the Design of LLM-based Agent in Enhancing Self-disclosure Among the Older Adults. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
- Christina N Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
- Randy Allen Harris. 2004. *Voice interaction design: crafting the new conversational speech systems*. Elsevier.
- Aike C Horstmann, Till Schubert, Lea Lambrich, and Clara Strathmann. 2023. Alexa, I Do Not Want to be Patronized: A Qualitative Interview Study to Explore Older Adults' Attitudes Towards Intelligent Voice Assistants. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. 1–10.
- Yaxin Hu, Yuxiao Qu, Adam Maus, and Bilge Mutlu. 2022. Polite or direct? Conversation design of a smart display for older adults based on politeness theory. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
- Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.
- Alyssa Hwang, Natasha Oza, Chris Callison-Burch, and Andrew Head. 2023. Rewriting the Script: Adapting Text Instructions for Voice Interaction. *arXiv preprint arXiv:2306.09992* (2023).
- Manuel Jesús-Azabal, Javier Rojo, Enrique Moguel, Daniel Flores-Martin, Javier Berrocal, José García-Alonso, and Juan M Murillo. 2020. Voice assistant to remind pharmacologic treatment in elders. In *Gerontechnology: Second International Workshop, IWOG 2019, Cáceres, Spain, September 4–5, 2019, Revised Selected Papers 2*. Springer, 123–133.
- Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 18, 16 pages. doi:10.1145/3544548.3581503
- Brittne Kakulla. 2021. *2021 tech trends and the 50+ Top 10 biggest trends*. Technical Report. AARP, Washington, DC.
- Sunyoung Kim and Abhishek Choudhury. 2021. Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study. *Computers in Human Behavior* 124 (2021), 106914.
- Masatomo Kobayashi, Akihiro Kosugi, Hironobu Takagi, Miyuki Nemoto, Kiyotaka Nemoto, Tetsuaki Arai, and Yasunori Yamada. 2019. Effects of age-related cognitive decline on elderly user interactions with voice-based dialogue systems. In *Human-Computer Interaction—INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part IV 17*. Springer, 53–74.
- Jaroslaw Kowalski, Anna Jaskulska, Kinga Skorupska, Katarzyna Abramczuk, Cezary Biele, Wieslaw Kopeć, and Krzysztof Marasek. 2019. Older adults and voice interaction: A pilot study with google home. In *Extended Abstracts of the 2019 CHI Conference on human factors in computing systems*. 1–6.
- Lucian Leahu, Marisa Cohn, and Wendy March. 2013. How categories come to matter. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3331–3334.
- Jiaan Li, Yulu Hu, Shihan Fu, Zhiqing Wu, Pan Hui, Hailiang Wang, and Mingming Fan. 2024. 'It's like having a friend': Exploring the Opportunity of Large Language Models (LLMs) in Assisting Retirees In Their Daily Lives. In *Proceedings of the Twelfth International Symposium of Chinese CHI*. 463–475.
- Senior Lifestyle. 2022. Digital assistants for seniors. <https://www.seniorlifestyle.com/resources/blog/digital-assistants-for-seniors/>. Accessed: 2023-11-9.
- Jessica Lis. 2022. US Voice Assistants and Smart Speakers Forecast 2022. <https://www.insiderintelligence.com/content/us-voice-assistants-smart-speakers-forecast-2022/>. Accessed: 2023-11-9.
- Chao Liu, Mingyang Su, Yan Xiang, Yuru Huang, Yiqian Yang, Kang Zhang, and Mingming Fan. 2025. Toward Enabling Natural Conversation with Older Adults via the Design of LLM-Powered Voice Agents that Support Interruptions and Backchannels. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.

- Ying-Chieh Liu, Chien-Hung Chen, Yu-Sheng Lin, Hsin-Yun Chen, Denisa Irianti, Ting-Ni Jen, Jou-Yin Yeh, Sherry Yueh-Hsia Chiu, et al. 2020. Design and usability evaluation of mobile voice-added food reporting for elderly people: randomized controlled trial. *JMIR mHealth and uHealth* 8, 9 (2020), e20317.
- Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. 2019. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science* 51, 4 (2019), 984–997.
- Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5286–5297.
- Amama Mahmood, Shiye Cao, Maia Stiber, Victor Nikhil Antony, and Chien-Ming Huang. 2025a. Voice Assistants for Health Self-Management: Designing for and with Older Adults. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.
- Amama Mahmood, Jeanie W Fung, Isabel Won, and Chien-Ming Huang. 2022. Owning mistakes sincerely: Strategies for mitigating AI errors. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–11.
- Amama Mahmood, Bokyung Kim, Honghao Zhao, Molly E Atwood, Luis F Buenaver, Michael T Smith, and Chien-Ming Huang. 2025b. Conversational Agents in Behavioral Sleep Medicine: Designing Self-Report and Analytics Tools. *arXiv preprint arXiv:2509.15378* (2025).
- Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2025c. User interaction patterns and breakdowns in conversing with llm-powered voice assistants. *International Journal of Human-Computer Studies* 195 (2025), 103406.
- Niharika Mathur, Kunal Dhodapkar, Tamara Zubatiy, Jiachen Li, Brian Jones, and Elizabeth Mynatt. 2022. A collaborative approach to support medication management in older adults with mild cognitive impairment using conversational assistants (CAs). In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–14.
- Lina Mavrina, Jessica Szczuka, Clara Strathmann, Lisa Michelle Bohnenkamp, Nicole Krämer, and Stefan Kopp. 2022. "Alexa, You're Really Stupid": A Longitudinal Field Study on Communication Breakdowns Between Family Members and a Voice Assistant. *Frontiers in Computer Science* 4 (2022), 791704.
- Roger K Moore, Hui Li, and Shih-Hao Liao. 2016. Progress and Prospects for Spoken Language Technology: What Ordinary People Think.. In *Interspeech*. San Francisco, CA, 3007–3011.
- Linda Mortensen, Antje S Meyer, and Glyn W Humphreys. 2006. Age-related effects on speech production: A review. *Language and Cognitive Processes* 21, 1-3 (2006), 238–290.
- Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–7.
- Katherine O'Brien, Anna Liggett, Vanessa Ramirez-Zohfeld, Priya Sunkara, and Lee A Lindquist. 2020. Voice-controlled intelligent personal assistants to support aging in place. *Journal of the American Geriatrics Society* 68, 1 (2020), 176–179.
- Bruna Oewel, Tawfiq Ammari, and Robin N Brewer. 2023. Voice Assistant Use in Long-Term Care. In *Proceedings of the 5th International Conference on Conversational User Interfaces*. 1–10.
- Young Hoon Oh, Kyungjin Chung, and Da Young Ju. 2020. Differences in interactions with a conversational agent. *International journal of environmental research and public health* 17, 9 (2020), 3189.
- OpenAI. 2023. ChatGPT can now see, hear, and speak. <https://openai.com/blog/chatgpt-can-now-see-hear-and-speak>
- Cathy Pearl. 2016. *Designing voice user interfaces: Principles of conversational experiences*. O'Reilly Media, Inc.
- Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.
- Alisha Pradhan, Leah Findlater, and Amanda Lazar. 2019. "Phantom Friend" or "Just a Box with Information" Personification and Ontological Categorization of Smart Speaker-based Voice Assistants by Older Adults. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.
- Alisha Pradhan, Amanda Lazar, and Leah Findlater. 2020. Use of intelligent voice assistants by older adults with low technology use. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 4 (2020), 1–27.
- Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident" Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on human factors in computing systems*. 1–13.
- Sandeep Puro, Haijing Hao, and Chenhang Meng. 2021. The use of smart home speakers by the elderly: exploratory analyses and potential for big data. *Big Data Research* 25 (2021), 100224.
- Sandeep Puro and Chenhang Meng. 2019. Data capture and analyses from conversational devices in the homes of the elderly. In *Advances in Conceptual Modeling: ER 2019 Workshops FAIR, MREBA, EmpER, MoBiD, OntoCom, and ER Doctoral Symposium Papers, Salvador, Brazil, November 4–7, 2019, Proceedings* 38. Springer, 157–166.
- Daniel Rausch. 2023. Previewing the future of Alexa. <https://www.aboutamazon.com/news/devices/amazon-alexa-generative-ai>

- John Rudnik, Sharadhi Raghuraj, Mingyi Li, and Robin N Brewer. 2024. CareJournal: A voice-based conversational agent for supporting care communications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–22.
- Jamie Sanders and Aqueasha Martin-Hammond. 2019. Exploring autonomy in the design of an intelligent health assistant for older adults. In *Proceedings of the 24th International conference on intelligent user interfaces: companion*. 95–96.
- Marcia Y Shade, Kyle Rector, Rasila Soumana, and Kevin Kupzyk. 2020. Voice assistant reminders for pain self-management tasks in aging adults. *Journal of gerontological nursing* 46, 10 (2020), 27–33.
- Ajwa Shahid, Jane Chung, and Seongkook Heo. 2025. Exploring Older Adults Personality Preferences for LLM-powered Conversational Companions. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.
- Elizabeth Smith, Petroc Sumner, Craig Hedge, and Georgina Powell. 2023. Smart-speaker technology and intellectual disabilities: agency and wellbeing. *Disability and Rehabilitation: Assistive Technology* 18, 4 (2023), 432–442.
- Brodrick Stigall, Jenny Waycott, Steven Baker, and Kelly Caine. 2019. Older adults' perception and use of voice user interfaces: a preliminary review of the computing literature. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. 423–427.
- Milka Trajkova and Aqueasha Martin-Hammond. 2020. "Alexa is a Toy": exploring older adults' reasons for using, limiting, and abandoning echo. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
- Pooja Upadhyay, Sharon Heung, Shiri Azenkot, and Robin N Brewer. 2023. Studying Exploration & Long-Term Use of Voice Assistants by Older Adults. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–11.
- Eleftheria Vaportzis, Maria Giatsi Clausen, and Alan J Gow. 2017. Older adults perceptions of technology and barriers to interacting with tablet computers: a focus group study. *Frontiers in psychology* 8 (2017), 1687.
- Linda Wulf, Markus Garschall, Julia Himmelsbach, and Manfred Tscheligi. 2014. Hands free-care free: elderly people taking advantage of speech-only interaction. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*. 203–206.
- Satoshi Yamada, Daisuke Kitakoshi, Akihiro Yamashita, Kentarou Suzuki, and Masato Suzuki. 2018. Development of an intelligent dialogue agent with smart devices for older adults: a preliminary study. In *2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. IEEE, 50–53.
- Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–35.
- Ja Eun Yu, Natalie Parde, and Debaleena Chattopadhyay. 2023. "Where is history": Toward Designing a Voice Assistant to help Older Adults locate Interface Features quickly. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
- Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Völkel, Johannes Schöning, Rainer Malaka, and Yvonne Rogers. 2022. Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity Dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–14.

Appendix

A Methods

A.1 ChatGPT+Alexa Skill: Activation Phrases

Following are the simple activation phrases used for the skill, similar to our prior work [Mahmood et al., 2025c]:

- Alexa, I have a question
- Alexa, question
- Alexa, let's chat
- let's talk
- Alexa, talk to me

B Data Analysis: Codebook

The codebook used for labeling our interaction data is given in Table 11.

Table 11. Our finalized codebook. Each one-turn query was labeled with codes listed below.

Code	Definition
Interaction type	The purpose of the query. Some interaction types had subcodes (e.g., "questions" were further labeled with question type).
Action	The action associated with the interaction type, if applicable. For instance, music and radio had actions for "play," "pause," "stop," etc.
Interaction	The start and end of multi-turn interactions were marked to count the number of turns (queries) in each interaction.
Social response/reaction	The immediate social response or reaction from the user.
Intended for Alexa	Whether the user directed the query at Alexa or not.
Alexa handled	Whether the query was transcribed by Alexa and Alexa acted or responded.
Alexa responded	Whether Alexa responded verbally to the query or not.
Success	Whether Alexa handled and responded to the query accurately.
Not handled	The user query was captured accurately but Alexa did not respond.
Private conversation about Alexa	The user engaged in a private conversation with themselves or others about Alexa during or immediately after the query.
Private conversation about same topic	The user engaged in a private conversation with themselves or others about the same topic queried to Alexa during or immediately after the query.
Private conversation about how to use Alexa	The user engaged in a private conversation with themselves or others about how to interact with Alexa during or immediately after the query.
Overlap	We coded separately whether the user talked over Alexa or vice versa.
Interruption	We coded separately whether the user interrupted Alexa or vice versa.
For each erroneous interaction	
Code	Definition
Error type	The type of error in the user query as determined by the coder from the transcribed logs.
Manifestation	Whether the coder could tell if an error occurred from the transcribed logs.
Identification	Whether the coder could tell if the user identified the error by their verbal response, reaction, or immediate action to fix the error.
Verbal/vocal reaction	The user's verbal or vocal reaction to the manifested error during or immediately after the query.
Retry	Whether the user tried to fix the error in the next immediate query.
Recovery strategy	The strategy employed by the user (if any) in the next immediate query to try and fix the error.
Resolution	Whether the error was resolved in the next immediate retry (if any) by the user.
Retries (n)	The number of retries taken by the user to fix a compounding error. The final query was marked with the total number of retries, signifying the conclusion of retries at that point for the original error.

Table 12. Field study demographics: General. All participants were fluent in English. All were retired except P10a (employed full-time) and P11a (employed-part time). Note: eq. denotes “equivalent diploma.”

P	Gender	Age	Ethnicity	Highest Degree	Profession	Disabilities/Aids	QOL Overall
Community Center – Assisted Living (by self)							
2	M	66	Prefer not to say	Bachelor’s degree	Physician assistant	Wheelchair	Alright
3	F	82	Caucasian	High school or eq.	Accountant clerk	Hearing aid	Alright
4	M	79	Caucasian	Master’s degree	Actuary	Wheelchair	Good
Community Center – Independent Living (by self)							
5	F	77	Caucasian	Bachelor’s degree	Social work		Good
6	F	81	Caucasian	Master’s degree	CIA	Cane/Rollator	Alright
7	F	74	African American	Bachelor’s degree	Administrator		
14	F	73	African American	Bachelor’s degree	Senior claims		Good
15	F	73	African American	Master’s degree	Educator/Pastor		Good
Homeowner (by self)							
8	F	84	Caucasian	Bachelor’s degree	Writer/Editor	Walker	Alright
9	M	75	Caucasian	Bachelor’s degree	Case management		Very good
13	M	80	Caucasian	High school or eq.	Management	Vision issues	Alright
Homeowner (couple, with spouse)							
1a	M	76	Caucasian	Bachelor’s degree	Business owner		Good
1b	F	75	Caucasian	Nursing diploma	Nurse		
10a	M	68	Caucasian	Bachelor’s degree	Staff engineer		Good
10b	F	94	Caucasian	Master’s degree	Education	Walker/Scooter	Bad
11a	M	72	Caucasian	Master’s degree	Urban planning		Very good
11b	F	71	Asian	Master’s degree	Social work/Teacher		Very good
12a	M	75	Caucasian	Bachelor’s degree			Very good
12b	F	72	Caucasian	Nursing diploma	Registered nurse		Good

Table 13. Field study demographics: Voice assistant technology use and familiarity.

P	Placement	Technology use
Community Center – Assisted Living (by self)		
2	Studio	Familiar with “the Alexa” in the dining room of community center. Has Alexa app on iPad.
3	Studio	Does not have smart phone. Never used VA before.
4	Studio	Had Alexa a long time ago, but not anymore.
Community Center – Independent Living (by self)		
5	Living room	Uses speech-to-text for messaging. Owns tablet.
6	Living room	Uses TV remote via voice commands.
7	Living room	Never used VA before.
14	Living room	Uses Siri on iPhone to ask questions and for directions, locations, and restaurant ratings. Owns Apple Watch and iPad, as well.
15	Living room	Never used VA before.
Homeowner (by self)		
8	Bedroom	Does not have smart phone but owns iPad. Not familiar with VAs.
9	Dining room	Owns a speaker compatible with Alexa, used mostly for music. Used occasionally for weather updates but primarily uses phone for this purpose.
13	Dining room	Owns tablet. Not familiar with VAs.
Homeowner (couple, with spouse)		
1	Living room	Has Alexa smart speakers in living room, kitchen, bedroom, study, and at daughter’s home next door. Used for communicating between rooms and timers while cooking, etc.
10	Living room	Not familiar with VAs.
11	Kitchen	Uses Siri on iPhone for calls (especially when driving), sending messages, asking questions, and Google Maps.
12	Living room	Has Alexa smart speakers in kitchen (will move to laundry room) and bedroom. Used for communicating (drop-ins) between rooms, playing music, and listening to radio.