arXiv:2408.16465v1 [cs.HC] 29 Aug 2024

Human and LLM-Based Voice Assistant Interaction: An Analytical Framework for User Verbal and Nonverbal Behaviors

SZEYI CHAN, Northeastern University, USA
SHIHAN FU, Northeastern University, USA
JIACHEN LI, Northeastern University, USA
BINGSHENG YAO, Northeastern University, USA
SMIT DESAI, Northeastern University, USA
MIRJANA PRPA, Northeastern University, USA
DAKUO WANG*, Northeastern University, USA

Recent progress in large language model (LLM) technology has significantly enhanced the interaction experience between humans and voice assistants (VAs). This project aims to explore a user's continuous interaction with LLM-based VA (LLM-VA) during a complex task. We recruited 12 participants to interact with an LLM-VA during a cooking task, selected for its complexity and the requirement for continuous interaction. We observed that users show both verbal and nonverbal behaviors, though they know that the LLM-VA can not capture those nonverbal signals. Despite the prevalence of nonverbal behavior in human-human communication, there is no established analytical methodology or framework for exploring it in human-VA interactions. After analyzing 3 hours and 39 minutes of video recordings, we developed an analytical framework with three dimensions: 1) behavior characteristics, including both verbal and nonverbal behaviors, 2) interaction stages—exploration, conflict, and integration—that illustrate the progression of user interactions, and 3) stage transition throughout the task. This analytical framework identifies key verbal and nonverbal behaviors that provide a foundation for future research and practical applications in optimizing human and LLM-VA interactions.

CCS Concepts: • **Human-centered computing → User studies**; **Sound-based input / output**; **Auditory feedback**; **Empirical studies in HCI**; **Interaction techniques**.

Additional Key Words and Phrases: User Study, Exploratory Study, Large Language Model-Based Voice Assistant, Verbal Interaction, Nonverbal Interaction, Human Behavior

ACM Reference Format:
Szeyi Chan, Shihan Fu, Jiachen Li, Bingsheng Yao, Smit Desai, Mirjana Prpa, and Dakuo Wang. 2024. Human and LLM-Based Voice Assistant Interaction: An Analytical Framework for User Verbal and Nonverbal Behaviors. 1, 1 (August 2024), 33 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding author: d.wang@northeastern.edu

Authors' addresses: Szeyi Chan, chan.szey@northeastern.edu, Northeastern University, USA; Shihan Fu, sh.fu@northeastern.edu, Northeastern University, USA; Jiachen Li, li.jiachen4@northeastern.edu, Northeastern University, USA; Bingsheng Yao, b.yao@northeastern.edu, Northeastern University, USA; Smit Desai, sm.desai@northeastern.edu, Northeastern University, USA; Mirjana Prpa, m.prpa@northeastern.edu, Northeastern University, USA; Dakuo Wang, d.wang@northeastern.edu, Northeastern University, 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 ACM 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.
© 2024 Association for Computing Machinery.
XXXX-XXXX/2024/8-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Voice assistants (VAs) have become increasingly popular thanks to their ability to converse in natural language [10]. These intelligent systems can now be integrated into a wide range of devices, from smartphones and computers to speakers, glasses, and even watches. Users regularly turn to VAs to for assistance with a diverse array of tasks, ranging from simple tasks such as querying about the weather or playing music to more complex tasks like navigation assistance and cooking support [97]. Most of the current popular VAs (e.g., Alexa and Siri) employ a conversational model backend that is primarily reliant on detecting a predefined set of intents and responding with a predefined set of responses to the user. This approach makes them well-suited for **simple tasks** [94]. Still, it can lead to high error rates, out-of-scope exceptions, unnatural conversational experience, and lower user satisfaction when it comes to more **complex tasks** that may involve multi-turn interactions[1] [110, 118].

Recognizing these limitations, researchers in the field of Human-Computer Interaction (HCI) have been exploring the integration of large language models (LLMs) to enhance voice-based interactions. This includes the development of LLM-based voice assistants (**LLM-VA**[2]) [41, 48, 115], which aim to enable more natural and human-like communication [23, 64, 64]. With advancements in LLM technology, LLM-VAs now enable more continuous and human-like interactions with users [114], allowing for conversations that are more fluid and natural, resembling human-human interaction patterns [44]. LLM-VAs have been effectively utilized in diverse domains such as in-vehicle systems [33, 48], healthcare applications [115], robotics [53] and cooking scenarios [21]. These advanced VAs mitigate the limitations of commercial VAs by engaging users in longer, more continuous interactions to assist with complex tasks [5, 9, 27]. As these technologies evolve, a critical question emerges: how can we further enhance **human and LLM-VAs interaction** design?

Taking inspiration from the ideals of human-human interaction, wherein **nonverbal cues** play a crucial role, often revealing more authentic thoughts and emotions than the **verbal behavior** alone [8, 43, 77]. These nonverbal behaviors have the potential to augment the capabilities of VAs, helping to better predict user intentions and create more seamless, personalized experiences [2].

Current studies on human-VA interactions often overlook the integration of nonverbal interactions or focus on them in isolation (e.g., studies involving eye-gaze [49]), despite research in other AI agent modalities have demonstrated the value of holistically incorporating nonverbal cues. AI agents using avatars can create a sense of closeness, transforming the perception of the agent from a mere tool to a companion [16]. In Human-Robot Interaction (HRI), social robots with human-like attributes evoke more natural and human-like emotional responses from users [72, 96]. The success of these AI agents highlights the importance of integrating verbal and nonverbal behaviors in human-agent interactions. Given that LLM-VAs are now capable of more natural, human-like interactions, incorporating verbal and nonverbal information can provide them with a more nuanced understanding of user behavior, especially when performing complex multi-turn tasks. This integration can lead to the development of more socially aware and intuitive VA technologies.

However, in human and LLM-VA interactions, the full potential of leveraging nonverbal cues remains unexplored, largely due to the lack of a systematic analytical framework for analyzing the integration of nonverbal behaviors in LLM-VA interactions. Therefore, our study aims to address this research gap by exploring human behavior in interactions with LLM-VAs and proposing a systematic analytical framework to examine users' both verbal and nonverbal behaviors. By

[1]Despite that users can download and configure third-party Skills or Actions to supplement the basic capabilities of commercial VA platforms, few users are aware of it [94].
[2]In this paper, we are referring to LLM-based voice assistants as LLM-VA.

leveraging these insights, we seek to support the future design of LLM-VAs, creating more natural and effective human-LLM-VA interactions.

In this paper, we focused on cooking as our task scenario. Cooking is a commonly chosen experiment scenario in prior research [45, 50, 58, 95, 109] for its inherent complexity, requiring extended back-and-forth interactions between the user and the LLM-VA. Moreover, the hands-on nature of cooking tasks increases the likelihood of participants using nonverbal signals [50], providing a rich environment for observing both verbal and nonverbal behaviors. This choice of scenario allows us to explore the full spectrum of human-LLM-VA interactions in a practical, real-world context. Building on our previous research that explored user perceptions of interacting with LLM-VA during cooking [21], we conducted a focused re-examination utilizing a subset of the data distinct from the prior study. This paper's analysis focuses on manual coding of 3 hours and 39 minutes of video recordings, identifying the verbal and nonverbal interaction of 12 participants with an LLM-VA as they prepared a salad in a real-world kitchen environment. These participants were assisted by our LLM-VA, *Mango Mango* (MM), which can provide step-by-step cooking instructions, respond to cooking-related general inquiries, and can provide social chit-chat functionalities. From this analysis, we propose an analytical framework incorporating both verbal and nonverbal elements to identify interaction stages.

We developed an analytical framework (shown in Figure 1), structured around three key dimensions: 1) Behavior Characteristics — this includes both verbal and nonverbal behaviors observed during interactions with the LLM-VA, 2) Interaction Stages — defined as **Exploration**, **Conflict**, and **Integration**, these stages delineate the progression of user interactions, and 3) Stage Transition — this dimension investigates how stages transit throughout the task, specifically focusing on patterns of stage transition, such as the ascending from **conflict** stage to **integration** stage and regressing from **integration** to **conflict** stage. The first dimension pinpoints the specific characteristics of users' verbal and nonverbal behaviors. The second dimension categorizes these behaviors into the defined stages, each stage characterized by unique behavioral patterns. The third dimension explores how these stages transition during the interaction, offering insights into the dynamic nature of user engagement with LLM-VAs.

This framework systematically considers both the verbal and nonverbal behaviors of participants engaging with the LLM-VA. It is designed to highlight user communication's dynamic and evolving nature, providing insights into how these interactions change over time. By continuously analyzing these shifts, the framework offers a comprehensive view of stage transition throughout the task, facilitating a deeper understanding of user engagement and improving the system's adaptability. As interest in the application of LLM-based voice assistants grows, our study seeks to provide a foundation for future research and implementation by offering a framework for analyzing user interactions with LLM-based voice assistants.

In summary, our paper makes the following contributions:

(1) We propose an analytical framework with three dimensions—behavior characteristics, three interaction stages, and stage transition—by identifying key verbal and nonverbal behaviors of users during communication with LLM-VAs. This framework provides a foundation for future research and practical applications in optimizing human-LLM-VA interactions.
(2) We outline design implications to guide researchers and system designers in utilizing this analytical framework to develop more effective and fluent interactions between humans and LLM-VAs.

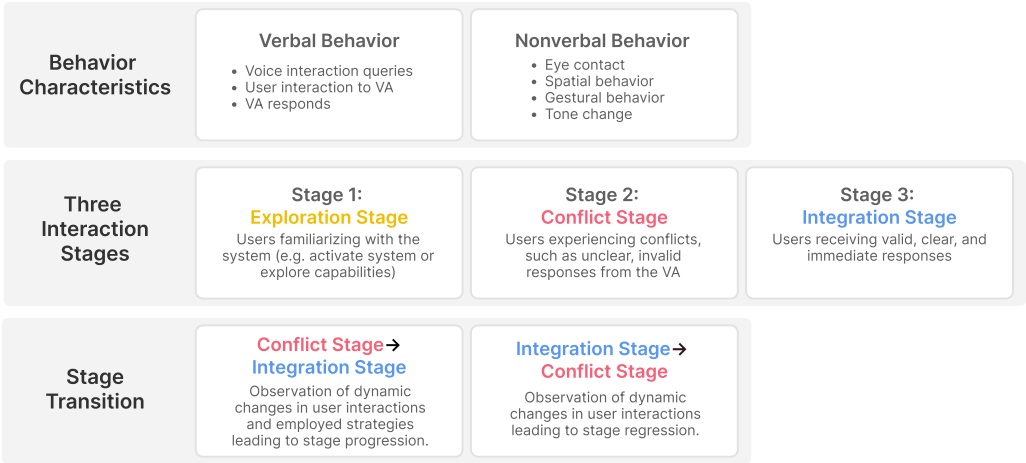


Fig. 1. The proposed analytical framework consists of three dimensions: 1) behavior characteristics, 2) the three interaction stages, and 3) stage transition.

2 RELATED WORK

To provide a thorough understanding of our study, we begin by introducing the importance of both verbal and nonverbal behaviors, covering interactions in both human-to-human and human-to-VA communication in Section 2.1. Following this, we explore various aspects of VAs, with a special focus on those designed for cooking tasks, as detailed in 2.2.

2.1 Vebal and Nonverbal Behaviors in communication

2.1.1 Behaviors in Human-Human Communication. Human communication is a complex and dynamic process involving multiple modalities. In human-human communication, verbal and nonverbal behaviors intersect [1]: verbal behaviors, delivered in real-time, serve as crucial social signals that can reveal the status of a conversation [79]; nonverbal behaviors, such as gestures and posture, can complement but also contradict verbal messages [7]. Different interpretations of behaviors can sometimes complicate the dynamic of human-to-human interactions across cultures, as certain gestures are seen as friendly in one culture but offensive in another [6, 70]. Consequently, the perception of others is shaped by a combination of verbal behavior and nonverbal expressions [101]. Understanding the relationship between verbal and nonverbal behavior is essential for effective human-human communication, integrating the spoken word and the unspoken signals to interpret messages.

In human-human communication, nonverbal communication significantly influences interactions. Nonverbal communication extends beyond words [55], encompassing behaviors that convey information, such as body language and visual cues, which enhance understanding within a shared context [28]. Prior studies identified four key nonverbal modalities: tone change (e.g., pitch variations), eye contact (e.g., gaze direction), gestural behavior (e.g., waving), and spatial behavior (e.g., distance and orientation between individuals) [8, 43], each adding depth to verbal exchanges and enriching interactions.

In scenarios where physical nonverbal cues are absent, such as in text-based user interfaces (e.g., chatbots) used for human-to-human communication through technology, the integration of verbal and nonverbal elements remains essential. Emoticons and emojis serve as digital stand-ins for the

nonverbal behavior commonly missing in online conversations, effectively mimicking face-to-face human interactions [59, 111]. The elements, from simple keyboard-created faces to the wide array of symbols for conveying emotions and reactions, help to bridge the communication gap in digital exchanges. Additionally, just as gestures and expressions differ significantly from one culture to another, the usage and interpretation of emojis also reflect cultural distinctions. For example, France's top emojis are heart-related, while other countries prefer face emojis. [67]

While human-human communication serves as a foundational communication model, there is ongoing debate about its relevance for human-VA interactions. Some research studies on VAs apply human communication principles to narrow the communication gap between humans and technology, including verbal and nonverbal behaviors. In contrast, some research argues the opposite. For example, Porcheron et al. citeporcheron2018voice suggest that the device's responses do not have the same conversational status as human dialogue. Our study contributes to this argument by proposing a formative approach to evaluating VUIs, with the goal of ultimately enhancing human-VUI interactions.

2.1.2 Behaviors in Human-VA Communication. Recognizing the significance of both verbal and nonverbal behaviors in communication, the Human-Computer Interaction (HCI) community has mainly concentrated on exploring both behaviors in two main aspects of human-VA interactions: first, integrating nonverbal behaviors into VAs, enabling the VA to display gestures or expressions; second, enabling VAs to interpret and respond to multimodal inputs, such as verbal and nonverbal behaviors from the user.

Exploring the integration of nonverbal behaviors into VAs is an increasingly researched area in the HCI community, showing significant potential to enhance interaction quality [20, 38, 40]. Previous studies have investigated the embodiment of VAs, enabling them to use visible speech, facial expressions, or body language, thereby adding supplementary communication modalities [68]. These behaviors enhance synchronicity and fluency between the speaker and listener, elevating the conversational quality and fostering social connections [22, 39]. Moreover, studies also demonstrate that users tend to interact with these advanced VAs similarly to how they would with other humans. These advanced VAs are capable of generating responses similar to those triggered by actual human interactions, highlighting their effectiveness in mimicking human-like interactions [78].

VAs predominantly rely on voice commands, showing a limited ability to perceive and engage with nonverbal communication [31]. Therefore, human verbal and nonverbal behavior as inputs to enhance interactions with VAs are another growing research area in the HCI community. In the Human-Robot Interaction (HRI) community research, the researchers have been trying to reduce the error rate of robots' capabilities to interpret and respond to human gestures and expressions accurately, such that it allows the interactions to be more intuitive and user-friendly by mirroring the complex dynamics of human communication [56, 57]. Building on the advancements in HRI, recent research explores the potential of VAs to detect errors using nonverbal behavior, aiming to overcome existing limitations and enhance the efficacy of human-technology interactions [32]. For example, TAMA is a system that incorporates gaze to enhance conversational interactions. The system detects a user's gaze direction and moves its "head" to establish mutual gaze with users, simulating eye contact during communication. Study results showed that this design increased engagement and led to higher rates of repeated queries and extended interaction times with the VA [74]. Researchers have also investigated gestural triggers for conversational agents by eliciting a range of gestures to pinpoint five viable options. A user study was conducted to evaluate users' acceptability and the effort needed to perform them effectively [88]. Results showed that among the five options, participants most preferred the snap and wave gestures.

To further explore human-VA communication dynamics, our study investigates the dynamic transition of user behavior, encompassing verbal and nonverbal elements, during interactions with LLM-VA. Exploring the characteristics and transition of users' verbal and nonverbal behaviors throughout the conversation process can enhance the design of human-VA interactions. By understanding the dynamics of human behavior in natural conversations with LLM-VAs, designers can improve the user experience and bridge communication gaps between humans and LLM-VAs in scenarios involving complex tasks, such as cooking.

2.2 Traditional Voice Assistant vs. LLM-Based Voice Assistant

Traditional VAs are becoming important in our daily routines, especially at home environments [12, 13, 60, 69, 97], with households interacting with them about 4.1 times daily on average [15] for services like playing music, setting alarms, controlling IoT devices, etc. [5, 15, 29, 54, 66, 83, 90, 105]. The VAs also show promise in self-care [19], education [85, 102, 117], and healthcare [11, 17, 46] areas. Despite the advantages of VAs, some users discontinue VA usage due to its limited capabilities, leading to users' frustration and disappointment [26].

In the current usage of commercial VAs, users frequently face challenges, including inaccuracies in speech recognition, out-of-scope queries, and dependence on external sources for query responses [14, 32, 61, 71, 75]. These challenges are mainly due to VAs' reliance on conventional language models (e.g., task-specific models trained under supervised settings) and predefined heuristics (e.g., rules for filtering keywords) [80, 110, 118]. As a result, the user interactions are primarily limited to single-turn questions or basic command-response structures [12, 54, 63]. The limitations restrict the flexibility and capacity of VAs to engage flexibly in complex and multi-turn conversations such as in cooking scenarios [89]. Jaber et al. [50] also highlights the importance of context-aware voice interactions in cooking. By evaluating interactions with a wizard-led context-aware VA, they found the interactions to be more fluid. These scenarios demand extended interaction and a deep contextual understanding, highlighting the importance of flexibility in the VAs' responses and actions. Furthermore, research on user behavior during interactions with VAs often focuses on straightforward tasks [31] and exploring how users navigate errors or unsatisfactory experiences [14, 32, 71]. However, this approach provides only a limited view of user behavior in more complex scenarios.

To deepen our comprehension of user behavior within complex scenarios, the development of LLMs, such as LLaMA [104] and GPT-4 [84], provides new pathways for researchers to explore and overcome the challenges of previous VA technologies. LLMs are effective at processing large text inputs, interpreting instructions, and generating quality responses [3, 92, 98]. LLMs can also manage tasks beyond the scope of traditional language models, including multi-turn conversations and complex queries [114]. The capability of these models in prompt engineering and other areas opens up new opportunities for diverse applications of LLMs in various fields [34, 51, 62, 107, 112, 113].

Integrating LLMs into VA for cooking illustrates an example of how technologies can enhance daily activities by providing real-time assistance and guidance. Cooking is a daily task that requires a foundation in basic cooking techniques and knowledge [81]. Particularly when trying new recipes or for those new to cooking, external assistance or recipe guidance is often essential [58, 109]. These needs become more pronounced when the cooking process requires multitasking or when hands are preoccupied [65].

While existing AI solutions have attempted help users learn and execute cooking tasks more efficiently, most of the research has focused mainly on overcoming practical challenges in the cooking process itself [24, 45, 82, 95]. There is, however, a research gap in understanding users' interaction behaviors with LLM-VAs during cooking, both verbally and nonverbally. In our study, we aim to develop an analytical framework that identifies the verbal and nonverbal behaviors that

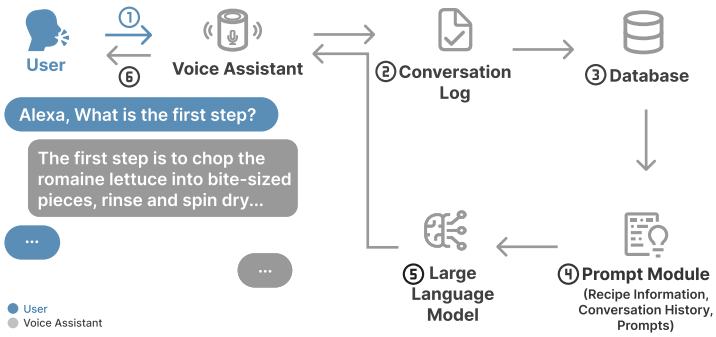


Fig. 2. System diagram of "Mango Mango" (MM). The process begins with users providing voice input to Alexa. Then, Alexa performs a speech-to-text conversion and adds the transcribed input to the conversation log. This log is saved to a database. Next, the conversation histories are processed in the prompt module. The completed prompt is then sent to GPT-3.5 Turbo, and finally, the resulting response is sent back to Alexa, where it is converted into speech for the user, completing the system loop. The complete and detailed system flow is described in previous work [21].

enhance communication between users and LLM-VAs during complex tasks such as cooking. By identifying and examining interactions across different interaction stages, i.e., **exploration**, **conflict**, and **integration** stage, we seek to understand not only the verbal commands and responses, but also the nonverbal behaviors. We aim to examine these interactions from a detailed view to a broader perspective to understand how they change and evolve over time. These insights will provide a foundation for future research and practical applications, optimizing human and LLM-VA interactions. Our comprehensive approach emphasizes the importance of understanding these dynamics to inform the development of more intuitive and responsive VA systems that cater effectively to users' needs during complex tasks.

3 SYSTEM DESCRIPTION: THE "MANGO MANGO" SYSTEM

Our study employed the LLM-VA "Mango Mango" (MM)[3] developed from previous work [21]. MM was created using the Alexa Skill platform, and its workflow is illustrated in Figure 2. MM was integrated into Amazon Smart Speakers, taking advantage of the flexibility of Text-to-Speech (TTS) and Speech-to-Text (STT) technologies to enhance application efficiency and effectiveness.

MM was enhanced by integrating with the GPT-3.5-Turbo model, significantly upgrading its natural language processing capabilities. GPT-3.5-Turbo excels in both understanding and generating natural language. A key feature of this model is its ability to handle extensive input content, allowing for the inclusion of past conversation histories. This capability leads to more coherent and contextually appropriate multi-round conversations. Furthermore, the GPT-3.5-Turbo model provides robust and stable API support, which is instrumental in ensuring the smooth execution of our laboratory experiments.

This system features a prompting module designed specifically for cooking scenarios. This module organizes and reconstructs user input before sending it to the LLM (GPT-3.5-Turbo model). The input comprises all the essential information about the recipe and customized instructions in response to user queries. For our experiment, we accurately transcribed a YouTube video recipe for

[3]In this paper, MM refers to the LLM-based system, while Alexa refers the original system or the device itself.

Participant	Gender	Age	Cooking Frequency	VA Usage Frequency
P2	F	18-24	At least once per week	At least once per month
P3	F	25-34	At least once per week	Rarely
P4	M	25-34	Daily	At least once per week
P5	M	18-24	Daily	At least once per month
P6	M	25-34	Daily	Daily
P7	F	25-34	Daily	At least once per week
P8	M	18-24	Daily	Daily
P9	F	25-34	At least once per week	At least once per week
P10	F	25-34	Daily	Rarely
P11	M	25-34	At least once per week	Rarely
P12	M	18-24	Daily	Daily
P13	M	25-34	Daily	Daily

Table 1. Demographics of participants (Female = F, Male = M). The ages of participants were collected in forms of ranges. The table listed the frequency of users' cooking activities and their usage of VAs.

Chicken Mango Avocado Salad, ensuring the originality of the instructions. This structure ensures precise and relevant assistance in our cooking-related tasks.

4 METHODS

We performed a focused reexamination of data collected from an existing study [21] to explore how users' behaviors emerge and evolve during interactions with LLM-VAs. While the previous work concentrated on the design, development, and user perceptions of the "Mango Mango" system, the current study shifts focus to analyzing user behavior using a different subset of the study data. From this analysis, we propose an analytical framework for understanding nonverbal behaviors in LLM-VA interactions.

This section provides an overview of our study details and methodology. Section 4.1 explains our recruitment process, Section 4.2 outlines the design and procedure of our study, and Section 4.3 details our data collection and analysis methods.

4.1 Recruitment and Participants

In the original study, participants were recruited through social media platforms and email by sending recruitment posters, which included details explaining our research objectives, a direct link, and a QR code that directed interested individuals to the screening questionnaire. The screening questionnaire ensured that participants met specific criteria: at least 18 years old, fluent in English, experienced in cooking, comfortable with audiovisual recording during the experiment, and no allergies to the ingredients used in the study. A total of 13 participants were successfully enrolled in our study. One data point was removed from the data set due to a system change, resulting in a final pool of 12 participants. The demographics of the participants are shown in Table 1.

To ensure safety, the cooking process was designed to avoid using ovens, sharp knives, stoves, or any other potentially hazardous tools and appliances. In recognition of the participant's time and contribution, each participant received a $30 Amazon e-gift card. The study's design and protocol received approval from the university's Institutional Review Board (IRB).

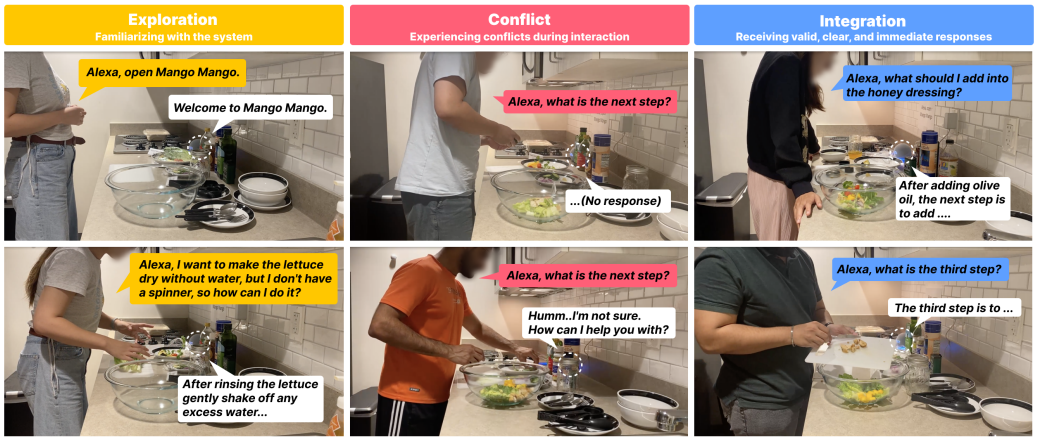


Fig. 3. Pictures of participants actively engaged in the experiment at different stages of their experience with dialogue history. The study took place in the fully operational kitchen in the smart home laboratory. The Alexa device was placed on the left side of the participants (circled in white).

4.2 Experimental Setup and Procedure

The original study was conducted in a smart home laboratory designed as a one-bedroom apartment, featuring a fully equipped kitchen and monitoring cameras. Figure 3 shows examples of participants engaging in the experimental task at different stages and provides a view of the kitchen setup. The Alexa device was positioned on the left-hand side of the participants. Informed consent for recording was obtained from all participants prior to the experiment.

Each experimental session lasted less than an hour and included the following procedure: 1) A researcher first began with a brief introduction to the study. 2) Then, participants received a five-minute tutorial on how to interact with MM, which aimed to familiarize the participants with MM's functionalities and interaction for the cooking task. The tutorial included demonstrations of three utterance examples: *"What is the first step?"*, *"What if I don't have chicken, what should I do?"*, and *"What did I just ask?"*. The researcher also show demonstrations of activating and stopping MM. 3) Next, the participants watched instructional YouTube tutorials on making chicken mango avocado salad. While memorization of the content was not mandatory, they were encouraged to familiarize themselves with the recipe. 4) Finally, participants proceeded to prepare the salad while freely interacting with MM. The anticipated time for completing the cooking task was approximately 20 minutes; however, participants were given a strict maximum time limit of 30 minutes. When this time limit was exceeded, the cooking activity would be concluded. During this time period, participants had to follow the steps from the video without being able to re-access it. Instead, participants could request assistance from the VA for clarifications on recipe steps or ingredient preparation.

4.3 Data Collection and Analysis

Participants' cooking process was video-recorded with their oral consent. In total, 3 hours and 39 minutes of audiovisual data were captured, which were transcribed by using an automated transcription service, resulting in a detailed interaction log of participant engagement with MM. Researchers manually checked the transcripts to ensure accuracy. The video coding process for this study involved three annotation cycles and used the ELAN annotation software [100], as shown in

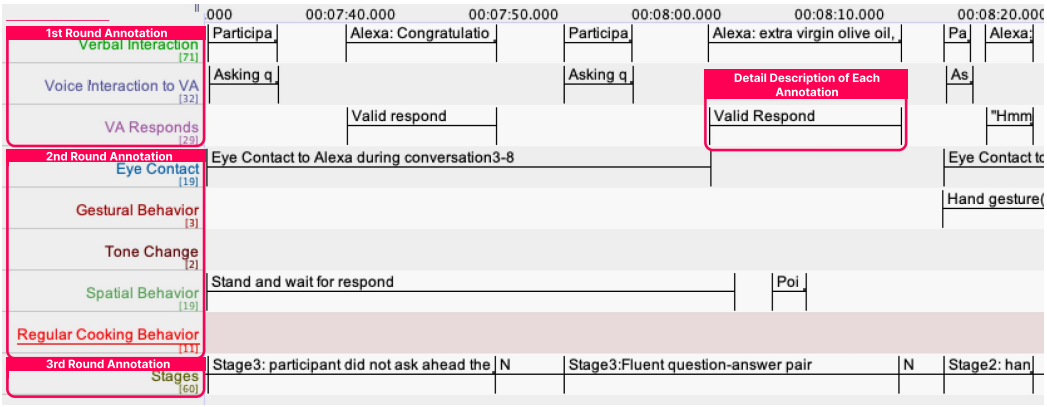


Fig. 4. Screenshot of ELAN annotation software showing three annotation cycles. All layers mentioned Section 4.3 can be seen here.

Figure 4. Researchers first separate the interactions into categories corresponding to each cycle's focus, then create annotation tiers for each category in ELAN. The first annotation cycle focuses on verbal interactions between participants and the VA. The second cycle concentrates mainly on participants' nonverbal behaviors. After identifying behavioral characteristics in the first two cycles, researchers collaboratively use these characteristics in the third cycle to analyze behaviors using grounded theory, applying both inductive and deductive coding methods [42, 103].

We employed a deductive approach by adapting previous research on communication behaviors in our coding process. Specifically, we adapted characteristics of nonverbal behavior categorized from Argyle [8] and Fiske [43]. These characteristics guided our analysis of nonverbal behaviors shown by participants. Additionally, we adapted prior research on verbal interactions with VAs, particularly from Vtyurina and Fourney [106], who categorized conversational cues by tracking the frequency of specific intents. We also incorporated findings from Mavrina et al. [71], detailed common requests and responses during verbal communication breakdowns when interacting with VAs. These previous works allowed us to have predefined categories on how users navigate interactions with VAs and how they respond to verbal and nonverbal communication challenges. This predefined codebook formed the initial structure of our analysis.

We recognized that existing literature has limitations in capturing the nuances of user behavior when interacting with LLM-VAs, a relatively new technological domain. While prior studies have documented human behavior with VAs, there remains a gap in understanding how users behave in this novel context. To address this, we employed an inductive coding approach to allow new patterns and behaviors directly from the data. This method allowed us to identify behaviors that may not have been observed in previous research and develop themes grounded in observations from the LLM-VA interactions. This combination of inductive and deductive approaches ensured that our analysis was grounded in prior knowledge and allowed room for discovering other behaviors unique to interactions with LLM-VAs.

The coding process involved continuous collaboration between two researchers. Each researcher initially coded independently in each annotation round and then convened to discuss and compare their codes. Any remaining disagreements were resolved using the negotiated agreement method [18] with the entire research team. Researchers verified the accuracy of the time labels for each interaction to ensure correctness. For example, the transcriptions of question-answer pairs were accurately labeled with the respective speakers.

5 BEHAVIOR CHARACTERISTICS

The first dimension of the framework is to identify the characteristics of users' verbal and nonverbal behaviors. This was accomplished during the first two annotation cycles.

In the first annotation cycle, researchers focused on verbal interactions between participants and the VA. Researchers classified verbal behavior into 3 categories: 'Voice Interaction', 'User Voice Interaction', and 'VA Responds'. 'Voice Interaction' refers to the entirety of the conversation, encompassing all queries. The *'User Voice Interaction'* tier indicates user queries, labeling behaviors such as simply asking a question, cutting off the VA's response, or instructing the VA to stop. Finally, 'VA Responds' refers to the VA's responses, which were further classified as valid if the VA answered correctly or provided a relevant answer. The response was annotated as invalid if unrelated to the participant's question.

The second round of the annotation cycle focused mainly on participants' nonverbal behaviors. Researchers classified these nonverbal behaviors into four categories: *'Eye Contact'*, *'Gestural Behavior'*, *'Tone Change'*, and *'Spatial Behavior'*. *'Eye Contact'* refers to instances when participants make direct or indirect eye contact, signaling attention or engagement. *'Gestural Behavior'* encompasses movements such as hand gestures or head nods that supplement verbal communication. *'Tone Change'* involves variations in vocal pitch, loudness, or intonation, indicating different emotional states. *'Spatial Behavior'* pertains to participants' physical proximity and orientation, including actions like leaning closer or moving away. This annotation scheme is based on previous research [8, 43], which provides a categorization of nonverbal human behavior observed in conversational contexts. Annotations included the duration and specific details of the behaviors, such as recording a participant leaning closer under *'Spatial Behavior'*. Additionally, time periods when the user was cooking independently without interacting with the VA were annotated as *'Regular Cooking Behavior'*.

In the third annotation cycle, researchers defined the three interaction stages of user interaction with the VA during the cooking process based on existing literature (e.g., [71, 106]) and observed user behavior: **exploration**, **conflict** and **integration**. These stages will be described in detail in the next section.

6 THE THREE INTERACTION STAGES

Based on our observation, we found the characteristics of human verbal and nonverbal behaviors within the interaction flow, and categorized them into three distinct interaction stages as the second dimension of the framework: **exploration**, **conflict**, and **integration**, as detailed in Table 2. Each stage defines unique behaviors that users show when interacting with LLM-VAs.

Specifically, 'stage' denotes distinct periods characterized by users' engagement in specific behaviors rather than static phases. User interaction within these stages is dynamic due to frequently transitioning across various stages throughout the execution of the task. This dynamic reflects the shifts that occur as users navigate through the experience.

6.1 Exploration Stage

6.1.1 Definition & Characteristics. In the **exploration** stage, users begin their initial familiarization with the system, either by activating it or exploring its capabilities [106].

For Verbal Behavior. Users may engage in conversation such as awakening the VA for the first time, forgetting to use the wake word (e.g., "Alexa"), or forgetting to initiate the correct skill (e.g., "Open Mango Mango"). They may also ask questions that go beyond the recipe for the first time, indicating an exploration of the system's capabilities.

Stage	Definition	Theme	Scenario
Exploration	Users initially interact with the VA system.	Awake System	- Awaken Alexa for the first time. - Fail to call "Alexa" on the initial attempt. - Forget to say "Open Mango Mango".
		Explore System Capability [106]	- Asking questions beyond the recipe for the first time.
Conflict	The VA system's response is unclear, invalid, or delayed.	Incorrect Response	- The system misinterprets user questions. - The system incorrectly comprehends the input, leading to incorrect responses. - Exit the system.
		No Response	- The system fails to respond.
		Clarification Response	- The system requests clarification due to misunderstood input.
		Conversation Conflict[86]	- User Interrupts VA. - VA interrupts user.
		Work Independently [106]	- Murmuring. - Proceeds with tasks without waiting for confirmation. - Does not follow the recipe.
Integration	The VA system provides valid, clear, and immediate response.	Receive Valid Response [93]	- Ask a question beyond the recipe. - Receive actionable steps. - Ask a follow-up question.

Table 2. Qualitative code book and description of characteristics of each stage.

For Nonverbal Behavior. During the **exploration** stage, eye contact with the VA emerged as the most frequent nonverbal behavior among all participants (M = 1.33, SD = 0.65), demonstrating an instinctual effort to establish a visual connection with their new conversational partner. Spatial behaviors were almost absent (M = 0.25, SD = 0.45) during initial interactions, such as leaning in closer to the VA, indicating a neutral communication style adopted by users. There was no significant gestural and tone change behavior was observed at this stage, indicating that users did not rely on gestures and tone to communicate with the VA initially.

6.1.2 Observations. In our study, we observed diverse interaction patterns among participants as they explored the ways to initialize the system. As mentioned in Section 4.2, participants were introduced to MM through a brief tutorial on prompted questions. Despite receiving brief training, participants still struggled with awakening the system to initiate communication.

A common issue observed was the failure to begin requests with the wake word "Alexa" or the specific command "Open Mango Mango," which is a crucial step for initial system engagement. This requirement may add an extra cognitive load to the participants during initial use as this action deviated from the participant's usual conversational patterns, creating a barrier to seamless initial interaction with the system. It is important to note that there were instances where system errors occurred, and participants did not realize that they were no longer interacting with MM. However, this stage does not account for cases where users were unaware of Alexa disengaging and failing to reactivate it.

In this stage, we also observed a significant shift in some participants' engagement with MM as they explored its functionalities more deeply. The participants were informed MM was powered by LLM at the beginning of the experiment, which suggested to participants that they could interact

with MM with no limitations regarding the scope of inquiries. Encouraged by this understanding, participants started to test the system's boundaries by asking questions unrelated to the cooking process, initiating a more exploratory interaction than initially anticipated. Through the inquiries, participants discovered that MM is more capable than simply guiding them through recipe steps and capable of responding to queries beyond the recipe's content, such as providing basic cooking knowledge. This expanded understanding of MM's capabilities encouraged a more inquisitive and experimental approach to interacting with the system, paving the way for a richer user experience.

6.2 Conflict Stage

6.2.1 Definition & Characteristics. In the **conflict** stage, users encounter responses from the VA that are unclear, invalid, or not provided promptly [86, 106].

For Verbal Behavior. During the **conflict** stage, participants often encounter difficulties executing the instructions suggested by MM. These challenges are mainly from communication breakdowns between the participants and MM. Participants often receive incorrect responses due to MM's failure to comprehend the user's input accurately, leading to misunderstandings or inappropriate replies. MM may also occasionally request further information or clarification from users to better understand their commands or queries. Another issue arises in instances of technical errors, where the VA may fail to recognize or respond to user inputs. Such situations can disrupt the user's experience, leading to frustration and potential disengagement from the task.

Furthermore, conversational conflicts also represent complexity in the interaction between users and the VA. We observed these conflicts occur on both the users' and VAs' sides. Such interruptions are mainly from misunderstandings, misinterpretations of commands, or the system's limitations in processing queries. Ultimately, faced with these challenges, some participants may bypass the system's assistance and proceed with the task independently.

For Nonverbal Behavior. All categories of nonverbal behavior are observed frequently in the **conflict** stage. Eye contact is particularly common (M = 7.92, SD = 4.42), as participants often look at the VA to check its status when it fails to deliver the expected response. Spatial behavior is notable (M = 3.00, SD = 2.80), with users moving closer to the device to ensure their commands are heard during communication breakdowns. Tone changes are also prominent (M = 1.92, SD = 1.38); users may raise their volume to express frustration or to regain MM's attention. Conversely, a drop in tone can indicate disappointment or impatience. Gestural behaviors, such as tapping the device or waving (M = 1.42, SD = 1.88), are employed to communicate with or reactivate MM.

6.2.2 Observation. In this stage, a frequently observed challenge is when MM responds incorrectly to participant queries. Often, the system misinterprets the user's questions, which can be attributed to errors in speech-to-text recognition. Such errors may prevent the system from comprehending the input accurately, leading to inappropriate responses. However, integrating LLMs into MM improves handling these challenges. LLMs enrich MM with a broader knowledge base and advanced comprehension capabilities. Consequently, even when misinterpretations occur, MM can navigate back to the conversation context by analyzing the query in light of the ongoing task and identifying that the misinterpreted query may not be related to the current task. When MM cannot fully comprehend an input, it requests further clarification from the user. For example, MM's response like *"I am sorry, but I didn't understand your question. Could you please rephrase it?"*. This is common in the experiment when the user request is ambiguous or the speech recognition system is uncertain about the user's request. Furthermore, sometimes issues arise when Alexa exits MM. In such scenarios, Alexa addresses queries using its standard functionalities, without including MM's specialized capabilities or specific knowledge base, such as detailed recipe information.

Interactions with MM can also lead to conflicts by interruptions in the dialogue flow. These interruptions appear either when participants proactively cut off MM or when MM interrupts the participants. We observed that participants commonly interrupt MM when the responses are not aligned with their expectations or when the responses are excessively lengthy. Conversely, MM may interrupt participants due to delays in processing responses performed by the LLM. Perceiving these delays as a non-response, participants might begin to repeat their queries or ask follow-up questions. However, MM might suddenly deliver a response to the initial query while the participant is still speaking, leading to overlaps and interruptions. While the LLM's advanced capabilities facilitate richer interactions by allowing for detailed explanations and the accommodation of follow-up inquiries, it also introduces complexities in maintaining a seamless and intuitive conversational flow due to the processing time, resulting in disruption of the natural flow of conversation.

Finally, even though participants consulted MM for recommended actions guided by MM's LLM during specific stages, some still chose to execute certain tasks on their own. This independent behavior was evident in actions such as murmuring to themselves. For instance, participants might quietly repeat the next ingredient, its quantity, or the forthcoming step. This helps the participant's reluctance to ask MM the same questions repeatedly. Participants engaged in self-dialogue, reiterating information provided by MM, and confirming the responses internally. This also helps solidify their memory of MM's instructions. Moreover, some participants proceeded without waiting for the system's confirmation or diverging from the provided recipe. An example of this is when a participant asks MM about a step in the recipe, either after they have already completed it or while they are in the process of doing so. Such an independent approach often indicates either a level of frustration with the system or a preference for self-guidance, effectively sidestepping the assistance offered by MM. However, it's important to note that taking independent steps without consulting MM is not considered in this observation.

6.3 Integration Stage

The **integration** stage is characterized by the VA system providing valid, clear, and immediate responses [93].

For Verbal Behavior. In the **integration** stage, users can confidently utilize the VA to assist with their cooking, asking questions that beyond the basic recipe instructions and executing actions advised by the VA, often including follow-up inquiries. The VA's responses, characterized by their validity, clarity, and immediacy, are indicative of successful integration. The interaction between the user and VA is marked by fluent conversations, with perfectly matched question-answer pairs. This stage signifies a shift as participants adeptly employ the VA, not just for recipe guidance but as a comprehensive cooking aid, extending their engagement to more complex culinary queries.

For Nonverbal Behavior. During the **integration** stage, eye contact remains the most observed nonverbal behavior (M = 11.75, SD = 4.43), indicating a stronger and more familiar connection with the VA. This sustained visual engagement suggests that users rely on visual feedback or confirmation as they interact with the VA. Spatial behavior is less frequent (M = 2.00, SD = 2.13), implying that users have found an optimal physical distance for interaction that they consistently adhere to, reflecting a stabilized interaction pattern. Gestural behavior is also reduced (M = 1.17, SD = 1.90), suggesting a decrease in the necessity for gestures as the interaction with the VA becomes more fluid and intuitive. Tone changes are rare in this stage (M = 0.58, SD = 0.67), signaling a comfortable and predictable rapport with the VA, and eliminating the need for users to modulate their tone to communicate effectively or prompt a specific response.

6.3.1 Observation. When the conversation progressed smoothly, we observed participants becoming more confident in asking questions beyond the recipe's scope. For instance, P7 asked about measuring 1/15 of a teaspoon. MM could adeptly provide answers without depending on the recipe information, even for queries unrelated to the recipe. Participants learned that they could explore topics outside the realm of recipe-specific questions and still obtain valuable, practical responses, which significantly enhanced the interactive quality of the conversation.

A noteworthy aspect of the interaction was the engagement of the participants in follow-up questions. These questions were naturally connected to participants' previous inquiries, creating a coherent dialogue. This interaction transcended a basic question-answer format, evolving into an enhanced interactive process. Previous studies highlighted that the iterative cycle of questioning and clarification played a pivotal role in decreasing uncertainties [47, 99]. This process clarified the ambiguities in the conversation, leading to a more accurate and effective exchange of information. Overall, MM consistently provided clear and practical instructions to participants, and maintained a smooth and informative interaction, enhancing the fluent conversation experience at this stage.

6.4 Quantitative Overview of Behaviors in the Three Interaction Stages

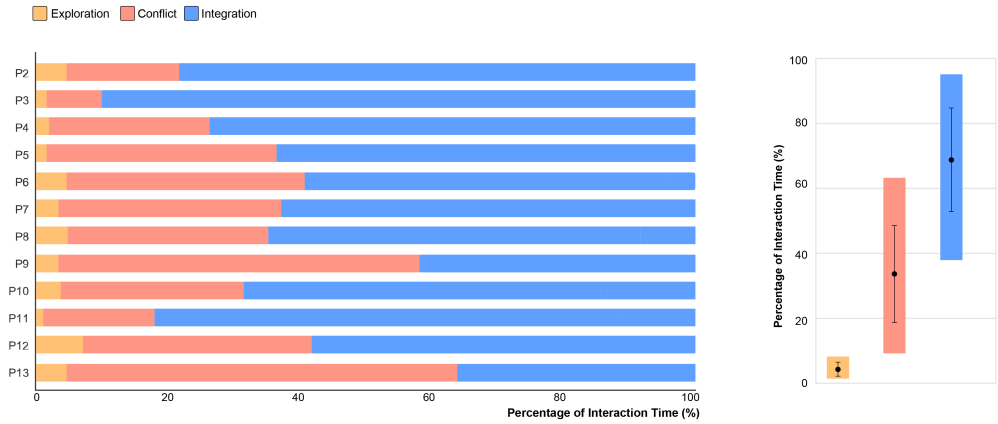
Participant	Total Question Asked	Question Answered Correctly by MM	Incorrect Answer From MM	% of Correct	Total Task Time
P2	29	23	6	79.30%	17:35
P3	32	25	7	78.10%	15:09
P4	27	22	5	81.50%	17:01
P5	35	28	7	80.00%	15:05
P6	48	28	20	58.30%	15:49
P7	30	20	10	66.70%	13:19
P8	36	28	8	77.80%	18:06
P9	40	13	27	32.50%	22:51
P10	75	59	16	78.70%	23:34
P11	19	17	2	89.50%	19:39
P12	47	27	20	57.40%	26:10
P13	34	10	24	29.40%	19:58

Table 3. The table provides an overview of the user performance on the salad-making task.

Following the characterization of the stages, we will analyze the communication dynamics among participants, focusing on both their verbal and nonverbal behaviors throughout the three stages.

6.4.1 Verbal Behavior. Of the total 447 queries in our study, 66.4% of valid queries received valid and accurate responses from MM. A query was considered valid only if it consisted of a complete question and its corresponding answer; queries lacking a response were excluded from consideration. A response was considered valid when it contained accurate content verifiable by the recipe and was expressed fluently. MM system errors and speech-to-text inaccuracies compromised 23.0% of queries. Furthermore, 10.6% of queries were invalid due to errors where the LLM would offer incorrect information sequencing or provide answers unrelated to the queries. Table 3 presents the accuracy rate of the queries for each participant.

In Figure 5a, we present the distribution of interaction time across participants as a normalized measure to facilitate comparative analysis. Given the intrinsic variability in absolute interaction times, we employed a percentage-based normalization method. This normalization technique



(a) (b)

Fig. 5. (a) Participant's time allocation of interaction across different stages with voice assistant. Each color represents a different stage, with the length corresponding to the duration of time spent in that particular stage of interaction. (b) Total time allocation of each stage among all participants. Each bar represents the range between the minimum and maximum occurrences recorded. The black dot indicates the mean frequency of these behaviors, while the vertical line extending from each bar denotes the standard deviation.

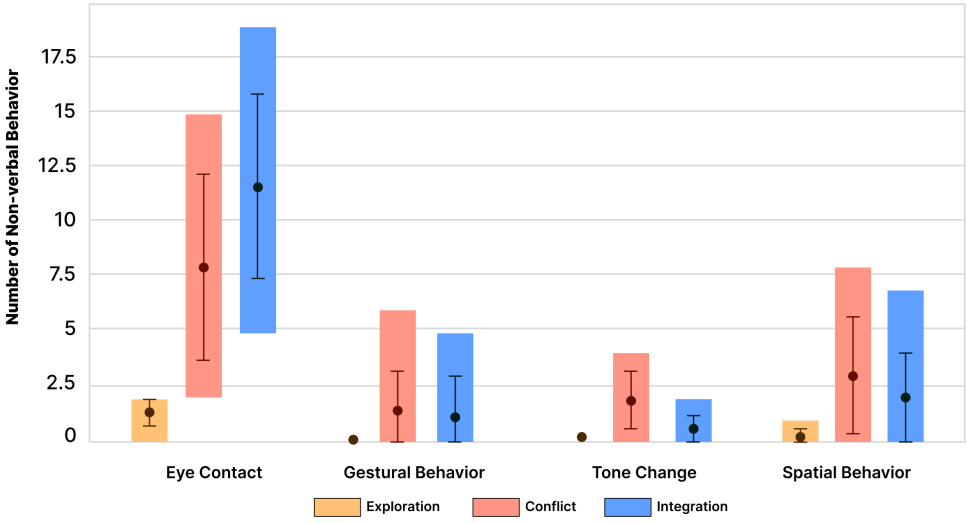


Fig. 6. A summary of the frequency of nonverbal behaviors across different stages among participants. The behaviors observed within the study included eye contact, gestural behavior, tone change, and spatial behaviors. Each bar represents the range between the minimum and maximum occurrences recorded. The black dot indicates the mean frequency of these behaviors, while the vertical line extending from each bar denotes the standard deviation.

Non-verbal Behavior	Stage	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
Eye Contact	Exploration	0	1	1	1	2	1	2	2	2	2	1	1
	Conflict	7	3	4	8	13	4	14	15	7	2	11	7
	Integration	12	7	15	16	8	12	19	10	5	9	18	10
Spatial Behavior	Exploration	0	0	0	0	3	0	0	0	0	0	0	2
	Conflict	6	0	0	2	2	0	1	4	0	1	0	1
	Integration	5	0	5	0	0	0	1	2	0	1	0	0
Gestural Behavior	Exploration	0	0	0	0	0	0	0	0	0	0	0	0
	Conflict	1	1	2	4	4	2	1	2	1	1	0	4
	Integration	0	0	1	0	1	1	2	1	0	0	1	7
Tone Change	Exploration	0	0	0	0	1	0	0	0	0	1	0	1
	Conflict	0	2	4	6	8	2	4	7	2	0	0	1
	Integration	2	0	0	7	1	2	1	5	1	2	3	0

Fig. 7. A heat map that visualizes the distribution of nonverbal behaviors across different interaction stages for each participant. Darker colors indicate a higher frequency of nonverbal behaviors, and each square contains a number representing the total count of these behaviors in each specific case.

ensures that the data are represented on a uniform scale, ranging from 0% to 100%. In examining the time distribution across other stages, the data (Figure 5b) shows a minimum engagement time of 1.31% in **exploration** stage, with a maximum of 7.89%. For the **conflict** stage, the interaction time varied ranging from a minimum of 19.35% to to a maximum of 58.51%. For the **integration** stage, the interaction time varied ranging from a minimum of 39.64% to a maximum of 76.77%. These results demonstrate that the varied ways in which users navigate through the stages of interaction with the system influence the effectiveness of communication.

6.4.2 Nonverbal Behavior. In our experiment, nonverbal communication was also an important part of user interactions. Figure 6 shows that the most frequently observed nonverbal behavior was eye contact (M = 21.00, SD = 7.24). This form of communication was particularly common during the **integration** stage when users consistently looked towards the device, especially in the cases of P4, 5, 8, 12, as shown in Figure 7. Such behavior was often noted when users asked questions to the VA, likely stemming from an instinctive tendency to "face" their conversational partner.

Additionally, spatial behavior was another significant nonverbal element, marked by users altering their body position relative to the device during interactions (M = 5.25, SD = 4.03). This shift was predominantly observed in the **conflict** stage. In instances of conflict, users commonly lean or move closer to the Alexa device to facilitate more precise communication. Specifically, P9, who spent a prolonged amount of time in the **conflict** stage and experienced lower response

accuracy, demonstrated the most pronounced spatial adjustments leaning closer to the Alexa device during this phase among other participants.

Furthermore, we also observed the use of gestural behavior among the users (M = 2.58, SD = 3.29). Users utilized gestures to reinforce their verbal interaction or to convey specific intentions, especially during the **conflict** stage. It appeared that users subconsciously incorporated hand gestures as a tool to assist their communication when verbal clarity was challenged.

Lastly, tone change also emerged as an element of nonverbal interaction among users (M=2.50, SD=1.68). During the **conflict** stage, 9 out of 12 participants raised their volume as an expression of frustration or in an effort to capture the voice assistant's attention when it is unresponsiveness (shown in Figure 7). Conversely, users softened their tone in response to receiving useful information in **integration** stage, sometimes coupling a gentler tone with expressions of gratitude such as saying "thank you" to the system.

7 STAGE TRANSITION

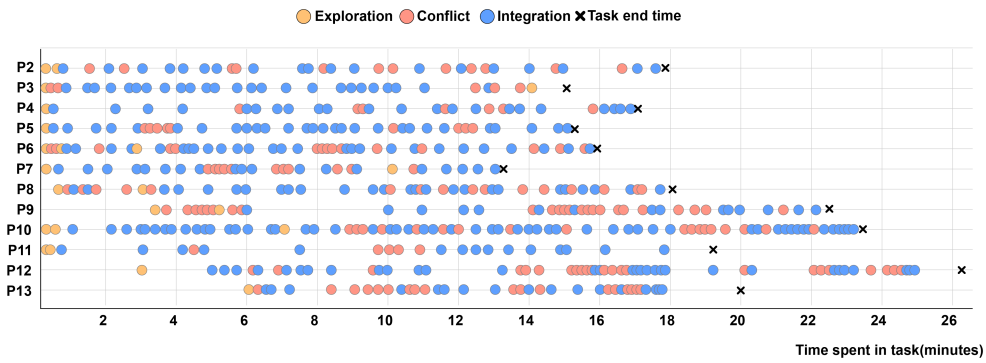


Fig. 8. The scatter plot timelines show a visual overview of the stages experienced by each participant throughout the task, laid out in a timeline format. The vertical axis lists the participants, while the horizontal axis shows the task time in minutes. While all participants begin at a zero point, their activities conclude at varying times. 'X' marks indicate task completion time. The distance between a zero point and an 'X' for each participant reflects the duration of task engagement.

In this section, we utilize the annotated data to explore the patterns of stage transition within our framework. This includes examining the ascending from **conflict** stage to **integration** stage and regressing from **integration** to **conflict** stage. These transitions represent the third dimension of our analytical framework.

The results from the three interaction stages identified in the previous section show the transition of user behavior toward the **integration** stage. This stage is considered ideal for participants to achieve during their interactions with MM, as the interaction is seamless and clear in this stage. However, our findings indicate that not all participants allocate significant amounts of time to the **integration** stage.

Despite this variability, we observed a trend in participant interaction times in the **integration** stage (shown in Figure 5b). On average, participants spent the most time in the **integration** stage (M=63.0%, SD=10.88). We especially observed that P3,4,5 dedicated a substantial portion of their time to the **integration** stage of the task, respectively, 74.36%, 75.09%, and 76.77%. However, P9 and P13, who encountered significant difficulties while using MM, spent more than half of their

total interaction time in the **conflict** stage, 50.97% and 58.51% correspondingly, and showed the least accurate responses in the communication.

Observing this trend, we aim to further analyze behavior as the third dimension in this framework, focusing on the transition of the stages. In this section, we will examine two key aspects: the progression from the Conflict Stage to the Integration Stage and the regression from the Integration Stage back to the Conflict Stage. Within the progression aspect, we will also explore the strategies participants employed to evolve toward the Integration Stage. Figure 8 illustrates an overview of each participant's progression through various stages labeled according to the code book established in the previous section. By examining participants' behaviors during these transitions, we aim to uncover the methods that facilitate their successful engagement with LLM-VAs.

7.1 Ascending From Conflict Stage to Integration Stage

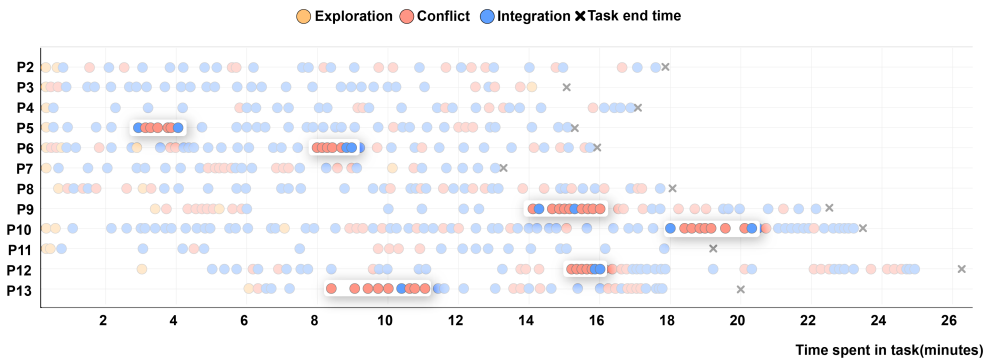


Fig. 9. The scatter plot showcases instances where participants have shown ascending transition of the stage. Each highlighted segment on the timeline signifies a transitional moment in the participant's journey, with the sequence of these segments denoting an ascending shift from the conflict stage to the integration stage.

In our study, participants navigated through the stages dynamically during the experimental task, including evolving between the **conflict** and **integration** stages. The most frequent stage transitions observed among participants involved experiencing multiple **conflict** stages prior to advancing to the **integration** stage. We observed this transition with participants who encountered multiple conflicts while communicating with MM, leading to the continued presence of the **conflict** stage, examples shown in Figure 9. This transition often occurred as users faced multiple conflicts during interactions with MM, resulting in continued conflict. Typically, participants first identified the reasons that cause the conflict, which may be caused by the LLM's extensive knowledge base leading to the interpretation of the question related to a different topic, queries that exceed MM's capabilities, or Alexa system issues unrelated to the MM system itself. During this process, participants tried to extract useful information from their interactions and discover potential solutions. For instance, it took P10 seven rounds of dialogue and P12 five rounds of dialogue to realize they had exited the MM. In this process, participants engaged in a cycle where they continued refining their strategies based on insights from previous experience, which will be discussed later in this section. This process allowed them to adjust their communication techniques effectively, facilitating a smooth transition to the **integration** stage.

Spatial behavior emerged as a nonverbal behavior in this transition of user interaction. This change in posture was primarily evident as the participants navigated the change in their experience.

For example, to figure out the cause of the conflict, participants would lean or move closer toward the Alexa device. This propensity to lean forward was noted in P5, P6, and P9 as a reaction to the absence of feedback from MM. Thus, spatial behavior, especially in the form of leaning closer to the device during **conflict** stages, indicates the change in interaction.

Overall, participants are actively seeking to move beyond **conflict** stage, aiming for an advancement to **integration** stage. We observed that participants employed different strategies to facilitate the stage transition throughout the task.

7.1.1 Re-frame the Question. We observe that participants often modified their queries when interacting with VAs, particularly when they received responses from the system that indicated a misunderstanding or that were incorrect. Participants made a total of 63 attempts using these strategies of re-framing the question in different ways. Although only 46.03% of the time was successful in getting valid responses, participants must experiment with different ways of reframing questions to learn the appropriate strategies over time. In these cases, the participants typically adapted by rephrasing or reformulating their queries. This behavior was an attempt by participants to make their requests more understandable to the system, thereby increasing the likelihood of reaching the **integration** stage. Through a process of trial and error, participants gradually learned their own most effective communication methods with MM. These modifications might include counting the steps, describing the current situation in the queries, and even interrupting the system to pose a refined question.

Participants sometimes enhanced their interactions with MM by including step numbers in their queries, leading to more structured communication and potentially more accurate responses from MM. This approach was demonstrated by P4 , who explicitly used step numbers in their questions to ensure they received the specific instruction they sought (shown in Figure 10). For instance, P4 began by generally asking about the next step. However, when an error occurred, the participant found it beneficial to directly request the next step by specifically referencing the “fifth step". This reliance on step numbers proved to be a practical method for obtaining precise responses, particularly in situations where maintaining the sequence of cooking steps was crucial.

```
01   P4:   ((Counting the number of Ingredients placed with the knife.))
02   P4:   Alexa, what is the fifth step?
03   MM:   The fifth step is to place slices of cucumber into the salad bowl along with the other ingredients.
```

Fig. 10. P4 used a knife to count the step numbers and referenced the step numbers in the questions.

Participants also enhance their interactions with MM by supplementing their queries with context or descriptions of their current circumstances, aiming for more precise and situationally relevant responses. This approach is particularly common when the information needed is not readily available in a provided resource, such as a recipe. For example, participants may lack basic cooking knowledge. In these cases, they describe their background and the context of their situation

before asking their question. By employing this strategy, participants are more likely to receive effective responses that are tailored to their specific circumstances.

We also observed the use of gestural behavior among the participants when explaining the current situation to MM. Participants utilized gestures to reinforce their verbal interaction or to convey specific intentions, especially when participants are trying to evolve beyond **conflict** stage. Figure 11 shows an example of P2 employing hand gestures while interacting with MM.

```
01   P2:   Alexa, I want to (0.4) I want to make the lettuce dry without some water but I don't
           have a spinner, how can I do it?
           ((                              moving hands                                    ))
02   MM:   After rinsing the lettuce, gently shake off any excess water over the sink. Two, place
           the lettuce leaves on a clean kitchen towel or paper towel...
```

Fig. 11. P2 employing hand gestures during interaction with MM while engaged in the task.

When participants had already received part of the information they were seeking, but MM continued its response, they would often interrupt MM mid-sentence as soon as they heard the answer they were looking for. They utilized the wake word to pause the system effectively and promptly posed a new follow-up question that more closely aligned with the specific answer they were searching for. This strategy of interruption served to minimize the wait for the system to complete its response, thus facilitating quicker access to the required information. Especially in situations where participants had specific queries, this method effectively directed the conversation toward a more relevant and expedient resolution.

```
01   P12:   Alexa, what's the last step?
02   (6.1) ((Continue working on the salad task.))
03   P12:   Alexa, what's the last step?
04   (11.5) ((Continue working on the salad task.))
05   P12:   Alexa, what are we making?
     ((Continue working on the salad task.))
06   MM:   Hmm. I don't know that one.
07   P12:   Alexa, stop. (2.4) Alexa, open mango mango.
     ((Continue working on the salad task.))
```

Fig. 12. P12 conversation fragment.

7.1.2 Reactivated MM. Participants who encountered irregular activities or conversations unrelated to their current task often reactivated the device using various input methods. These included a combination of voice commands, touch inputs, or visual cues, which enhanced the interaction and promoted clear communication. During such instances, participants discovered that restarting the system was beneficial. This reboot typically helped align the system's responses with their

expectations and led to more accurate and relevant information related to the working task. Participants perceived the blue status lights around Alexa as indicators of system activity. However, these lights did not always provide sufficient information. For example, the blue status light was sometimes not informative when Alexa was not operating within the MM system. In these cases, participants relied on conversational interaction to discern Alexa's current state. For instance, Figure 12 shows that P12 learned the technique of reactivating MM over time, making it easier to recognize signs indicating that a system restart might be necessary to obtain accurate answers. As the task progressed, P12 increasingly employed this technique.

7.1.3 Repeat the Question. In addition to reframing queries to obtain valid responses, we noticed a pattern where participants often repeated their questions, particularly when they received no feedback from MM. Participants are actively exploring different possibilities to understand the lack of response. Several factors might contribute to this situation; for example, MM might not have heard the question due to system or environmental issues or failed to recognize the wake word. In such instances, humans frequently repeat their questions to ascertain whether the listener comprehends them and can proceed with the task at hand. This strategy reflects an adaptive approach to communication. By adjusting their interactions in real-time, participants aim to overcome these challenges, ensuring that the transition to the **integration** stage of the dialogue is as seamless as possible.

7.1.4 Calling the Wake Word-"Alexa". Another common strategy among participants was quickly re-engaged with MM after forgetting to use the wake word in their initial interaction. This behavior was identified in 7 out of 12 participants. Notably, the participants effectively rectified 84.2% of the attempts. We observe that participants typically recognized and rectified this oversight almost instantaneously, where participants noticed Alexa's non-response in merely a second and subsequently corrected their approach. This is because participants engage with Alexa using patterns akin to human-human conversation, wherein the necessity to utter a wake word before every command is often overlooked.

7.2 Regressing From Integration Stage to Conflict Stage

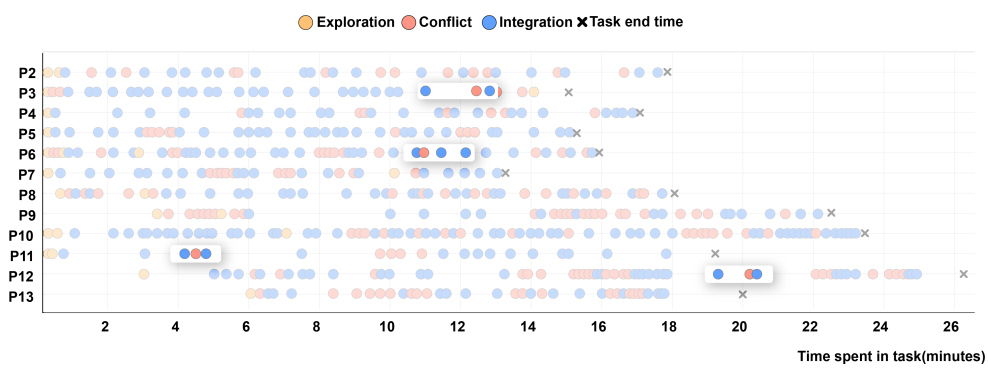


Fig. 13. The scatter plot showcases instances where participants have shown regressing transition in stage. Each highlighted segment on the timeline signifies a transitional moment in the participant's journey, with the sequence of these segments denoting a regressing shift from the integration stage to the conflict stage.

We also observed that a few participants unexpectedly transitioned out of the **integration** stage and into the **conflict** stage after a period of successful engagement, examples shown in Figure 13.

These shifts commonly resulted from one-time events, such as unexpected errors from LLM due to its extensive knowledge base. Figure 14 shows that P3 experienced unexpected errors when MM misinterpreted the question, leading to a response that did not align with the expectations, causing a transition from the **integration** stage to the **conflict** stage.

```
01   P3:   Alexa, am I done?
02   MM:   No, you are not done. Everyone has different
           knowledge and experiences. Asking questions
           is a great way to learn and improve your
           understanding. Keep asking questions and never
           hesitate to seek knowledge.
```

Fig. 14. P3's conversation fragment on experiencing unexpected errors.

In additional instances, P6 and P12 faced regressions when the LLM incorrectly ordered the steps necessary for task completion, leading to the participants unexpectedly receiving incorrect responses. Moreover, P11 encountered a situation where the MM prompted with "Would you like to know what the next steps are?" To which the participant affirmatively responded with "Yes, please," not realizing the necessity of invoking a wake word to initiate further communication.

Despite encountering conflicts that led them unexpectedly out of the **integration** stage, participants can still effortlessly maintain their dialogue with MM. The ability of LLM to handle multi-turn interactions and maintain a record of the conversation history enabled the conversation to be carried on effectively. This ability to quickly resolve disruptions maintains the interaction flow and ensures that participants can smoothly return to the **integration** stage. Consequently, MM demonstrated dynamic adaptability, effectively managing minor conflicts that participants experienced and maintaining engagement with participants.

8 SUMMARY OF FINDINGS

To summarize, we analyzed participant interactions with the LLM-VA in a cooking scenario using our analytic framework's three dimensions: Behavior Characteristics, Three Interaction Stages, and Stage Transitions.

The first dimension, Behavior Characteristics, focuses on verbal and nonverbal behaviors. Verbal behaviors, such as user queries and interactions with the VA and VA responses, shifted as participants progressed through the stages. Nonverbal behaviors were analyzed through eye contact, spatial movements, gestures, and tone changes.

The second dimension, Three Interaction Stages, captured how participants interacted with the VA across the exploration, conflict, and integration stages. In the exploration stage, participants were still familiarizing themselves with the system, often encountering difficulties activating the VA and asking questions beyond the recipe to understand its capabilities. Eye contact was the dominant nonverbal behavior as participants engaged with the VA. In the conflict stage, communication breakdowns occurred, including incorrect responses, misunderstandings, and incomplete recognition of user input. This stage had the richest mix of verbal and nonverbal behaviors, as participants expressed frustration, interrupted the VA, and proceeded without confirmation. Nonverbal cues like tone changes, spatial adjustments (e.g., moving closer to the VA), and gestures (e.g., tapping or waving) were frequently used as participants tried to resolve issues and move the task forward. In the integration stage, participant interactions became more focused and efficient. Verbal behaviors included sophisticated queries and follow-up questions, while nonverbal behaviors, such as eye contact, were used to confirm understanding. This stage represented the most seamless interaction, as participants had adapted to the system.

The third dimension, Stage Transitions, demonstrated how participants adjusted their communication strategies when moving between stages. While the integration stage was the most ideal and where participants spent the most time, this was not true for everyone. Some participants remained in the conflict stage for extended periods. Transitioning from conflict to integration typically required participants to refine their communication, such as rephrasing queries, providing additional context, and adjusting spatial behaviors to resolve issues. These strategies helped them progress to the more efficient integration stage. However, transitions were not always smooth. Some participants reverted back to the conflict stage due to unexpected errors or lapses, such as forgetting to use the wake word during extended periods of smooth interaction.

9 DISCUSSION

In this section, we summarize the key contributions of our study, highlight design implications from the research, and discuss potential future directions. Finally, we address the limitations of our research and future work.

9.1 Design Implications

Our analytical framework, focusing on behavior characteristics, three interaction stages, and stage transition, provides a foundational basis for future research and practical applications of LLM-VA. By leveraging both verbal and nonverbal behaviors, it offers valuable insights into user engagement with LLM-VA systems during tasks. These insights help developers and researchers create more intuitive and user-centric technologies, emphasizing the importance of integrating these behaviors into the design and assessment of interactive systems. Based on this knowledge, we present design implications to guide the development of more effective human-VA interactions.

9.1.1 Understanding User: Behavior & Emotion Prediction. In the analysis of the three interaction dimensions, the observation reveals that user emotions fluctuate throughout the task flow (e.g., experiencing confusion when the LLM-VA does not respond). Previous research has incorporated nonverbal signals into models for computing emotions, identifying specific emotional states like happiness or anxiety, and providing corresponding emotional support [76]. Our framework builds upon this by integrating nonverbal behaviors, enabling LLM-VAs to achieve a more comprehensive understanding of human behavior and establish user behavioral patterns.

Additionally, long-term observation of repeatable user behavior patterns is advantageous for understanding the unique behavioral characteristics of different individuals and developing personalized user behavior predictions. Recognizing specific nonverbal signals at various stages can help LLM-VAs adopt more targeted strategies. These strategies might include guiding conversations more proactively or in a calming manner, based on the user's emotional state and behavior. Through this approach, LLM-VAs can recognize immediate emotional states and use the dimensions in the framework to predict potential shifts in user emotions based on the progression of their interactions, making adjustments tailored to the current situation.

9.1.2 Responding More: Real-time Emotional Feedback from VA. VAs that analyze users' nonverbal behavior and apply our established behavioral stages can enhance the quality of interaction by providing emotionally attuned responses. Previous research has looked into the use of single multimodal input and output systems in VA design, like gaze and gesture [4, 49, 73]. However, more input could be used within this process. For example, if a LLM-VA detects signs of anxiety or frustration during the **conflict** stage, it could adopt a reassuring tone, showing concern for the user's emotions and a commitment to assisting with the problem. Conversely, in response to a user's excitement or joy, the LLM-VA could mirror this positivity with an energetic and cheerful tone. Accurately interpreting different stages of behavioral analysis and users' nonverbal behavior

allows VAs to provide precise and empathetic emotional responses. These approaches align with the expectation that humans have for receiving social responses from computers [91]. Prior research has [36, 87] highlighted users' preference for continued interaction with computer agents as well, which demonstrated empathy, sympathy, and active listening, especially in moments of frustration.

9.1.3 Bridging Minds: Theory of Mind Implications for LLM-VA Design. Our findings have important implications for LLM-VA design when viewed through the lens of the mutual theory of mind (ToM) [108]. The observed interactions across the **exploration**, **conflict**, and **integration** stages reveal how users develop and refine their understanding of the VA's "mind" and capabilities. In the **exploration** stage, users build their initial ToM of the VA, as evidenced by participants asking questions beyond the recipe scope to gauge the VA's knowledge boundaries. This suggests that VAs should provide clear feedback about their capabilities early on. The **conflict** stage, marked by communication breakdowns and rich nonverbal behaviors, represents moments when users' ToM is challenged. For instance, in Figure 15, a participant is shown using a hand stop gesture while expressing frustration with the VA's overly detailed response. This nonverbal cue, coupled with the participant's verbal request to simplify the instruction ("Only tell me the step after adding the vinegar"), exemplifies how users attempt to manage and recalibrate their mental model of the VA during these challenging interactions. Such behaviors suggest that VAs must recognize and respond to these cues to better support users and reduce frustration. The **integration** stage demonstrates a more mature ToM, where participants used sophisticated queries and efficient eye contact to confirm understanding. The dynamic nature of ToM, evident in participants reverting to the **conflict** stage after forgetting the wake word, implies that VAs should maintain behavioral consistency while handling occasional lapses.

```
01   P2:   °Okay° Alexa, tell me the rest of the steps after adding the vinegar.
02   (6.1)
03   MM:   Once you have added the vinegar, the next step is to add half teaspoon of Dijon mustard
           and half teaspoon of honey to the mason jar. After that proceed to add the minced
           garlic, one quarter garlic clove or one quarter teaspoon minced garlic, followed by
           [one quarter teaspoon of sea salt and one sixteenth]
04   P2:   [Wow wow wow wow. (0.8) Alexa.]
           (( Making hand stop gesture ))
05   (1.7)
06   P2:   Only tell me the step after adding the vinegar.
```

Fig. 15. P2 holding hand stop gesture.

Additionally, LLM-VAs should be designed to develop their own theory of mind about users (also called the "human state" [37]) to provide more effective support. In the **exploration** stage, VAs should track the types and complexity of user questions to gauge their familiarity with the system and adjust explanations accordingly. During the **conflict** stage, VAs need to recognize and interpret users' nonverbal cues, such as tone changes or spatial adjustments, as indicators of frustration or confusion. Building on prior research that has demonstrated the significant

impact of communication styles (e.g., formal or informal) on user perceptions [25, 30], our findings from the **integration** stage suggest that VAs should be designed to adapt their communication approach based on the needs of the users. Specifically, VAs should learn from successful interactions, identifying and adopting the communication styles and query response patterns that prove most effective for individual users. This adaptive behavior could be implemented through reinforcement learning techniques, allowing the VA to continuously refine its user model. To address the dynamic nature of interactions, VAs should maintain a flexible, real-time model of the user's cognitive and emotional state, potentially using techniques from affective computing. This model should inform the VA's responses, allowing it to preemptively offer help when it detects potential confusion or to provide more advanced features as it recognizes increasing user expertise. By incorporating these design elements, LLM-VAs can build and maintain their own theory of mind about users, leading to more personalized, context-aware, and ultimately more effective interactions.

9.1.4 Adaptive LLM-VA Personas. Our findings extend previous research on adaptive VA personas [35, 36, 52] by demonstrating how VAs' responsiveness to users' verbal and non-verbal signals across interaction stages can significantly enhance user experience and interaction efficiency. During the **exploration** stage, when users pose questions beyond the recipe scope, VAs should recognize this as a cue to adopt a more informative and patient persona, perhaps switching to a 'teacher' mode with detailed explanations. In the **conflict** stage, where users express frustration through tone or word choice, VAs could shift to an empathetic, problem-solving persona, using reassuring language and offering step-by-step guidance. Perhaps depending on individual users, humor can be used to diffuse the tension as well [116]. Non-verbal cues, particularly evident during conflicts, such as users moving closer to the device or making exaggerated gestures, should prompt VAs to become more attentive and responsive, potentially increasing response speed or adjusting tone to match the user's urgency. As users transition between stages, especially from **conflict** to **integration**, VAs should smoothly adjust their persona to match their evolving state, adopting a more collegial approach as users demonstrate increased proficiency. Throughout these adaptations, VAs should maintain a core persona for continuity, balancing subtle yet meaningful changes to address changing user needs. Over time, VAs should learn from successful interactions and user preferences to develop personalized adaptive strategies, accounting for which persona adjustments worked best for specific user states and applying them proactively in future interactions.

9.2 Generalizability

Cooking serves as a fundamental and practical context for exploring human and VA interactions, offering initial insights into interactions with LLM-VAs [81]. While our analytical framework was initially studied within cooking scenarios, its applicability shows abilities to extend beyond this context. Our findings can inform subsequent tests and validations of LLM-VA interactions in various natural communication settings, such as education, healthcare, driving, and more. Our analytical framework can guide the design of scenario-based tests to evaluate metrics like user experience. By using the analytical framework, we can identify the disadvantages in the interaction process between human and LLM-VA. These insights can be strategically leveraged to enhance the design of LLM-VAs in various task-oriented settings. Our findings indicate that as task complexity increases, users are likely to experience more frequent **conflict** stages in their interactions with the LLM-VA, likely due to the increasing complexity of their voice commands. Therefore, collaborative scenarios involving humans and LLM-VAs, such as driving, predicting task complexity, and providing users with preemptive guidance on formulating voice commands could lead to smoother interactions. For instance, our findings show that as task complexity increases, users are likely to encounter more frequent difficulties during interactions with the VA due to the increasing complexity of their

voice commands. By applying this analytical framework, researchers and system designers can identify disadvantages in the interaction between humans and LLM-VA and strategically use these insights to enhance the design of LLM-VAs across various task-oriented settings.

9.3 Limitation and Future Work

While our study provides valuable insights into human interactions with LLM-VAs in cooking tasks, it has certain limitations that must be acknowledged.

First, our study was conducted with limited participants, which may not fully represent the diverse range of users interacting with LLM-VAs. Additionally, focusing exclusively on cooking tasks may limit the applicability of our findings to other contexts where LLM-VAs are used. Cooking is a specific activity with its own challenges and user interactions, which might not fully encapsulate all the scenarios in which people use LLM-VAs. Furthermore, this analytical framework was developed based on observations in this specific context. Further validation may be required to assess its applicability to other tasks, technologies, or user groups.

To build on this research, future studies could address these limitations in several ways. Conducting the study with a larger and more diverse group of participants, including different age groups, cultural backgrounds, and varying levels of familiarity with technology, would enhance the generalizability of the findings. Additionally, applying the study to different contexts and tasks where LLM-VAs are used could provide insights into how interaction patterns vary across different activities. By addressing these limitations, future research can further contribute to the field of human-computer interaction, particularly in enhancing the design and functionality of LLM-VAs for various practical applications.

10 CONCLUSION

In this research, we bridge the gap between VAs and the nuanced, multi-modal nature of human communication by developing an analytical framework that captures the dynamic interplay of verbal and nonverbal behaviors in human interactions with LLM-VAs. Through our study, we demonstrated that users naturally rely on both forms of communication when performing complex tasks, such as cooking, with LLM-VAs. We identified key behavior characteristics, developed a three-stage model (Exploration, Conflict, and Integration), and provided insights into how users transition between these stages.

These findings highlight the potential of incorporating multimodal inputs into future designs, enabling LLM-VAs to be more personalized, socially aware, and responsive to users' mental states and communication styles. As LLMs become increasingly integrated into voice assistants, our work sets the stage for deeper exploration of human-LLM-VA interactions and offers practical guidance for designing more adaptive, user-friendly LLM-VAs that enhance both user experience and system performance.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Institute On Minority Health And Health Disparities of the National Institutes of Health under Award Number R01MD018424. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

[1] Drew H Abney, Rick Dale, Max M Louwerse, and Christopher T Kello. 2018. The bursts and lulls of multimodal interaction: Temporal distributions of behavior reveal differences between verbal and non-verbal communication. *Cognitive science* 42, 4 (2018), 1297–1316.

[2] Hassan Ali, Philipp Allgeuer, and Stefan Wermter. 2024. Comparing Apples to Oranges: LLM-powered Multimodal Intention Prediction in an Object Categorization Task. *arXiv preprint arXiv:2404.08424* (2024).

[3] James Allen. 1995. *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc.

[4] Víctor Alvarez-Santos, Roberto Iglesias, Xose Manuel Pardo, Carlos V Regueiro, and Adrián Canedo-Rodriguez. 2014. Gesture-based interaction with voice feedback for a tour-guide robot. *Journal of Visual Communication and Image Representation* 25, 2 (2014), 499–509.

[5] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. 2019. Music, search, and IoT: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 3 (2019), 1–28.

[6] Dane Archer. 1997. Unspoken diversity: Cultural differences in gestures. *Qualitative sociology* 20 (1997), 79–105.

[7] Dane Archer and Robin M Akert. 1977. Words and everything else: Verbal and nonverbal cues in social interpretation. *Journal of personality and social psychology* 35, 6 (1977), 443.

[8] Michael Argyle. 2013. *Bodily communication*. Routledge.

[9] 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.

[10] Brooke Auxier. [n. d.]. 5 things to know about Americans and their smart speakers. https://www.pewresearch.org/fact-tank/2019/11/21/5-things-to-know-about-americans-and-their-smart-speakers/

[11] Vince Bartle, Janice Lyu, Freesoul El Shabazz-Thompson, Yunmin Oh, Angela Anqi Chen, Yu-Jan Chang, Kenneth Holstein, and Nicola Dell. 2022. "A Second Voice": Investigating Opportunities and Challenges for Interactive Voice Assistants to Support Home Health Aides. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[12] Diana Beirl, Y Rogers, and Nicola Yuill. 2019. Using voice assistant skills in family life. In *Computer-Supported Collaborative Learning Conference, CSCL*, Vol. 1. International Society of the Learning Sciences, Inc., 96–103.

[13] Erin Beneteau, Ashley Boone, Yuxing Wu, Julie A Kientz, Jason Yip, and Alexis Hiniker. 2020. Parenting with Alexa: exploring the introduction of smart speakers on family dynamics. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[14] 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.

[15] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.

[16] Michael Bonfert, Nima Zargham, Florian Saade, Robert Porzel, and Rainer Malaka. 2021. An Evaluation of Visual Embodiment for Voice Assistants on Smart Displays. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. 1–11.

[17] 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.

[18] John L. Campbell, Charles Quincy, Jordan Osserman, and Ove K. Pedersen. 2013. Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement. *Sociological Methods & Research* 42, 3 (Aug. 2013), 294–320. https://doi.org/10.1177/0049124113500475

[19] Clare Carroll, Catherine Chiodo, Adena Xin Lin, Meg Nidever, and Jayanth Prathipati. 2017. Robin: enabling independence for individuals with cognitive disabilities using voice assistive technology. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 46–53.

[20] Justine Cassell, Yukiko I Nakano, Timothy W Bickmore, Candace L Sidner, and Charles Rich. 2001. Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. 114–123.

[21] 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).

[22] Tanya L Chartrand and John A Bargh. 1999. The chameleon effect: The perception–behavior link and social interaction. *Journal of personality and social psychology* 76, 6 (1999), 893.

[23] Chaoran Chen, Bingsheng Yao, Yanfang Ye, Dakuo Wang, and Toby Jia-Jun Li. [n. d.]. Evaluating the LLM Agents for Simulating Humanoid Behavior. ([n. d.]).

[24] Jen-Hao Chen, Peggy Pei-Yu Chi, Hao-Hua Chu, Cheryl Chia-Hui Chen, and Polly Huang. 2010. A smart kitchen for nutrition-aware cooking. *IEEE Pervasive Computing* 9, 4 (2010), 58–65.

[25] Jessie Chin, Smit Desai, Sheny (Cheng-Hsuan) Lin, and Shannon Mejia. 2024. Like My Aunt Dorothy: Effects of Conversational Styles on Perceptions, Acceptance and Metaphorical Descriptions of Voice Assistants during Later Adulthood. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 88:1–88:21. https://doi.org/10.1145/3637365

[26] Janghee Cho and Emilee Rader. 2020. The role of conversational grounding in supporting symbiosis between people and digital assistants. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.

[27] Minji Cho, Sang-su Lee, and Kun-Pyo Lee. 2019. Once a kind friend is now a thing: Understanding how conversational agents at home are forgotten. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 1557–1569.

[28] Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. (1991).

[29] 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.

[30] Samuel Rhys Cox and Wei Tsang Ooi. 2022. Does Chatbot Language Formality Affect Users' Self-Disclosure?. In *Proceedings of the 4th Conference on Conversational User Interfaces*. ACM, Glasgow United Kingdom, 1–13. https://doi.org/10.1145/3543829.3543831

[31] 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.

[32] Andrea Cuadra, Hansol Lee, Jason Cho, and Wendy Ju. 2021. Look at Me When I Talk to You: A Video Dataset to Enable Voice Assistants to Recognize Errors. *arXiv preprint arXiv:2104.07153* (2021).

[33] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 902–909.

[34] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. arXiv:2209.01390 [cs.HC]

[35] Smit Desai, Mateusz Dubiel, and Luis A. Leiva. 2024. Examining Humanness as a Metaphor to Design Voice User Interfaces. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3640794.3665535

[36] Smit Desai and Michael Twidale. 2023. Metaphors in Voice User Interfaces: A Slippery Fish. *ACM Transactions on Computer-Human Interaction* 30, 6 (Sept. 2023), 89:1–89:37. https://doi.org/10.1145/3609326

[37] Sandra Devin and Rachid Alami. 2016. An Implemented Theory of Mind to Improve Human-Robot Shared Plans Execution. In *The Eleventh ACM/IEEE International Conference on Human Robot Interation (The Eleventh ACM/IEEE International Conference on Human Robot Interation)*. IEEE Press, Christchurch, New Zealand, 319–326. https://doi.org/10.1109/HRI.2016.7451768

[38] Philip R Doyle, Leigh Clark, and Benjamin R Cowan. 2021. What do we see in them? identifying dimensions of partner models for speech interfaces using a psycholexical approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[39] Starkey Duncan and Donald W Fiske. 2015. *Face-to-face interaction: Research, methods, and theory*. Routledge.

[40] Paul Ekman and Wallace V Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica* 1, 1 (1969), 49–98.

[41] Iván Esteban-Lozano, Álvaro Castro-González, and Paloma Martínez. 2024. Using a LLM-Based Conversational Agent in the Social Robot Mini. In *International Conference on Human-Computer Interaction*. Springer, 15–26.

[42] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods* 5, 1 (2006), 80–92.

[43] John Fiske. 2010. *Introduction to communication studies*. Routledge.

[44] Rose Guingrich and Michael SA Graziano. 2023. Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits in machines. *arXiv preprint arXiv:2311.10599* (2023).

[45] Reiko Hamada, Jun Okabe, Ichiro Ide, Shin'ichi Satoh, Shuichi Sakai, and Hidehiko Tanaka. 2005. Cooking navi: assistant for daily cooking in kitchen. In *Proceedings of the 13th annual ACM international conference on Multimedia*. 371–374.

[46] 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.

[47] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. 2018. Interactively picking real-world objects with unconstrained spoken language instructions. In *2018 IEEE*

International Conference on Robotics and Automation (ICRA). IEEE, 3774–3781.

[48] Shaoshuai Huang, Xuandong Zhao, Dapeng Wei, Xinheng Song, and Yuanbo Sun. 2024. Chatbot and Fatigued Driver: Exploring the Use of LLM-Based Voice Assistants for Driving Fatigue. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[49] Razan Jaber. 2023. *Towards Designing Better Speech Agent Interaction: Using Eye Gaze for Interaction*. Ph. D. Dissertation. Department of Computer and Systems Sciences, Stockholm University.

[50] Razan Jaber, Sabrina Zhong, Sanna Kuoppamäki, Aida Hosseini, Iona Gessinger, Duncan P Brumby, Benjamin R Cowan, and Donald Mcmillan. 2024. Cooking With Agents: Designing Context-aware Voice Interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–13.

[51] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. Promptmaker: Prompt-based prototyping with large language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–8.

[52] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (Oct. 2020), 163:1–163:26. https://doi.org/10.1145/3415234

[53] Callie Y Kim, Christine P Lee, and Bilge Mutlu. 2024. Understanding large-language model (llm)-powered human-robot interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 371–380.

[54] 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.

[55] Mark L. Knapp, Judith A. Hall, and Terrence G. Horgan. 2013. *Nonverbal Communication in Human Interaction*. Cengage Learning.

[56] Dimosthenis Kontogiorgos, Andre Pereira, Olle Andersson, Marco Koivisto, Elena Gonzalez Rabal, Ville Vartiainen, and Joakim Gustafson. 2019. The effects of anthropomorphism and non-verbal social behaviour in virtual assistants. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 133–140.

[57] Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne van Waveren, and Joakim Gustafson. 2020. Behavioural responses to robot conversational failures. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 53–62.

[58] Thomas Kosch, Kevin Wennrich, Daniel Topp, Marcel Muntzinger, and Albrecht Schmidt. 2019. The digital cooking coach: using visual and auditory in-situ instructions to assist cognitively impaired during cooking. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 156–163.

[59] Franklin B Krohn. 2004. A generational approach to using emoticons as nonverbal communication. *Journal of technical writing and communication* 34, 4 (2004), 321–328.

[60] Emily Kuang, Ehsan Jahangirzadeh Soure, Mingming Fan, Jian Zhao, and Kristen Shinohara. 2023. Collaboration with Conversational AI Assistants for UX Evaluation: Questions and How to Ask them (Voice vs. Text). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[61] Duong Minh Le, Ruohao Guo, Wei Xu, and Alan Ritter. 2023. Improved Instruction Ordering in Recipe-Grounded Conversation. *arXiv preprint arXiv:2305.17280* (2023).

[62] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus* 15, 6 (2023).

[63] Q Vera Liao, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, N Sadat Shami, and Werner Geyer. 2018. All work and no play?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[64] Zhe Liu, Chunyang Chen, Junjie Wang, Mengzhuo Chen, Boyu Wu, Xing Che, Dandan Wang, and Qing Wang. 2024. Make llm a testing expert: Bringing human-like interaction to mobile gui testing via functionality-aware decisions. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. 1–13.

[65] Robert H Logie, AS Law, Steven Trawley, and Jack Nissan. 2010. Multitasking, working memory and remembering intentions. *Psychologica Belgica* 50, 3-4 (2010), 309–326.

[66] 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.

[67] Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. 770–780.

[68] Dominic W Massaro, Michael M Cohen, Sharon Daniel, and Ronald A Cole. 1999. Developing and evaluating conversational agents. In *Human performance and ergonomics*. Elsevier, 173–194.

[69] 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.

[70] David Matsumoto and Hyisung C Hwang. 2013. Cultural similarities and differences in emblematic gestures. *Journal of Nonverbal Behavior* 37 (2013), 1–27.

[71] 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.

[72] Derek McColl, Alexander Hong, Naoaki Hatakeyama, Goldie Nejat, and Beno Benhabib. 2016. A survey of autonomous human affect detection methods for social robots engaged in natural HRI. *Journal of Intelligent & Robotic Systems* 82 (2016), 101–133.

[73] Donald McMillan, Barry Brown, Ikkaku Kawaguchi, Razan Jaber, Jordi Solsona Belenguer, and Hideaki Kuzuoka. 2019. Designing with Gaze: Tama – a Gaze Activated Smart-Speaker. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 176 (nov 2019), 26 pages. https://doi.org/10.1145/3359278

[74] Donald McMillan, Barry Brown, Ikkaku Kawaguchi, Razan Jaber, Jordi Solsona Belenguer, and Hideaki Kuzuoka. 2019. Designing with gaze: Tama–a gaze activated smart-speaker. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[75] 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.

[76] Shrikanth Narayanan and Panayiotis G Georgiou. 2013. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proc. IEEE* 101, 5 (2013), 1203–1233.

[77] Clifford Nass and Li Gong. 2000. Speech interfaces from an evolutionary perspective. *Commun. ACM* 43, 9 (2000), 36–43.

[78] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[79] Clifford Nass and Jonathan Steuer. 1993. Voices, boxes, and sources of messages: Computers and social actors. *Human Communication Research* 19, 4 (1993), 504–527.

[80] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435* (2023).

[81] Nils Neumann and Sven Wachsmuth. 2021. Recipe Enrichment: Knowledge Required for a Cooking Assistant.. In *ICAART (2)*. 822–829.

[82] Elnaz Nouri, Adam Fourney, Robert Sim, and Ryen W White. 2019. Supporting complex tasks using multiple devices. In *Proceedings of WSDM'19 Task Intelligence Workshop (TI@ WSDM19)*.

[83] 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.

[84] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023). https://api.semanticscholar.org/CorpusID:257532815

[85] Cansu Oranç and Azzurra Ruggeri. 2021. "Alexa, let me ask you something different" Children's adaptive information search with voice assistants. *Human Behavior and Emerging Technologies* 3, 4 (2021), 595–605.

[86] Jamie Pearson, Jiang Hu, Holly P Branigan, Martin J Pickering, and Clifford I Nass. 2006. Adaptive language behavior in HCI: how expectations and beliefs about a system affect users' word choice. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 1177–1180.

[87] Rosalind W Picard and Jonathan Klein. 2002. Computers that recognise and respond to user emotion: theoretical and practical implications. *Interacting with computers* 14, 2 (2002), 141–169.

[88] Patryk Pomykalski, Mikołaj P. Woźniak, Paweł W. Woźniak, Krzysztof Grudzień, Shengdong Zhao, and Andrzej Romanowski. 2020. Considering Wake Gestures for Smart Assistant Use. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3383089

[89] 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.

[90] 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.

[91] Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK* 10, 10 (1996).

[92] Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering* 3, 1 (1997), 57–87.
[93] Minjin Rheu, Ji Youn Shin, Wei Peng, and Jina Huh-Yoo. 2021. Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human–Computer Interaction* 37, 1 (2021), 81–96.
[94] Aafaq Sabir, Evan Lafontaine, and Anupam Das. 2022. Hey Alexa, Who Am I Talking to?: Analyzing Users' Perception and Awareness Regarding Third-party Alexa Skills. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3517510
[95] Ayaka Sato, Keita Watanabe, and Jun Rekimoto. 2014. MimiCook: a cooking assistant system with situated guidance. In *Proceedings of the 8th international conference on tangible, embedded and embodied interaction*. 121–124.
[96] Shane Saunderson and Goldie Nejat. 2019. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics* 11 (2019), 575–608.
[97] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. 2018. " Hey Alexa, What's Up?" A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 designing interactive systems conference*. 857–868.
[98] Paul Semaan. 2012. Natural language generation: an overview. *J Comput Sci Res* 1, 3 (2012), 50–57.
[99] Mohit Shridhar and David Hsu. 2018. Interactive visual grounding of referring expressions for human-robot interaction. *arXiv preprint arXiv:1806.03831* (2018).
[100] Han Sloetjes and Peter Wittenburg. 2008. Annotation by category-ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*.
[101] Siegfried Ludwig Sporer and Barbara Schwandt. 2006. Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 20, 4 (2006), 421–446.
[102] George Terzopoulos and Maya Satratzemi. 2020. Voice assistants and smart speakers in everyday life and in education. *Informatics in Education* 19, 3 (2020), 473–490.
[103] David R Thomas. 2003. A general inductive approach for qualitative data analysis. (2003).
[104] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[105] 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.
[106] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the role of conversational cues in guided task support with virtual assistants. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–7.
[107] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
[108] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445645
[109] Johanna Weber, Margarita Esau-Held, Marvin Schiller, Eike Martin Thaden, Dietrich Manstetten, and Gunnar Stevens. 2023. Designing an Interaction Concept for Assisted Cooking in Smart Kitchens: Focus on Human Agency, Proactivity, and Multimodality. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 1128–1144.
[110] Rainer Winkler, Matthias Söllner, Maya Lisa Neuweiler, Flavia Conti Rossini, and Jan Marco Leimeister. 2019. Alexa, can you help us solve this problem? How conversations with smart personal assistant tutors increase task group outcomes. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.
[111] Sarah Wiseman and Sandy JJ Gould. 2018. Repurposing emoji for personalised communication: Why means "I love you". In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–10.
[112] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–22.
[113] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2023. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. arXiv:2307.14385 [cs.CL]
[114] Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Hong Yu, James Hendler, Anind K Dey, and Dakuo Wang. 2023. Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385* (2023).
[115] 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.

[116] Nima Zargham, Vino Avanesi, Leon Reicherts, Ava Elizabeth Scott, Yvonne Rogers, and Rainer Malaka. 2023. "Funny How?" A Serious Look at Humor in Conversational Agents. In *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3571884.3603761

[117] Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. 2022. Storybuddy: A human-ai collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–21.

[118] Yaxi Zhao, Razan Jaber, Donald McMillan, and Cosmin Munteanu. 2022. "Rewind to the Jiggling Meat Part": Understanding Voice Control of Instructional Videos in Everyday Tasks. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–11.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009