"*Mango Mango*, How to Let The Lettuce Dry Without A Spinner?": Exploring User Perceptions of Using An LLM-Based Conversational Assistant Toward Cooking Partner

SZEYI CHAN*, Northeastern University, USA
JIACHEN LI*, Northeastern University, USA
BINGSHENG YAO, Northeastern University, USA
AMAMA MAHMOOD, Johns Hopkins University, USA
CHIEN-MING HUANG, Johns Hopkins University, USA
HOLLY JIMISON, Northeastern University, USA
ELIZABETH D MYNATT, Northeastern University, USA
DAKUO WANG†, Northeastern University, USA

The rapid advancement of Large Language Models (LLMs) has created numerous potentials for integration with conversational assistants (CAs) assisting people in their daily tasks, particularly due to their extensive flexibility. However, users' real-world experiences interacting with these assistants remain unexplored. In this research, we chose cooking, a complex daily task, as a scenario to explore people's successful and unsatisfactory experiences while receiving assistance from an LLM-based CA, *Mango Mango*. We discovered that participants value the system's ability to offer customized instructions based on context, provide extensive information beyond the recipe, and assist them in dynamic task planning. However, users expect the system to be more adaptive to oral conversation and provide more suggestive responses to keep them actively involved. Recognizing that users began treating our LLM-CA as a personal assistant or even a partner rather than just a recipe-reading tool, we propose five design considerations for future development.

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

Additional Key Words and Phrases: user study, exploratory study, large language model-based conversational assistant

ACM Reference Format:
Szeyi Chan, Jiachen Li, Bingsheng Yao, Amama Mahmood, Chien-Ming Huang, Holly Jimison, Elizabeth D Mynatt, and Dakuo Wang. 2024. "*Mango Mango*, How to Let The Lettuce Dry Without A Spinner?": Exploring User Perceptions of Using An LLM-Based Conversational Assistant Toward Cooking Partner. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 35 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Current conversational assistants (CAs), such as Amazon's Alexa, Apple's Siri, and Google Assistant, are important in our daily lives, especially in home-based settings [5, 54, 112]. The "hands-free"

*Both authors contributed equally to this research.
†Corresponding author d.wang@northeastern.edu

CSCW '25, Oct 2025, Bergen, Norway

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

and "eyes-free" design enables users to effortlessly access information through voice commands for simple question-answering tasks, including setting reminders, providing weather updates, and searching for recipes [94, 118, 149].

Nevertheless, CAs face limitations and challenges when instructing users with hands-on tasks in family-centered scenarios, such as experimenting with new recipes for special family dinners, resolving urgent plumbing problems, or collaboratively assembling new furniture. These tasks often require fundamental knowledge in areas in which family members may not always have expertise, leading them to seek guidance through online instructional videos or product manuals [15, 73]. In particular, cooking requires steps like preparing food, finding ingredients, measuring the correct amount, and planning, all while the cook's hands are occupied with food preparation [31, 98, 99]. Unfortunately, current CAs cannot provide comprehensive and continuous support with these tasks [117]. Existing CAs rely on predefined dialogue logic and often struggle with language comprehension, prohibiting natural back-and-forth conversations for complex tasks [6, 7, 23, 30, 91, 117, 122]. Therefore, exploring new approaches is essential to effectively address these challenges and enhance the support provided by CAs in such complex, interactive scenarios.

Recent advancements in language models, particularly large language models (LLMs), for example, GPT-3.5/4 [101], LLaMA [121], and PaLM [24], show the ability to overcome the limitations of language models used in current CAs. Existing works have shown LLMs have natural language understanding (NLU) [4] and generation (NLG) [109, 113] capabilities to understand users' lengthy text input and accommodate multi-turn dialogues [141]. Despite significant advancements, the integration of LLMs into CAs for real-world scenarios remains underexplored. Specifically, there is limited understanding of whether CAs with LLMs can address key challenges faced by traditional CAs, such as adapting to diverse user needs and supporting complex, domain-specific tasks like cooking or collaborative problem-solving. While LLMs are good at understanding and generating text, their practical application in CAs requires further explorations to tailor responses to dynamic contexts and ensure usability across diverse interactions.

To explore the integration of LLMs into CAs, our research consists of two parts: 1) developing an LLM-based system, *Mango Mango*, specifically tailored to help individuals cook at home and 2) conducting a mixed-method in lab exploratory study to evaluate users' experiences through preparing for a salad. Following the study, we performed both qualitative and quantitative research analyses with semi-structured interviews, surveys, and system logs. Our research is guided by two primary questions: **(1) How do users perceive LLM-based CA in cooking scenarios through their interaction experiences?** and **(2) What are the design implications of LLM-based CAs aimed at assisting users in real-world practices like cooking?**

The questionnaire results from the study indicate that participants generally have a positive experience using *Mango Mango*. Users' feedback from the interviews shows appreciation for features including receiving aid beyond the recipe, recollection of the current cooking status, personalized instructions, task planning, free control of the cooking process by user preference, etc. However, some design aspects require improvement, including managing information overload from responses, addressing issues with understanding oral expressions, minimizing redundant interactions with the system, facilitating more engaging dialogues with the CAs, and more. Additionally, the study found that users' perceptions of *Mango Mango* changed during their interaction, from perceiving CAs simply as a tool, to a personal assistant, and to a partner. Based on the findings, we discussed design considerations for leveraging LLMs' NLU and NLG capabilities to enhance the effectiveness and usability of LLM-based CAs specifically in cooking applications.

The main contributions of our paper are summarized as follows:

(1) We developed a conversational assistant system that integrates a widely deployed LLM (GPT 3.5-Turbo) to guide users in cooking scenarios.

(2) We conducted a mixed-methods exploratory study with 12 participants in a home kitchen setting to better understand user experiences when using LLM-based CAs in cooking tasks.
(3) We summarized the key themes of successful and unsatisfactory user experiences based on semi-structured interviews.
(4) We provided design implications for future LLM-based CAs in cooking scenarios.

2 RELATED WORK

We first focus on recent developments in CAs designed to meet real-world demands in Section 2.1. Additionally, we discuss the evolution of language models and recent applications developed with LLM in Section 2.2. Lastly, we touch on existing work that utilizes AI techniques to enhance cooking scenarios in Section 2.3.

2.1 CAs for Human: Real-World Challenges

Researchers have been exploring using CAs with language models in real-world situations to assist people in accomplishing daily tasks. CAs applications like chatbots [8, 43, 47, 136, 137, 139] have been developed and tested to successfully assist people in completing various activities. For example, smart CAs have shown promising capability as reliable healthcare technologies for elders [10, 11, 16, 19, 48, 107]. CAs are also used for other scenarios, such as travel [20, 106], music [6], education [26, 35, 37, 38, 40, 60, 142, 147], home [12–14, 112], etc., showing promising utilities [56, 118, 124].

However, challenges in developing CAs are identified primarily due to disparities in user perceptions of the system's capabilities. Issues like speech detection failure and faulty recognition can occur [96, 104]. The use of heuristics in most existing commercial CAs limits the scope of questions that can be answered and constrains the support of basic interaction functionalities (e.g. setting reminders), which can potentially cause users to feel discouraged and lower their expectations of the technology's capabilities [6, 7, 23, 30, 91, 122]. Additionally, current CAs face challenges in responding to queries about external sources, lapses in providing comprehensive details, and lack of ability to provide broader context [58, 76]. Jaber et al. [59] highlights the challenges and importance of context awareness when working on complex tasks such as cooking. Current commercial VAs often fail due to a lack of contextual awareness, leading to irrelevant responses. This underscores the need for developing CAs that can maintain and use shared context during the interaction to improve interaction quality and task support.

The aforementioned limitations are related to LM-based CA, and the advancement of LLM offers the potential to effectively address and mitigate these issues. To unlock the potential of LLM, previous work explored that designing effective prompting [138] and facilitating information retrieval within conversational contexts [82] would provide natural user experiences. However, the question of how people adapt these benefits of LLM with CAs in real-life tasks remains an important yet unexplored topic. Our research aims to fill the gap in exploring user experiences using LLM-based CAs, focusing on cooking in a home kitchen setting, which we will describe the rationale for and previous work on in the next section.

2.2 Leveraging the Potential of LLM in Everyday Applications

Current language models require substantial amounts of data for training, facing challenges like fine-tuning a system to generate responses with varying tones [3]. However, innovative methods and algorithms, such as instructional-finetune [27, 131] and reinforcement learning with human feedback (RLHF) [25, 102], have revolutionized the potential of LLMs such as LLaMA [121], FLAN [27, 131], PALM [24], InstructGPT [102], and GPT-4 [101]. These models are fine-tuned on various natural language tasks, enabling them to effortlessly comprehend all instructions and

generate high-quality text content [17, 100, 110]. LLMs also show the ability to handle lengthy text input (e.g., GPT-4 [101] can take 32,000 tokens) to perform tasks that traditional LMs cannot handle, such as multi-turn conversations.

As LLM technology advances, researchers are actively exploring various potential applications [51, 63, 69, 77, 78, 81, 87, 119, 126, 135]. Researchers are particularly interested in utilizing LLM's ability to process inputs through prompt engineering and generate outputs that combine extensive dataset knowledge to make these applications come true [34], including qualitative analysis with cultural context comprehension [138], connecting LLM to robots for executing complex real-world tasks with task planning [2], co-creating tools for story and sketch generation [28], software engineering tools for code generation [64], and tools for mental health awareness [74, 140].

However, an underexplored area remains in utilizing LLMs for everyday home-based tasks, such as cooking. Our work aims to leverage the advantages of LLM technology and incorporate conversational assistance to bridge the gap between LLM capabilities and the lack of consideration for system design implications from a human-computer interaction perspective in everyday scenarios.

2.3 AI for Cooking

Cooking is a common daily task that requires the execution of sequential steps and multitasking skills to enhance efficiency [31, 98, 99]. Individuals new to cooking or attempting to prepare a new recipe often turn to resources like cookbooks and YouTube videos for guidance [72, 89, 130]. Their hands are usually occupied during the cooking process, restricting their capability to gather and process information. Various AI cooking assistants have emerged to address the challenge by using multiple modes of communication, including text, video, and audio, across various devices such as screens, tablets, and computers [22, 111]. For instance, AI-powered cooking assistants like "Cooking Nav" [46], "AskChef" [99], and "MimiCook" [111] provide multi-tasking planning, step-by-step guidance, and interactive ingredient weight projections. These tools help individuals optimize their cooking process and maintain their hands during the cooking process.

While there has been an increase in the number of AI-powered cooking assistants available, many remain limited to providing guidance strictly based on pre-set recipes and relying on multimodal inputs for context [22, 46, 98, 111]. Researchers explore using smart CAs for cooking assistants, but traditional language models and pre-determined heuristics may limit their flexibility, ability to answer questions, and multi-turn conversation capacity [134, 148]. Our study explores the potential of LLM-powered cooking CAs for a seamless, interactive cooking experience through voice commands, user experience, and design considerations for further development, allowing users to complete tasks at their own pace and receive immediate assistance.

3 METHODS

To attain insights into users' expectations and feedback during interactions with LLM-CA and to frame design suggestions that best utilize the unique strengths of LLMs in real-world scenarios, we conducted an exploratory user study utilizing a mixed-methods approach. This section presents an overview of our approach, including the implementation details of the system we developed and the user study specifics.

We will first explain the development and design of our LLM-based CA system, providing a detailed overview of the system pipeline and prompt design in Section 3.1. This will be followed by a discussion of the experiment design and procedure in Section 3.2. Next, we will describe the recruitment process and participant demographic information in Section 3.3 and 3.4. We will then outline the data collection methods in Section 3.5, including semi-structured interviews, surveys, and system logs. Finally, we will detail the analysis process in 3.6.

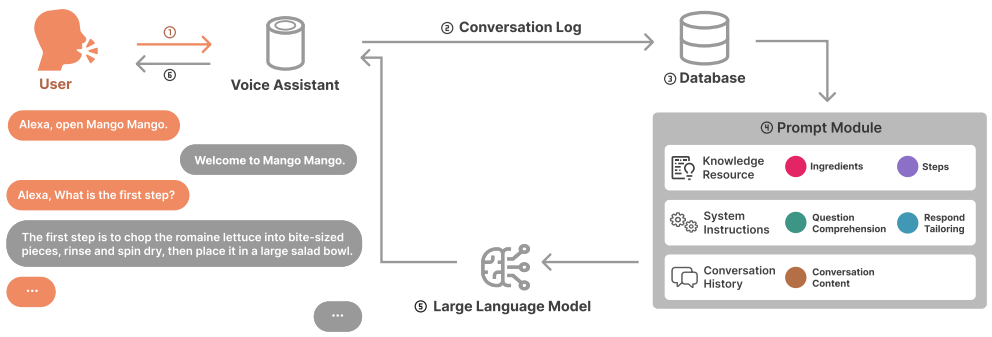

Fig. 1. Simplified system diagram of *Mango Mango*. The flow of the system is as follows: (1) Users speak to Alexa as voice input, then the text-to-speech process; (2) The transcribed inputs are saved in the conversational log (database); (3) The conversational log, along with the updated conversational history, was processed in the prompt module. The prompt module included knowledge resources, instructions, and conversation history; (4) The completed prompt is then sent to the GPT-3.5 Turbo; (5) The resulting response is sent back to Alexa; (6) Finally, Alexa converts the response into speech to the user to complete the system loop.

3.1 System Design

In this study, we introduced an LLM-based CA *Mango Mango* designed to assist users in completing a recipe. We selected Amazon Alexa as our voice-based conversational assistant (CA) and used the Alexa skill platform because of its flexible functionality and built-in features, particularly the text-to-speech conversion technology [84, 90]. Moreover, we have integrated it with the GPT-3.5-Turbo model, which has elevated its natural language processing capabilities. Figure 1 demonstrates the complete pipeline of our system.

3.1.1 Alexa Skill. The Alexa Skill Kit is a development framework for CA applications that can be integrated into Amazon smart speakers, such as Amazon Echo and Dot. This framework leverages Amazon's fundamental natural language and speech recognition technologies, such as Text-to-Speech (TTS), Speech-to-Text (STT), and intent recognition, to enable necessary speech recognition and text conversion functionalities for CAs, and allow users to customize the back-end application pipelines with a significant degree of freedom.

When a user activates the skill using a predefined invocation name, Amazon's STT technology transcription converts the user's spoken queries into text. The text is then sent to the backend of the skills, where it undergoes processing through our LLM system, as discussed in Section 3.1.2. Once the LLM has generated a response, it is sent back through the API and converted to synthetic voices using Amazon's TTS technology. The system awaits further user inputs after providing the response.

We will not delve into designing and implementing Alexa skills as it's not directly related to our research topics, but we plan to make the source code publicly available upon acceptance.

3.1.2 LLM Selection. Our LLM-CA system utilizes OpenAI's GPT-3.5-Turbo LLM in the backend. The selection of an LLM was guided by an evaluation of several key factors. Firstly, GPT-3.5-Turbo has demonstrated remarkable proficiency regarding both natural language understanding and generation, making it the backbone of the prevalent web-based chat assistant, ChatGPT. Furthermore, its capability to manage extensive input content enables us to send numerous previous rounds of conversation histories simultaneously, resulting in more coherent and suitable multi-round conversations.

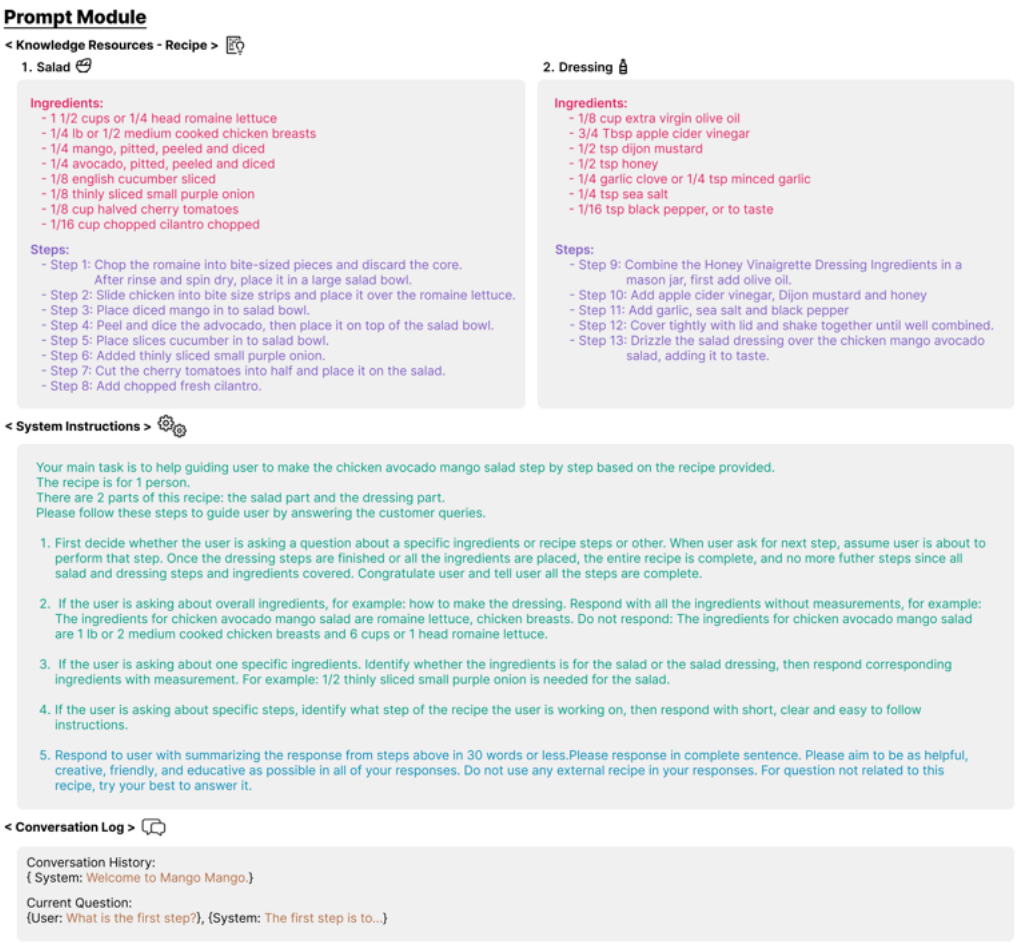

Fig. 2. Detail components of *Mango Mango* 's prompting module. The prompting module contains the instructions module (left) and the knowledge resources (right). The instructions module will understand the users' input from the conversation, then the model selects the appropriate knowledge resource based on the user's input. The Knowledge Resources cover all the necessary information related to the recipe, including ingredients and steps. Finally, return the tailored guidance or suggestions based on users' inquiries.

Secondly, GPT-3.5-Turbo provides comprehensive and stable API support, which is crucial to supporting the smoothness of our lab experiments. During the implementation of our system, we endeavored to utilize the GPT-4, a more advanced LLM, which boasts superior capabilities to its predecessor, GPT-3.5-Turbo. Regrettably, despite its acclaimed superiority, we observed a suboptimal response time from GPT-4 API, making it more prone to exceeding the Alexa Skill's backend waiting time limit. This caused the Alexa Skill to be forcibly terminated before the response from GPT-4 was generated and sent back. In summary, GPT-3.5-Turbo is an ideal LLM benchmark to provide stable support while providing the unique advantages of LLMs over traditional language models for our exploratory study.

After the user's voice input is captured and correctly recognized by the Alexa Skill, it is converted into text and sent to a database for conversation log storage. The conversation log is then forwarded

to the back-end prompting module, where the input text is organized and reconstructed into a complete query. This query is sent to the LLM, GPT-3.5-Turbo in our implementation, via API to generate a response.

3.1.3 Database. To effectively maintain and manage the conversation log, we incorporated a database into our system design, a method similarly utilized in previous studies [92, 143]. In our implementation, we integrated a shared Google Sheet as a middleware database. Each interaction is logged in real-time and stored in the Google Sheet, ensuring accessibility, transparency, and efficient tracking of conversational data. This logged data is dynamically forwarded to the prompt module, where it becomes part of the prompt input. By including conversation history in the prompt, the system leverages prior conversations to enhance contextual understanding to generate more relevant and personalized responses.

3.1.4 Prompting Module. Our prompting module is specifically designed to support the cooking scenario with recipes. In addition to the conversation log, it is structured around two additional core components: **Knowledge Resources** and **Instructions**. Figure 2 illustrates the complete prompt, developed based on the salad recipe used in our lab experiment. To ensure the system's functionality, we conducted three pilot studies within the research team to iteratively test and refine the system. These studies helped identify key areas for improvement and informed the final design of the module.

Knowledge Resources. In the cooking scenario used in our experiment, the Knowledge Resources were structured based on a design choice to ensure clarity and scalability. These resources consisted of two primary components: ingredients and steps, encompassing all necessary information related to the recipe. During our pilot study, we observed that GPT struggled to distinguish whether an ingredient and steps were intended for the salad or the dressing. Therefore, subcategories were created within each component to clarify distinct elements of more complex recipes. For example, a subsection for "salad" and "dressing" were added under the categories in our experiment, as its preparation was relatively independent of the main salad preparation process. This structure reflects a design decision aimed at maintaining flexibility and usability. By organizing the recipe in additional subcategories, the framework supports effortless scaling to accommodate various recipes, regardless of complexity, while ensuring that the information remains logically structured for users.

Regarding the data resources used to create the cooking steps, existing work [130] has discovered that people tend to search for recipes on the internet in real-life scenarios, especially YouTube recipe teaching videos. Therefore, we chose YouTube cooking teaching videos as recipe source data in our system. Specifically, we first transcribed the video to capture the ingredient list and each cooking step, ensuring the instructions' originality and identicality to the instructions from the video. We leveraged bullet points to list each individual ingredient information, such as the name and quantity of the ingredients required for the recipe, as well as individual step details, so that the LLM can more conveniently and accurately locate the sequence of instructions and the details of each item. Additionally, the distinct component in this recipe was the dressing of salad, so we separated the list of ingredients for the dressing into a subsection of special ingredients. The process of transferring YouTube videos into the Knowledge Resources that could be used in our system is highly replicable and scalable to different recipes. For our experiment, we selected the chicken avocado mango salad, and we will delve into the recipe specifics in Section 3.2.

Cooking tasks vary widely due to the unique requirements of different recipes. The structures of Knowledge Resources were designed to be manually adaptable, catering to various recipes like sandwiches, cocktails, or no-bake desserts, in line with the procedures outlined in the earlier

section. While the content of ingredients and cooking steps differ, the method of inputting relevant information into these sections in the Knowledge Resources remains consistent. However, we still recognize that some unique recipes and scenarios might pose challenges in scaling up *Mango Mango*, which will be discussed in the limitations section.

Instructions. LLMs possess an exceptional natural language generation ability and access an almost boundless wealth of knowledge, enabling them to answer a wide range of questions. However, LLMs also pose the difficulty of limiting the content they produce, which can lead to information overload, as highlighted in previous studies [29, 58, 70]. To optimize the LLM's natural language capabilities for cooking-related inquiries, we designed a detailed instruction pipeline in the prompting module, allowing for proper information retrieval from the Knowledge Resource to generate accurate responses based on requests. This comprises question comprehension and two aspects of response customization, namely recognition and targeted adaptation for different question types, as well as guidance on generating content more akin to human conversation.

We understand that cooking questions from users need different levels of detail and response methods. When users ask about necessary ingredients, it can be challenging to provide every detail, such as names, quantities, and specifics. Instead of overwhelming them with too much information, providing a list of ingredients is more effective. If users want details about a particular ingredient or step, they should ask additional questions. The model should provide specific responses based on the knowledge available in the knowledge resources module. The model should do more than just provide recipes. It should also respond to non-recipe-related inquiries. For example, users might ask practical questions about the cooking process, such as how to use kitchen tools or convert measurement units.

Therefore, as shown in the left module in Figure 2, we first require the model to understand user's input. Based on different users' inquiries, we provide targeted response guidance and suggestions for the model. When the user inquires about the required ingredients, we instruct the model to provide only a list of ingredients without specifying the quantity. We also provide a response template for such queries. If the user wants to know specific details about an ingredient, such as quantity, weight, or measurement conversion, the model needs to identify whether the user is referring to dishes or seasonings. Based on the user's input, the model selects the appropriate knowledge resource to provide an accurate response. This design can correctly identify and respond to user inquiries about specific ingredients in dishes that share common ingredients with condiments. In regard to recipes, it is imperative that users receive descriptive and succinct guidance when inquiring about a specific step.

In addition to our tailored guidance for different question types, we've compiled some general tips to enhance the naturalness of the AI-generated responses. Our analysis revealed that the model often produces verbose and redundant content, which can overwhelm the recipient and disrupt the exchange's coherence. Furthermore, it is important to note that in certain instances, despite the fact that the LLM's response is comprehensive, Alexa Skill may truncate extended responses when speaking back to the user, leaving them incomplete mid-sentence. It is imperative to ensure that responses remain concise in order to avoid this issue.

As a result, we asked the model to prioritize brevity, aiming for responses that are no longer than 30 words, whenever feasible. We require the model to limit its scope to answering only recipe-related questions from the given knowledge resource. However, when the user's questions exceed the boundaries of the recipe itself, we expect the model to leverage its world knowledge to provide comprehensive guidance.

(a) (b)

Fig. 3. Our study took place at the smart home laboratory (a). It was designed with a one-bedroom apartment floor plan with a fully functional kitchen and monitoring cameras. (b) Picture of a participant working on the experiment in the kitchen. Alexa is marked with a red circle on the left side of the table.

3.2 Experiment Design and Procedure

To gain insights into users' experiences while interacting with *Mango Mango* during cooking, we organized an in-lab user study to simulate real-world scenarios and collect valuable feedback. Our study took place in a smart home laboratory, designed with a one-bedroom apartment floor plan that included a fully functional kitchen and equipped with monitoring cameras, as illustrated in Figure 3a. Participants should have completed demographic questionnaires and relevant surveys as part of the initial screening process. Upon arrival, researchers provide participants with an informative sheet detailing the data collection method and data storage and the participant protocol for the study. The researcher then asked for verbal consent from participants regarding the recording of their participation during the experiment. Subsequently, participants received a tutorial session guided by the research team. This session included a brief tour of the kitchen space and a trial interaction with *Mango Mango* to familiarize them with the Alexa voice assistant. Following this, participants viewed instructional YouTube videos that demonstrated how to prepare a chicken mango avocado salad. They were not required to memorize the video content but were encouraged to become acquainted with the recipe. After viewing the video once, participants no longer had access to it, relying solely on our system, *Mango Mango*, for assistance when needed. Figure 3 shows the tabletop setup for the experiment. They then proceeded to prepare the salad while freely interacting with *Mango Mango*, without intervention from the researchers. Throughout this process, researchers observed the interactions from the control room and collected video recordings. Upon completing the dish, participants engaged in semi-structured interviews and surveys to reflect on their experiences. Our experiment was approved by the university's Institutional Review Board. To minimize any potential risks during the cooking process, we intentionally excluded using ovens, sharp knives, stoves, or any other appliances and tools that could threaten the participants' safety.

3.2.1 Rationale. We opted to utilize YouTube videos as our primary data resource for the following reasons. YouTube videos are immensely popular due to their detailed descriptions and rich visual cues. However, they lack voice interaction and sometimes require manual touch and scrolling for video control. On the other hand, voice assistants support hands-free interaction but may lack detailed information. Recognizing this disparity, we divided the use of these two tools into two phases: watching a video **before** cooking and interacting with the cooking assistant **during** the cooking process. Consequently, as previously described, we developed a workflow to translate

ID	Gender	Age	Cooking Frequency	Recipe Searching Frequency	CA Frequency	CA Recipe Search Frequency	CA Usage
1	Female	18-24	≥1/week	≥1/week	≥1/month	Rarely	Check weather, Look up information, Small talk, Set alarm
2	Female	25-34	≥1/week	Rarely	Rarely	Never	Check weather, Home device control, Music, Set alarm, Check time
3	Male	25-34	Daily	≥1/week	≥1/week	Rarely	Check weather, Home device control, Set alarm
4	Male	18-24	Daily	≥1/week	≥1/month	Rarely	Check weather, Music
5	Male	25-34	Daily	≥1/week	Daily	Rarely	Check weather, Music, Look up information, Set alarm, Check time, Reminder
6	Female	25-34	Daily	≥1/week	≥1/week	Rarely	Check weather, Music, Look up information
7	Male	18-24	Daily	Daily	Daily	Rarely	Check weather, Music, Look up information, Set alarm, Reminder
8	Female	25-34	≥1/week	≥1/week	≥1/week	Never	Check weather, Music, Set alarm, Check time
9	Female	25-34	Daily	Rarely	Rarely	Rarely	Music, Look up information
10	Male	25-34	>1x/wk	≥1/week	Rarely	Rarely	Music, Set alarm, Reminder
11	Male	18-24	Daily	≥1/month	Daily	Never	Check weather, Music, Reminder
12	Male	25-34	Daily	≥1/week	Daily	Never	Home device control, Music, Set alarm

Table 1. Overview of participant demographics and CA usage for our 12 participants. This table summarizes user habits related to cooking, recipe searches, and CA usage. Frequencies are categorized as Daily, ≥1/week (at least once per week), ≥1/month (at least once per month), Rarely, and Never.

video content into prompts. The full prompts can be found in Tables 5 and 6 in the appendix. In the user study, following this workflow, we initially presented participants with a YouTube video, followed by their interaction with *Mango Mango* for real-time in-situ assistance during cooking.

We chose the recipe for a chicken mango avocado salad for our study due to its relatively short preparation time, with all the steps typically completed within 30 minutes. However, this recipe presents a cognitive challenge for users because of its numerous ingredient measurements, often necessitating external assistance [72, 89, 130]. Furthermore, to address safety concerns, the recipe does not require the use of an oven, stove, or sharp knife (instead, a table knife is used), ensuring the ethical compliance of our study.

3.3 Recruitment Process

Participants for this study were recruited via social media platforms and email. Recruitment posters were shared along with a comprehensive description of our research objectives, a direct link and QR code leading to the screening questionnaire. The screening questionnaire was used for participant selection and included questions related to demographic information, allergy history, prior usage of CAs, and participants' cooking experiences.

A total of 12 participants were successfully enrolled in our study, each meeting the following eligibility criteria: being 18 years of age or older, fluent in English, possessing prior cooking experience, comfortable with audiovisual recording during the experiment, and having no known food allergies to the ingredients used in the study. The experimental session's duration was less than one hour. Each participant was compensated with a $30 Amazon e-gift card for acknowledging and contributing their time to participate in our study.

3.4 Participant Demographic Information

We recruited a total of 12 participants. The sample consisted of 58.3% male and 41.7% female participants, majority aged between 25 and 34. Most participants reported cooking daily and searching for recipes at least once per week. CA usage patterns varied, with 33.3% using it daily and others less frequently. Common CA activities included listening to music (83.3%), checking the weather (75%), and setting alarms (66.7%), while small talk was rare (8.3%). Recipe searches were primarily conducted on YouTube and recipe-sharing websites, with limited use of CA for cooking assistance. Detailed demographic information is available in Table 1.

3.5 Data Collection

3.5.1 Semi-Structured Interview. We designed and conducted a semi-structured interview after participants finished with their post-study questionnaire. The interview covered simple questions on participants' experience using existing CAs and our system *Mango Mango*, cooking habits, and how they envision using *Mango Mango* in the future. Each interview lasted between 15 - 40 minutes. These semi-structured interviews provided information for researchers to understand participants' overall experiences with CAs, particularly in the cooking task.

3.5.2 Survey Measures. In this study, we utilize different methods to collect results from interaction, performance, subjective workload, and participants' feedback to explore our RQ1. A pre-study questionnaire was provided to collect participants' context on basic demographics, cooking background, and usage of voice assistants. Participants were also requested to complete a five-question survey that we designed to assess their perceptions of the CAs' capabilities.

The post-study questionnaires consisted of four elements: the Voice Usability Scale (VUS), a 12-question scale that assesses the usability of the voice interface [150]; the Explainable AI survey (XAI), a six-question scale that evaluates the trustworthiness of explainable AI systems' output from users [53]; the NASA-Task Load Index (NASA-TLX), a six-question scale used to measure participants' subjective workload in six dimensions [49]. Participants were also asked to complete the same survey provided before the study to evaluate any potential changes in their perceptions of the current capabilities of the voice assistant. The Results section will provide a detailed analysis of the survey results.

3.6 Data Analysis Process

The collected survey responses were based on a 5-point Likert scale. Mean and standard deviation were calculated for each question to summarize user responses, providing a general sense of performance. Higher scores indicated better performance or more positive user experiences. The quantitative data served as a supplement to the qualitative findings, offering context to the results derived from thematic analysis.

A total of 4 hours and 35 minutes of interview audio, along with 3 hours and 39 minutes of audiovisual recordings from the experiment were collected. These recordings were transcribed using an automated service for further analysis. The co-authors independently conducted open coding for the first two participants, employing thematic analysis to identify initial themes related to the research questions.

After initial coding, the co-authors collaboratively reviewed, discussed, and categorized the codes to establish a preliminary codebook. This codebook was refined through iterative discussions and reviews of emerging codes, ensuring consistency and agreement across interpretations. The final codebook was collaboratively refined and finalized after achieving agreement among the co-authors. One author then applied the finalized codebook to the remaining transcripts following Grounded Theory [18, 41].

4 RESULTS

In addressing RQ1, this section presents quantitative results to provide an objective understanding of user experience. We then discuss the themes of users' experience, including successful and challenging experiences with LLM-based CAs in cooking tasks.

4.1 Quantitative Results

To better understand user experiences with our system, *Mango Mango*, we evaluated system accuracy through conversations (Section 4.1.1), analyzed users' task performance (Section 4.1.2),

and collected quantitative feedback through questionnaires (Section 4.1.3). Although the sample size was small, the results provide a snapshot of user impressions, offering measurable insights to complement our qualitative observations.

4.1.1 System Accuracy Evaluation. Evaluating the accuracy of *Mango Mango* in interpreting and responding to instructions is crucial. In our analysis, a response was considered valid when it contained accurate content verifiable by the recipe and was expressed fluently.Among 447 queries, 66.4% of queries received valid and accurate responses. 23.0% were invalid due to Alexa system errors and speech-to-text inaccuracies, while 10.6% were invalid due to LLM errors of incorrect sequencing or unrelated answers. Participants posed a total of 28 queries beyond the recipe's scope during the study. 75.0% of those queries received valid responses. Some invalid responses were due to the VA's inability to access specific information, such as the task's remaining time, and difficulty verifying the accuracy, such as the calorie count of the dish not being mentioned in the YouTube video or recipe.

Our accuracy calculation represents the maximum potential inaccuracy of our system, as any response that could not be verified against the original video or recipe was classified as invalid. Despite this, we observed that users often rephrased or repeated their questions to obtain valid answers, suggesting that initial invalid responses did not prevent them from continuing their tasks. Additionally, we acknowledge the inherent limitations of LLMs, such as the propensity to generate hallucinated responses, which will be discussed further in the limitations section.

4.1.2 User Task Performance Result. To explore *Mango Mango*'s efficacy, we analyzed key metrics such as task completion rates and performance efficiency among participants. All participants completed the assigned task within the 30-minute limit. The completion times ranged from approximately 13 minutes 19 seconds to 26 minutes 10 seconds (mean = 18 minutes 26 seconds). The average number of queries per participant was 37.7 queries, ranging from 19 to 75 questions.

Out of the 12 participants, all participants successfully prepared the dish accurately with all correct ingredients, relying on *Mango Mango* without direct access to the recipe or video while cooking. Notably, three participants precisely followed the procedures outlined in the video and instructions. As for the remaining nine participants, they introduced some sequencing errors. However, these mistakes did not impact the final dish's outcome. These deviations included multitasking and improvisation, like rearranging the order of adding garlic, sea salt, and black pepper after receiving the complete instructions (P1,2,4,6,8,10,12). Overall, every participant successfully completed the dish. In the following section, we will delve into their experiences interacting with the system by analyzing the survey results.

4.1.3 Survey Result. We employed four scales in this study: the Voice Usability Scale (VUS) to evaluate the usability, affectiveness, and recognizability & visibility of the voice interface [150], the XAI survey to assess the trustworthiness of system outputs [53], and the NASA-TLX to measure task workload, including mental, physical, and temporal demands, effort, frustration, and performance [50]. These surveys were adapted to fit *Mango Mango* by selecting relevant questions while omitting those unrelated to the experimental task. The complete set of questions and detailed results are provided in Appendix A.2, and responses were collected using a 5-point Likert scale, where higher scores indicated better performance.

Overall, participants perceived *Mango Mango* positively in cooking scenarios. The VUS results reflected overall satisfaction with the system's user experience, evaluating usability, affectiveness, and recognizability & visibility. The usability questions assessed the system's difficulty, resulting in a score of 2.03 on a 5-point Likert scale. For affectiveness, participants rated the system with a mean

score of 4.25 on a 5-point Likert scale. Finally, the mean score for recognizability and visibility was 3.25 on a 5-point Likert scale.

The XAI survey measured trustworthiness across four dimensions: predictability, reliability, efficiency, and believability [150]. Results for all questions exceeded a mean score of 3.9, with an overall trustworthiness score of 4.11 on a 5-point Likert scale. NASA-TLX results showed low levels of mental, physical, and temporal demand, as well as low frustration, alongside high-performance ratings. These quantitative findings align with qualitative interview feedback, where most participants described successful interactions with minor obstacles.

Beyond these scales, we explored whether participants' perceptions of CAs changed after interacting with *Mango Mango*. We included five exploratory questions asked both before and after the study. These questions focused on conversational fluency, memory, follow-up questions, integration into daily activities, and active collaboration. Results across all five aspects improved after participants used *Mango Mango*.

4.2 Themes of User's Experience with LLM-Based CAs

To further understand users' perceptions through their interaction experiences, we categorized users' experiences into several themes through thematic analysis. In the following sections, we will present these themes under two high-level categories: successful and unsatisfactory, based on the interaction between users and *Mango Mango*. Successful experiences were those where users effectively leveraged the LLM's capabilities, receiving clear, actionable instructions that enhanced their cooking tasks and met or exceeded their expectations. In contrast, unsatisfactory experiences occurred when users encountered difficulties in utilizing the capabilities of LLM during their cooking tasks.

4.2.1 ***Successful*** *Experience When Using LLM-Based CAs for Cooking Tasks.*

From the survey results, we have confirmed that participants had an overall successful experience using *Mango Mango*. In this section, we will elaborate on specific aspects of participants' usage and experiences they were satisfied with, particularly those related to the LLM powered capabilities.

Firstly, many participants asked *Mango Mango* for **information that extended beyond the scope of the recipe and received satisfactory answers**. These inquiries often revolved around fundamental cooking tips, which might be unrelated to the specific recipe and were not included in the original instructions. These were particularly helpful, especially for novice cooks lacking essential cooking knowledge. For instance, P5 inquired, "How do you peel an avocado?" Such information not only aided in the immediate task but also contributed to participants' overall cooking skills. Another category of information sought by users pertained to the recipe but was not explicitly provided in the original instructions, such as nutrition information. For example, P8 asked, "How many calories are in the salad?" *Mango Mango* responded with an estimate: "This might answer your question, 224 calories," despite the absence of this specific data in the original recipe. Despite the lack of explicit information, ChatGPT is capable of estimating the results and providing a suggestion. Importantly, participants posed these questions naturally, demonstrating their recognition of the system's ability to address such inquiries.

Another common type of question participants frequently asked was **next-step instructions**, such as "What's the current step?" This pattern of inquiry suggests that users had recognized the system's capability to remember the ongoing status and the history of the conversation. P6 pointed out, "It follows up on your previous question… It sticks to the track, so it's like one continuous flow." Similarly, P1 was impressed by the system's ability to stay on track, stating, "It (*Mango Mango*) can memorize which step you are in right now. And you can continue to the next one instead of starting over from the very beginning." When interacting with the system, users quickly accepted

the fact that *Mango Mango* could retain this information, indicating a high level of confidence in its retrieval capabilities as a 'machine'.

Mango Mango also excelled in **tailoring solutions to meet users' specific requirements**, and our participants quickly took advantage of this feature to receive instructions based on their own settings. For instance, during the experiment, some participants encountered challenges when a specific tool demonstrated in the video or recipe was unavailable. In these situations, our system provided valuable assistance, even when these occurrences were not explicitly outlined in the original instructions given to our system. P1 noticed a missing tool and asked for help from *Mango Mango*, stating, "Alexa, I want to make the lettuce dry without a spinner, but I don't have one. How can I do it?" *Mango Mango* offered tailored, step-by-step guidance on completing the task without the missing tool. Moreover, *Mango Mango*'s responses could be further personalized based on the specific setting and individual user preferences. In a different instance during the experiment, P10 asked, "Alexa, give me all the vegetables and leafy greens that I need to chop," Instead of following the procedure described in the original recipe, *Mango Mango* responded with customized instructions: "You will need to chop one-quarter head of romaine lettuce, English cucumber, and thinly sliced purple onions for the salad." This shows that our participants had both needs and confidence in *Mango Mango*'s ability to reorganize existing information to tailor it to users' needs.

Due to its extensive capability to customize instructions according to users' requests, participants also realized its ability to assist them in **planning cooking tasks and dynamically controlling the workflow**. Participants could adjust the order of tasks based on real-time situations or even plan for multitasking with the assistance of *Mango Mango*. For example, P10 preferred to inquire about tasks a few steps in advance, stating, "I always used to ask it a few steps before. So when I'm cutting the onion, I would ask what I need to do with a tomato." P10 also highlighted the advantages

Theme	Sub-theme	Example
Receive Extensive Information Beyond the Recipe by the LLM	Fundamental Cooking Tips	"Alexa, how to peel avocado?"(P5) "Alexa, tell me that amount of teaspoon if I want one quarter tablespoon"(P1)
	Nutrition Information Related to The Dish	"I asked how many calories are in the in the salad" (P8)
Contextual Memory & Task Awareness	Current Step	"There was one question I asked which step am I at right now? And he told me on step four." (P5)
Adaptive Contextual Personalization	Lack of Tools	"Alexa, I want to I want to make the lettuce dry without some water but I don't have a spinner so how can I do it?" (P1)
Plan Tasks & Control Flow Dynamically	Support Multi-Tasking/ Task Planning	"I always used to ask it a few steps before. So when I'm cutting the onion, I would ask what I need to do with a tomato. " (P10) "When I focus on something I just asked, you know, *Mango Mango* whats the next step and then I was cutting stuff and it says the instruction."(P4)
	Change the Order of Tasks	"And basically, I could execute things in my order as well. I did not have to follow the same path, I could figure out my own path." (P10)
Culinary Learning through Recipes	New Recipes	"I would say it works well on beginners and people who have like a good experience with cooking but who are also new to certain recipes."(P10)
Conversational Engagement and Encouragement	Congratulation Messages	Q: Do you think this Alexa talks differently? P2: Expressions like enjoy your food.

Table 2. Qualitative code book and description of participants' **successful experience** and usage with the LLM-based CA.

of this approach with *Mango Mango*'s assistance, noting, "I wouldn't be standing there waiting for it to give me an answer. I would always be doing something… You get to ask a question one step ahead at a time, and that helps." Likewise, P4 adopted a similar strategy for task planning, saying, "When I focus on something, I just asked *Mango Mango* what the next step was, and then I was cutting stuff, and it gave me the said instruction." In summary, *Mango Mango*'s ability to promptly react to in-situ flow changes enables more efficient and dynamic flow control for users, especially those with advanced skills in task planning within cooking scenarios. However, it is important to note that instead of providing goals and letting *Mango Mango* plan the order of tasks, our participants tended to only ask for information and still did the planning themselves. This suggests a potential preference for a usage mode in cooking, which often requires complex task planning and extensive user controls.

In addition, participants praised the system's ability to help them learn a new recipe, which was the case in the experiment where all the participants made this specific salad for the first time. Acknowledging this learning potential suggests that our participants may view *Mango Mango* as a mentor-like system with extensive knowledge of the recipe and general cooking, capable of teaching them new things they were unaware of.

Lastly, an interesting response came from P2 when we asked, "Do you think Alexa talks differently?" They answered, "Expressions like 'enjoy your food'." While this information may not be necessary for completing the task, it mimics human-like conversation and fosters a sense of conversational engagement and encouragement, making the participant feel like they are interacting with a special CA. This experience shows that having such human-like can positively influence a participant's perception of the system, enhancing the overall user experience.

In summary, we explored participants' successful experiences and interactions with *Mango Mango*. In the following section, we will delve into some of the unsatisfactory experiences.

4.2.2 ***Unsatisfactory*** *Experience When Using LLM-Based CAs for Cooking Tasks.*

Although our participants benefited greatly from the assistance provided by *Mango Mango*, there were still many challenges during the interaction, many of which related to the disparities of perception in the LLM powered capability.

There are instances of dissatisfaction from users due to **information overload**. In our experiment, the recipe encompassed instructions for preparing the salad and crafting the dressing. Although all the necessary ingredients were provided, participants were tasked with measuring precise amounts for the dressing. The dressing was introduced all at once in the instructional video, and we followed a similar approach in our written recipe prompt. However, many participants expressed difficulties following *Mango Mango* 's instructions, primarily due to the presentation of multiple ingredients at once. For instance, P5 articulated this issue, stating, "The first was that it was giving too much information. For example, he's telling me salt and pepper together, where I have to measure one and then measure the other one. But when I measured the first one, I forgot about the other one" Additionally, P8 also highlighted the narrative speed was too fast, "When *Mango Mango* delivers the instructions, it tends to speak too rapidly, necessitating repeated requests for clarification" This indicates the system's inability to comprehend and deliver the appropriate amount of information, which, however, is a fundamental requirement for users to ensure fluent and informative conversation.

Furthermore, as users became more accustomed to natural conversations with *Mango Mango*, some **system constraints** became more evident, such as misunderstandings of oral expressions, the need to initiate conversations using the wake word, and increased cognitive load. For example, P7 encountered a linguistic error during the experiment and noted, "One area where I found a mistake was that I asked what's the 'last' instruction, meaning the 'previous' one, it took me to the 'very

last' instruction." This occurred due to the ambiguity of certain words, which can have multiple meanings, especially in oral versus written contexts. Additionally, when users are able to adopt the system's assistance for complex tasks like multitasking, it introduces a significant cognitive load compared to simply listening to instructions. In the case of cooking, this increased cognitive load could potentially hinder task completion and even lead to safety issues. In summary, we observed that as our conversations became more natural due to the extensive capabilities provided by the LLM, the system needed to adapt accordingly. It had to recognize that the dialogue had become more oral, making it more challenging for users to consistently use the wake word. Additionally, the increased cognitive load needed to be addressed.

Recognizing that the system is imperfect and sometimes does not behave as expected, some participants expressed a desire for our system to incorporate more user feedback for further verification before making decisions. P8 highlighted a specific suggestion: "I would like it (*Mango Mango*) to repeat my questions so that I can confirm that I'm providing the right instruction. It's much better than asking something, and it (*Mango Mango*) misunderstands me and gives the wrong answer." To address this issue more effectively, P8 expressed a preference for the system to "show me more itself or ask me for confirmation in my case" to minimize the occurrence of misunderstandings. We realized that although our system supports further iteration through follow-up questions, it was primarily designed as a question-solving system that often aims to provide an immediate answer rather than engaging in cooperative decision-making with users. As an LLM-based assistant, users expect it to be more communicative and involve them more in decision-making.

Similarly, as a question-initiated system, *Mango Mango* primarily provides **passive responses**. However, P10, for instance, expressed a desire for the system to "check on me before proceeding." We realized that this indicates an increased level of expectation from users. "Checking on users"

Theme	Sub-theme	Example
Information Overload	Too Many Ingredients at Once	"The first was that he was giving too much information. Like for example, he's telling me salt and pepper together where I have to measure one, measure the other one where we measure the first one. And I forgot about that." (P5)
	Speak Too Fast	"When *Mango Mango* actually provide me with the steps it's kind of speak too fast and I have to kept asking *Mango Mango* to repeat the instructions"(P8)
Misunderstanding of Oral Expressions	Linguistic Error for Oral Expressions	"One area where I found a mistake was that I asked what's the 'last' instruction, meaning the 'previous' one, it took me to the 'very last' instruction."(P7)
Increased Cognitive Load	Distracting Back and Forth Conversation	"Even though it's doing a very good job with interaction, sometimes you still need to try talk to it slowly so you can understand the answers. That requires like back and forth conversation. But while you were doing that, and if something is on the stove, that could be very distracting."(P4)
Expect More Dialogue with the System	Lack of Verification From Users	"I would like it to repeat my questions. So that I understand that I'm giving the right instruction. It's much better than I asked something, and it (*Mango Mango*) misunderstand me and gave the wrong answer."(P8)
		"I would like to show me more itself or asked me for confirmation in my case." (P8)
Only Passive Response	Lack of Auto-Tracking	"I would like a mode in which, for example, while making the dressing, instead of telling me all the ingredients one after the other and may not be able to catch up. Maybe if I asked her (*Mango Mango*) to like check on me before proceeding. That would be like an amazing step, amazing feature to have." (P10)
Feature Discovery Challenges	Lack of User Guidance on System's Features	"I don't know whether Alexa can help me to control the time or it can just tell me. I don't know like whether it can intelligently tell me when I should maybe do something and do the other things. I don't know whether he (*Mango Mango*) can do that." (P9)

Table 3. Qualitative code book and description of participants' **unsatisfactory experience** and usage with the LLM-based CA.

suggests a transition from the system being a passive assistant that waits for questions to an 'agent' that actively participates in the process and provides assistance. Note that P10 also mentioned "checking on me" rather than "telling me what to do," which aligns with the earlier statement about users preferring more dialog-like suggestions rather than direct instructions.

Finally, P9 raised a notable concern regarding **the absence of clear guidance on the available features** when using *Mango Mango*, which is unsurprising. Despite the advantages offered by *Mango Mango*, participants were constrained by a 30-minute time limit for task completion, coupled with a brief 5-minute tutorial provided by the research team before initiating the assignment. P9 articulated this issue by saying, "I don't know whether Alexa (*Mango Mango*) can help me control the time or intelligently tell me when I should maybe do something and do the other things. I don't know whether it (*Mango Mango*) can do that." Considering this, presenting the full range of *Mango Mango*'s capabilities could potentially empower users to use it more effectively and extract maximum benefits, especially in real-world contexts. However, how to design such a tutorial remains an issue that needs further discussion, which we will also explore in a later section.

In summary, we explored participants' unsatisfactory experience and interactions with *Mango Mango*. In the following section, we will discuss how this might influence future design.

5 DISCUSSION

In this study, we explored users' successful and challenging experiences while interacting with *Mango Mango* in a cooking scenario, specifically focusing on the different interactions facilitated by the LLM. Based on the insights from the user studies, we address RQ2: What are the design implications of LLM-based CAs aimed at assisting users in real-world practices like cooking? (Section 5.1). Additionally, we expand on RQ1 by discussing users' shift from traditional recipes to LLM-CAs as instructional tools for cooking (Section 5.2, 5.3). Next, we discuss other potentials and challenges of LLM-based CAs in real-world practices inspired by our study results, presenting directions for future works. (Section 5.4, 5.5, 5.6). Finally, we recognize the study's limitations (Section 5.7).

5.1 Design Considerations: Redesign LLM-CAs for Effective Collaboration

In this section, we will delve into how our findings guide the design of CAs capable of leveraging the full potential of LLM to assist users in accomplishing real-time tasks. We proposed five design implications for a future LLM-CA, offering actionable solutions and providing examples in the context of cooking.

5.1.1 Contextualize an Universal LLM-CA for Specific Tasks.

In contrast to typical CA applications that rely on techniques like intent recognition in NLP [46, 99, 111], leveraging LLM provides an easier approach to handling a wide array of inquiries beyond rigid rule-based frameworks through prompts engineering [57, 78, 126, 132]. In the past, various dialogue systems attempted to encompass as many user queries as feasible within their training resources due to limitations in understanding questions beyond those resources [52, 95, 125]. However, with the robust capability to comprehend 'common sense knowledge' [133], the challenge for an LLM-powered CA shifts to contextualizing a diverse range of inquiries. Past CAs are usually compartmentalized for specific applications. In contrast, LLM-CA often possesses the ability to manage multiple applications through a single agent, making context identification even more challenging. Although prompt engineering might offer partial solutions for this issue [88, 105, 127], the distinct challenge of understanding these questions in real-time voice-based tasks such as cooking persists due to the inherently oral nature of these inquiries. In our experiment, we noticed that users' queries were always oral and vague, often lacking direct references to the specific recipe

or the type of tasks. As a result, apart from describing the context in prompts [86], to ensure smooth context transitions and maintain relevance, we propose the following solution:

- Extract context from conversation history to help understand users' queries. In cooking scenarios, this can involve identifying the user's current cooking step based on the past conversation and offering relevant instructions accordingly.
- Proactively ask for user confirmation. The system could proactively verify the user's current stage during the cooking process. This check-in ensures the system's responses align with the user's actual progress and needs.

5.1.2 Augment Current LLM Knowledge Base.

LLM systems have the capability and should actively gather information to enrich their knowledge base, specifically related to the target tasks [33, 39]. In our experiment, by integrating relevant information such as ingredients, steps and instructions as an external knowledge base, *Mango Mango* performed well in responding to queries about making the salad. While the current system performs sufficiently, it could further enhance its effectiveness by integrating a border range of task-specific information. During the experiment, participants raised questions that extended beyond the recipes, such as asking about the salad's calories. Broadening the system's specialized knowledge base is recommended to address such inquiries effectively, enabling users to receive accurate responses to their related questions. Consequently, we propose:

- Incorporate special training materials. For cooking scenario, this could involve integrating nutritional information and fundamental cooking techniques into its knowledge base.
- Fine-tune the model according to the context of specific tasks.

5.1.3 Elicit Feature Discovery Instead of Focusing on the Expectation Alignment.

The advancements in knowledge of LLM-CA as described in the earlier section also necessitate a shift in how systems should engage with users to convey their usability. Prior studies have highlighted that voice assistants (without LLMs) often fall short of providing comprehensive knowledge, proposing the design implications for systems to focus on communicating their capabilities and limitations [52, 58]. However, this design consideration needs adaptation, as now the knowledge base of LLMs might occasionally exceed users' expectations instead of falling short. In such cases, instead of limiting users in their interactions with the voice assistant, the system should encourage diverse and creative inquiries. Therefore, LLM-CAs should proactively reveal hidden features that support dynamic, creative, and context-specific user interactions and encouraging users to explore these hidden features could increase engagement and satisfaction with the system.

A mechanism for supporting this shift from expectation alignment to feature discovery is prompt design, which defines how the LLM-CA responds to user queries and encourages user-driven exploration. While system developers design the initial prompt that establishes the assistant's tone, role, and operational logic, users also act as "prompt engineers" during interactions. Unlike traditional VAs, where interactions are pre-programmed intents, LLM-CAs allow users to issue open-ended, custom prompts. However, not all users know how to prompt effectively. Research shows that users are often unfamiliar with prompt engineering and struggle to phrase their requests in ways to get responses that they are looking for [144]. This knowledge gap can leave users unaware of system capabilities, ultimately limiting engagement. To bridge this gap, we propose:

- For system prompt design, the system's initial prompt should establish its role, such as a "cooking assistant", to offer task-relevant guidance, prevent unrelated responses, and support users in discovering relevant features.
- Add suggestions at the end of relevant responses to guide users on how to phrase their queries. For instance, after telling the user to add salt to the salad, the system could add a

note (e.g., *"You can also ask for help if you've added too much salt."*) to teach users effective query patterns.

5.1.4 Calibrate Trust in System Accuracy and Reliability.

While acknowledging the extensive capabilities of LLM-CAs, we also recognize the challenges posed by integrating LLMs into voice assistants. LLMs are known to suffer from issues such as hallucination caused by overconfidence in responses, sycophancy and more [65]. These challenges can potentially decrease user trust and result in harmful outcomes.

It is still important for humans not to fully rely on LLM-CAs, as humans can apply prior knowledge and common sense to validate responses [1, 114]. This becomes particularly critical when user input diverges from the LLM's responses, which may lead to hallucination, particularly when speech-to-text errors occur [71]. In our study, participants were shown a recipe video before using the system, giving them a baseline understanding of the steps expected from the LLM-CA. However, our system occasionally generated unreliable answers, though these instances were not frequent. One notable example occurred when P1 mentioned having only a quarter cup of olive oil, while the recipe required 1/8 cup. Instead of advising how to measure the required amount, *Mango Mango* incorrectly suggested altering the recipe to accommodate a quarter cup.

To enhance the collaborative experience, it is important to balance encouraging user exploration with effectively communicating the reliability of answers to users, thereby building a trustworthy system [32, 108, 116]. Maintaining transparency about the system's capabilities can familiarize users with LLM-CAs strengths and limitations, fostering more collaborative interactions [83, 145, 146]. As detailed in Sec. 4.1.1, these types of misguidance highlight the importance of careful design considerations when creating LLM-based voice assistants for complex tasks. Therefore we propose the following recommendations to mitigate such issues in LLM-CAs:

- Indicate the level of confidence in responses. In cooking scenarios, after providing an uncertain answer, the system could add a statement like, *"I am not very sure about this response since it's outside of the recipe,"*. This will help build user's trust in the system's reliability.
- Including brief reasoning in responses. This helps users actively engage with the system, especially in tasks where accuracy matters or trust is important. It allows users to quickly check if a response is correct and provide feedback or ask for clarification, supporting shared decision-making. In critical situations where mistakes could have serious consequences, such as cooking with food on the stove, involving users in validating the response builds trust and creates a stronger collaboration between the user and the system.

5.1.5 Implement an adaptive response style.

Traditional CAs, designed for simple tasks like setting timers or playing music, operated within rigid rule-based frameworks and limited question sets. Previous research recommended concise, straightforward responses for transactional interactions where flexibility or deeper engagement was unnecessary [45]. However, further studies on using traditional CAs for tasks like cooking revealed key challenges. For example, in cooking scenarios, prior investigations also revealed tradeoffs between information overload and insufficient detail in responses [29, 58, 70].

While *Mango Mango* utilized LLM to partially address the challenge by providing adaptive responses, the unique capabilities of LLMs also offered opportunities to re-evaluate and improve design implications. One significant opportunity is the ability to address the challenge of information overload better when working with CAs, especially in cooking scenarios [58]. Previous research suggested that to mitigate information overload, assistants should carefully manage how they process and deliver information [58, 59]. With LLMs, there is now greater flexibility to dynamically tailor responses based on the user questions with prompting instructions, including customizing

Design consideration	Example from experiment	Solution	Example in cooking
1. Contextualize a universal LLM-CA for specific tasks	Many inquiries are lack of direct references to the recipe and the type of tasks	**Extract context from conversation history** **Proactively ask for user confirmation**	Offer step guidance based on conversation history Proactively verify user's current cooking stage
2. Augment current LLM's knowledge base	Question like *"How many calories are in the salad?"(P8)* could not be answered accurately using current knowledge base	**Incorporate special training materials** **Fine-tune the model on specific task**	Incorporate information such as nutritional details and fundamental cooking techniques
3. Elicit feature discovery instead of expectation alignment	*Mango Mango* can provide guidance on correcting cooking mistakes, but no users are aware of this feature	**Add suggestions at the end of relevant response**	*System:"..., I can also help if you've added too much salt."*
4. Calibrate trust in system accuracy and reliability	*P1: "I've only got a quarter cup."* *MM: "In that case, you can use one quarter cup of extra virgin olive oil instead of 1/8 cup."*	**Indicate the level of confidence in responses**	*System:"..., I am not very sure about the response since it's outside of the recipe"*
5. Implement an adaptive response style	Trade-off between offering a list of ingredients in response and not enough details on how to complete a step	**Rephrase the response based on implicit expressions in user's request** **Respond as user's perception of the system role**	*User: "What are the ingredients again?"* *System:"Let me repeat with less information..."* System as tool, personal assistant or partner during cooking

Table 4. Summary of the five design considerations with interaction examples from our experiment, corresponding solutions, and example solutions in the context of cooking.

both the content and tone of responses. In our system design, we incorporated prompts instructing the system to answer questions by following the rule with fewer words and providing a response depending on the user's question. For example, the system should include the measurement details only when users ask about ingredient measurements. However, if users ask for required ingredients, the system should exclude measurement details to avoid unnecessary cognitive load. Similarly, the system adapts its tone based on the conversational context. For instance, in casual, low-urgency interactions, a humorous tone can enhance user engagement and learning [58]. In contrast, for more urgent or task-critical contexts, a clear, direct tone is preferred. However, some users in our study reported that they forgot the other ingredients mentioned after adding one ingredient. Therefore, the conversational context in complex tasks is also important [59]. Unlike traditional CAs, LLM-CAs in our study leverage conversation history to deliver more contextual and relevant responses. Using LLMs, we suggest an opportunity for greater flexibility in tailoring responses based on the task context. Tasks with low cognitive load, such as placing items in a bowl, can be grouped with related steps to enhance efficiency. Conversely, high cognitive tasks, like measuring ingredients, should be presented individually to minimize the risk of overwhelming the user.

In sum, with the capabilities of LLMs, we've introduced a novel response style—adaptive—that dynamically adjusts the tone and amount of information in responses throughout the duration of a task by integrating user feedback. We suggest three methods for implementing the adaptive response style:

- Rephrase the response through implicit expressions in user's request. Although not always explicitly stated as a command, the expressions in a user's request often reflect whether and how the future response style should be adjusted. For instance, when a user asks the system to repeat the necessary ingredients, it often implies that the previous response from the system contained too much information, prompting the need for a more concise instruction next time.
- Respond as user's perception of the system role. Through our analysis, we discovered that users perceived our system's role differently in their interactions. As a result, in real-time scenarios, the system needs to discern these roles and stages based on user responses. In our cooking experiment, we identified three roles (tool, personal assistant, partner) that should dictate tailored responses.
- LLM-CAs can be prompted to recognize the urgency and complexity of a situation based on contextual cues, such as when a user is handling tasks that require immediate attention. Instead of relying solely on the length of responses, the system should adjust the response style according to task completion metrics like task complexity, cognitive load, and urgency. For instance, even a brief response like an ingredient name can impose a high cognitive load, as users need to locate, measure, and add ingredients. In such cases, the system should avoid combining multiple instructions.

5.2 The Shift From Other Forms of Recipes to LLM-CAs

In our study, participants began by watching a YouTube version of the recipe before using the LLM-CA to prepare the salad. Watching the video gave participants a visual reference of the recipe's general flow, ingredients, and key techniques. However, many participants noted that they often needed to rewatch parts of the video while following the instructions when cooking with new recipes, especially for the ingredient measurements. In contrast, with *Mango Mango*, participants mentioned they no longer needed to refer back to the video repeatedly. Instead, they could receive step-by-step guidance directly from the LLM-CA, allowing them to maintain their cooking flow.

LLM-CAs differ from traditional recipe formats, such as YouTube videos, text-based instructions, and human guidance, in supporting users during cooking tasks. One notable shift participants highlighted was the ability to engage with the LLM-CA for on-demand, context-specific support. Unlike static video instructions, which require users to pause, rewind, or search for specific information, *Mango Mango* allowed users to ask questions during the cooking process without stopping. For instance, participants mentioned that if they had questions mid-cooking—like how many calories are in the salad or clarification about a step—they could directly ask the LLM-CA. This interaction level contrasts with searching for answers online, typically requiring stopping the cooking process, navigating a device, and sifting through search results. Furthermore, P5 mentioned perceiving LLM-CAs as more "teacher-like" than traditional recipe formats. Instead of passively following instructions, participants felt that *Mango Mango* acted as a supportive instructor, offering contextual guidance and adapting to user input. This is distinct from following a static text or video recipe, where the user must interpret instructions independently.

Despite these benefits, participants noted that LLM-CAs cannot fully replace human guidance. Both P7 and P12 mentioned that in situations where their mother is unreachable, they would be more inclined to turn to *Mango Mango* for guidance, especially when cooking from scratch. LLM-CAs serve as a secondary source of assurance, helping users validate their cooking process. However, they are not a full replacement for the experience of seeking advice from a mother, as human connection and expertise provide different connections that LLM-CAs cannot fully replicate.

5.3 LLM-CAs as Motivational and Instructional Tools in Cooking

In our study, we envisioned *Mango Mango* as both an instructional tool, guiding users through completing their cooking tasks. However, our analysis also revealed that *Mango Mango* could serve as a motivational tool, promoting a learning procedure while completing the cooking tasks. This finding aligns with prior research showing that CAs can support user learning by providing tailored guidance and fostering engagement [68]. With LLM-CAs, our participants envision learning customized content in new and meaningful ways. For example, one participant expressed a desire to use *Mango Mango* to preserve their mother's unique recipes and revisit them anytime. Its ability to adapt to user-specific content enables self-directed learning experiences [36, 85]. By tailoring responses to individual needs, *Mango Mango* can serve as a repository for culturally significant or family-specific recipes, creating a bridge between technology and personal heritage.

Participants also highlighted the potential of LLM-CAs to motivate and reinforce learning through interactive and engaging responses. For example, integrating features such as real-time feedback on user progress could enhance skill development during tasks like cooking. By actively engaging with users through tailored guidance and motivational cues, LLM-CAs could transform routine activities into opportunities for experiential learning. This ability to combine instructional guidance with motivational support underscores the unique potential of LLM-CAs to enhance learning in real-world contexts.

5.4 Integrating Multi-Modal Features and Addressing Privacy Concerns

In our study, participants highlighted the potential of integrating multi-modal features, such as a camera, to enhance the capabilities of *Mango Mango*. By incorporating real-time visual recognition from a camera, the LLM-CA can use this visual information as additional context to acknowledge the current step or status of the food, enabling the LLM-CA to provide more accurate and context-aware guidance. Advances in multi-modal LLMs, such as GPT-4V, offer a promising way to realize this potential. With its ability to process both visual and textual inputs, facilitating more collaborative, adaptive, and context-aware systems [128]. For example, in cooking scenarios, a camera could allow the system to identify whether the ingredients were placed or the progress of a recipe step. By processing visual input as contextual information with LLM, it can then provide immediate feedback aligned with users' specific actions, reinforcing skill development, improving task accuracy, and reducing cognitive load.

However, participants also raised concerns about the privacy implications of embedding a camera into the system. While multi-modal systems have been explored in prior research to enhance cooking assistance [46, 99, 111], integrating such features into LLM-based CAs introduces unique challenges. Using a camera has raised privacy concerns, particularly in home settings [21, 44, 97, 103]. For instance, users may be hesitant to adopt a system that continuously monitors their actions, even if it enhances the functionality.

Future research could explore strategies to mitigate these privacy concerns while maintaining the benefits of multi-modal capabilities. Possible solutions include implementing strict data processing and storage policies, offering transparent explanations of how visual data is used, and enabling users to toggle the camera on or off. Additionally, investigating user acceptance of multi-modal features in LLM-CAs compared to traditional CAs could provide insights into balancing enhanced functionality with user comfort and trust. Solving these challenges could enable multi-modal LLM-CAs to enhance collaboration while maintaining users' privacy.

5.5 Problem of Hallucination in LLM-Generated Content

Hallucination, where LLMs generate nonsensical, irrelevant, or unfaithful content to the user input, remains a significant problem [9, 55, 62]. These inaccuracies are concerning in various applications where incorrect information could lead to serious consequences. Prior research has explored the use of LLMs and identified the issue of hallucination in diverse scenarios, such as academic research [67], coding [120], medical [123], and data annotation [129]. For example, Kapania et al. [67] conducted a study on using LLMs for HCI research, where participants expressed concerns that relying on LLMs during the paper writing process could result in misinformation or incorrect references, potentially leading to significant issues in academic work. Another related challenge is sycophancy, where LLMs generate responses based on users' beliefs rather than objective facts [115]. When using LLM, the ability to engage in multi-round conversations allow the system to accumulates the conversation and uses it as context for future interactions [55]. Manzini et al. [93] found sycophancy could preventing users from critically assessing their own assumptions that can negatively impact human-AI interaction.

In our cooking experiment, we also observed participants encountering hallucination-related issues. When participants faced suggestions that did not align with their understanding of the recipe from the YouTube video, they would pause and seek clarification, often by rephrasing their query or expressing doubt about the system's response. Participants' prior knowledge of how cooking steps should look played a critical role in their ability to identify and question hallucinated responses from *Mango Mango*. Since users rely on accurate, step-by-step guidance during cooking, hallucinations in LLM-generated instructions can present significant challenges. For instance, if a user follows a suggested adjustment for ingredient measurements but later tastes the dish and encounters an unexpected taste. Unlike visual interactions, where users can check progress by inspecting a sauce's consistency, hallucinated suggestions from LLMs can mislead users without visual cues for validation. This reliance on system guidance increases the potential for user errors, especially when users have no immediate way to verify the correctness of the LLM's responses. However, participants' prior knowledge enabled them to detect inconsistencies between their expectations and the system's guidance. This observation aligns with existing research on tasks like data annotation [129], where human evaluation is critical in identifying and mitigating hallucinated outputs from LLMs.

5.6 Inclusive Accessibility

In this study, we conducted experiments with a general population and discovered that LLM-CAs offer multiple advantages over conventional cooking assistance methods. Key benefits include the ability to request information beyond the recipe and more adaptive support aligned with the natural flow of cooking. Beyond general use, we recognize that these capabilities could further support marginalized populations who are often excluded by conventional cooking tools for cooking assistance, offering them greater accessibility and support.

The potential of LLM-CAs to support diverse user group has been demonstrated in prior research across various domains. Conventional tools often fail to support individuals who face challenges with complex interfaces, visual demands, or rigid, linear instructions [80]. By enabling conversational, hands-free, and context-aware interactions, LLM-CAs make it easier for users to stay engaged in the task while reducing cognitive load. For example, Kaniwa et al. [66] leveraged LLMs to provide conversational guidance for visually impaired individuals navigating a shopping mall. Participants reported that the system offered easy access to various information, enabled them to ask follow-up questions, and had the ability to plan visit tours, leading to more effective exploration. Similarly, Gorniak et al. [42] utilized LLMs to facilitate conversational interaction for visually

impaired users to navigate visual data using voice commands. Beyond navigation and visual data interaction, LLM-CAs have also been leveraged in healthcare applications, such as supporting communication between older adults and their providers [143]. These LLM-based applications show flexibility and adaptability in supporting diverse accessibility needs across multiple domains.

Given the ample evidence in the literature regarding the potential of LLM-CAs in various domains, our findings suggest that LLM-CAs can also support marginalized groups in cooking contexts. The flexibility and responsiveness observed during our cooking experiments underscore this potential. For instance, older adults frequently rely on technology to track their cooking progress but often struggle to switch between tasks without external support [75]. Age-related cognitive or physical changes may further complicate the process of following multi-step instructions, increasing the need for real-time, accessible guidance [61]. LLM-CAs address this challenge through their conversational adaptability, allowing users to request clarification or task-specific guidance at any point during the cooking process. In addition, individuals with visual impairments face additional barriers when using traditional cooking tools, which often rely heavily on text, images, or video-based instructions [79]. Navigating these visually demanding formats can be difficult, especially when precision is required for cooking steps like ingredient measurements or timing. LLM-CAs offer advantage through voice-based interaction, enabling visually impaired users to access recipes, receive detailed verbal instructions, and clarify steps without relying on visual cues. By providing adaptive, real-time guidance, LLM-CAs demonstrate their potential as tools for enhancing independence and inclusivity in cooking tasks, particularly for those with unique accessibility needs.

5.7 Limitations of the Study

We acknowledge a few limitations in our system and exploratory lab study. While our system evaluation in Section 4.2 did not highlight prominent issues caused by LLM, we acknowledged that certain notorious problems of LLM, such as hallucination, could potentially affect users' experience with *Mango Mango*. Moreover, while we chose cooking as an example due to its representative nature as a sequential but complex real-time task where CAs are already extensively used, we still realized its limitation and tried not to over-generalize our findings to other scenarios, especially considering our small sample size (n=12). Future research could explore the possibility of applying these design implications in different scenarios to test their generalizability and further tailor the results to various use cases in a large-scale study.

Our study focused on a salad-making scenario to minimize safety concerns. As described in the methods section, we selected a relatively complex salad recipe with multiple food preparation and measurement steps to mimic a more comprehensive cooking experience. Simultaneously, we devised a streamlined process for transitioning from recipes to system prompts, aiming to maximize the scalability of our methods. However, the limitation of conducting experiments on only one recipe might still restrict the generalizability of our findings, especially some recipes might involve different cooking processes, for instance, those requiring a stove. Consequently, we aim to articulate the limitations of our experiment clearly. Future research is encouraged to explore various recipes to further enhance the scope of investigation.

In the previous section, we discussed the need for feature discovery of LLM-CA. In our study, we conducted a brief tutorial session beforehand to demonstrate some example questions like "What is the first step?", "What if I don't have chicken, what should I do?", and "What did I just ask?". Despite the training session, users might still require some trial and error to discover their preferred way of using and communicating with the system, which we summarized as one design implication, as well as pointed out as one of the limitations of our experiment.

Lastly, since our primary focus was not on quantitative results, and we did not conduct a comparative study that would provide a baseline for analysis, we mainly used those results to verify

our system's basic usability and general user performance. Future work could involve comparative studies to assess the effectiveness of such systems across various dimensions.

6 CONCLUSION

In this study, we explored users' experiences, thoughts, and expectations while interacting with an LLM-based CA system and synthesized design implications for future systems. To achieve this, we conducted a mixed-methods exploratory study with 12 participants and asked them to complete a salad recipe with assistance from our system. We then examined their experiences using surveys, interviews, and interactive logs. Our findings revealed that users quickly adapted to the LLM's capabilities to assist their cooking practices, including asking for extensive information, requesting personalized and context-aware assistance, and dynamically planning their tasks. However, users also expressed the desire for the system to facilitate more natural and oral conversations. Additionally, participants wanted to be more involved in the decision-making process of the CA, suggesting a potential shift in their perception of the system from a tool to a personal assistant and even a partner. Based on these observations, we synthesized design implications for a future LLM-CA.

REFERENCES

[1] Bhashithe Abeysinghe and Ruhan Circi. 2024. The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches. *arXiv preprint arXiv:2406.03339* (2024).

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. arXiv:2204.01691 [cs.RO]

[3] Majid Alfifi, Xiangjue Dong, Timo Feldman, Allen Lin, Karthic Madanagopal, Aditya Pethe, Maria Teleki, Zhuoer Wang, Ziwei Zhu, and James Caverlee. [n. d.]. Howdy Y'all: An Alexa TaskBot. ([n. d.]).

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

[5] Merav Allouch, Amos Azaria, and Rina Azoulay. 2021. Conversational agents: Goals, technologies, vision and challenges. *Sensors* 21, 24 (2021), 8448.

[6] 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.

[7] 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.

[8] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* (2023).

[9] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[10] Vince Bartle, Liam Albright, and Nicola Dell. 2023. "This machine is for the aides": Tailoring Voice Assistant Design to Home Health Care Work. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–19.

[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 *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 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 (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376344

[14] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 91 (sep 2018), 24 pages. https://doi.org/10.1145/3264901

[15] Pascal Bercher, Gregor Behnke, Matthias Kraus, Marvin Schiller, Dietrich Manstetten, Michael Dambier, Michael Dorna, Wolfgang Minker, Birte Glimm, and Susanne Biundo. 2021. Do it yourself, but not alone: companion-technology for home improvement—bringing a planning-based interactive DIY assistant to life. *KI-Künstliche Intelligenz* 35, 3-4 (2021), 367–375.

[16] Caterina Bérubé, Zsolt Ferenc Kovacs, Elgar Fleisch, and Tobias Kowatsch. 2021. Reliability of Commercial Voice Assistants' Responses to Health-Related Questions in Noncommunicable Disease Management: Factorial Experiment Assessing Response Rate and Source of Information. *Journal of Medical Internet Research* 23, 12 (Dec. 2021), e32161.

[17] Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. 2007. Large language models in machine translation. (2007).

[18] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[19] 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* 12, 2 (June 2022), 1–32.

[20] Julia Cambre, Alex C Williams, Afsaneh Razi, Ian Bicking, Abraham Wallin, Janice Tsai, Chinmay Kulkarni, and Jofish Kaye. 2021. Firefox Voice: An Open and Extensible Voice Assistant Built Upon the Web. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–18.

[21] George Chalhoub, Martin J Kraemer, Norbert Nthala, and Ivan Flechais. 2021. "It did not give me an option to decline": A Longitudinal Analysis of the User Experience of Security and Privacy in Smart Home Products. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[22] 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.

[23] 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.

[24] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[25] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[26] Jennifer Chubb, Sondess Missaoui, Shauna Concannon, Liam Maloney, and James Alfred Walker. 2022. Interactive storytelling for children: A case-study of design and development considerations for ethical conversational AI. *International Journal of Child-Computer Interaction* 32 (2022), 100403.

[27] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).

[28] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: visual sketching of story generation with pretrained language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–4.

[29] Eric Corbett and Astrid Weber. 2016. What can I say? addressing user experience challenges of a mobile voice user interface for accessibility. In *Proceedings of the 18th international conference on human-computer interaction with mobile devices and services*. 72–82.

[30] 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* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 43, 12 pages. https://doi.org/10.1145/3098279.3098539

[31] Fergus IM Craik and Ellen Bialystok. 2006. Planning and task management in older adults: Cooking breakfast. *Memory & Cognition* 34, 6 (2006), 1236–1249.

[32] 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.

[33] Emily Czekalski and David Watson. 2024. Efficiently updating domain knowledge in large language models: Techniques for knowledge injection without comprehensive retraining. (2024).

[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] Griffin Dietz, Jimmy K Le, Nadin Tamer, Jenny Han, Hyowon Gweon, Elizabeth L Murnane, and James A. Landay. 2021. StoryCoder: Teaching Computational Thinking Concepts Through Storytelling in a Voice-Guided App for Children. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 54, 15 pages. https://doi.org/10.1145/3411764.3445039

[36] Gilbert Dizon. 2024. ChatGPT as a tool for self-directed foreign language learning. *Innovation in Language Learning and Teaching* (2024), 1–17.

[37] Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. "Hey Google is It OK If I Eat You?": Initial Explorations in Child-Agent Interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children* (Stanford, California, USA) *(IDC '17)*. Association for Computing Machinery, New York, NY, USA, 595–600. https://doi.org/10.1145/3078072.3084330

[38] Yao Du, Kerri Zhang, Sruthi Ramabadran, and Yusa Liu. 2021. "Alexa, What is That Sound?" A Video Analysis of Child-Agent Communication From Two Amazon Alexa Games. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference* (Athens, Greece) *(IDC '21)*. Association for Computing Machinery, New York, NY, USA, 513–520. https://doi.org/10.1145/3459990.3465195

[39] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* 2 (2023).

[40] Radhika Garg and Subhasree Sengupta. 2020. He Is Just Like Me: A Study of the Long-Term Use of Smart Speakers by Parents and Children. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 11 (mar 2020), 24 pages. https://doi.org/10.1145/3381002

[41] Barney Glaser and Anselm Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

[42] Joshua Gorniak, Yoon Kim, Donglai Wei, and Nam Wook Kim. 2024. Vizability: Enhancing chart accessibility with llm-based conversational interaction. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–19.

[43] Jonathan Grudin and Richard Jacques. 2019. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300439

[44] Neilly H. Tan, Richmond Y. Wong, Audrey Desjardins, Sean A. Munson, and James Pierce. 2022. Monitoring pets, deterring intruders, and casually spying on neighbors: Everyday uses of smart home cameras. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–25.

[45] Gabriel Haas, Michael Rietzler, Matt Jones, and Enrico Rukzio. 2022. Keep it Short: A Comparison of Voice Assistants' Response Behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–12.

[46] 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.

[47] Songhee Han and Min Kyung Lee. 2022. FAQ chatbot and inclusive learning in massive open online courses. *Computers & Education* 179 (2022), 104395.

[48] 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 *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15.

[49] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[50] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[51] Masum Hasan, Cengiz Ozel, Sammy Potter, and Ehsan Hoque. 2023. SAPIEN: affective virtual agents powered by large language models. *arXiv preprint arXiv:2308.03022* (2023).

[52] Daniel Hocutt. 2021. Interrogating Alexa: Holding Voice Assistants Accountable for Their Answers. In *Proceedings of the 39th ACM International Conference on Design of Communication*. 142–150.

[53] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[54] Matthew B Hoy. 2018. Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical reference services quarterly* 37, 1 (2018), 81–88.

[55] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* (2023).

[56] Ting-Hao (Kenneth) Huang, Joseph Chee Chang, and Jeffrey P. Bigham. 2018. Evorus: A Crowd-Powered Conversational Assistant Built to Automate Itself Over Time. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173869

[57] Alyssa Hwang, Bryan Li, Zhaoyi Hou, and Dan Roth. 2023. Large Language Models as Sous Chefs: Revising Recipes with GPT-3. *arXiv preprint arXiv:2306.13986* (2023).

[58] Alyssa Hwang, Natasha Oza, Chris Callison-Burch, and Andrew Head. 2023. Rewriting the Script: Adapting Text Instructions for Voice Interaction. *arXiv preprint arXiv:2306.09992* (2023).

[59] 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.

[60] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[61] Agnieszka J Jaroslawska, Glen Bartup, Alicia Forsberg, and Joni Holmes. 2021. Age-related differences in adults' ability to follow spoken instructions. *Memory* 29, 1 (2021), 117–128.

[62] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[63] 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.

[64] Ellen Jiang, Edwin Toh, Alejandra Molina, Kristen Olson, Claire Kayacik, Aaron Donsbach, Carrie J Cai, and Michael Terry. 2022. Discovering the syntax and strategies of natural language programming with generative language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[65] Uday Kamath, Kevin Keenan, Garrett Somers, and Sarah Sorenson. 2024. LLM challenges and solutions. In *Large Language Models: A Deep Dive: Bridging Theory and Practice*. Springer, 219–274.

[66] Yuka Kaniwa, Masaki Kuribayashi, Seita Kayukawa, Daisuke Sato, Hironobu Takagi, Chieko Asakawa, and Shigeo Morishima. 2024. ChitChatGuide: Conversational Interaction Using Large Language Models for Assisting People with Visual Impairments to Explore a Shopping Mall. *Proceedings of the ACM on Human-Computer Interaction* 8, MHCI (2024), 1–25.

[67] Shivani Kapania, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, and Hong Shen. 2024. " I'm categorizing LLM as a productivity tool": Examining ethics of LLM use in HCI research practices. *arXiv preprint arXiv:2403.19876* (2024).

[68] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[69] Tae Soo Kim, DaEun Choi, Yoonseo Choi, and Juho Kim. 2022. Stylette: Styling the web with natural language. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[70] Yelim Kim, Mohi Reza, Joanna McGrenere, and Dongwook Yoon. 2021. Designers characterize naturalness in voice user interfaces: their goals, practices, and challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[71] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless Whisper: Speech-to-Text Hallucination Harms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1672–1681.

[72] 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.

[73] Matthias Kraus, Marvin Schiller, Gregor Behnke, Pascal Bercher, Michael Dorna, Michael Dambier, Birte Glimm, Susanne Biundo, and Wolfgang Minker. 2020. " Was that successful?" On Integrating Proactive Meta-Dialogue in a DIY-Assistant using Multimodal Cues. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 585–594.

[74] Harsh Kumar, Yiyi Wang, Jiakai Shi, Ilya Musabirov, Norman AS Farb, and Joseph Jay Williams. 2023. Exploring the Use of Large Language Models for Improving the Awareness of Mindfulness. In *Extended Abstracts of the 2023 CHI*

Conference on Human Factors in Computing Systems. 1–7.

[75] Sanna Kuoppamäki, Sylvaine Tuncer, Sara Eriksson, and Donald McMillan. 2021. Designing Kitchen Technologies for Ageing in Place: A Video Study of Older Adults' Cooking at Home. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 2 (2021), 1–19.

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

[77] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[78] Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. DAPIE: Interactive Step-by-Step Explanatory Dialogues to Answer Children's Why and How Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–22.

[79] Franklin Mingzhe Li, Ashley Wang, Patrick Carrington, and Shaun K Kane. 2024. A Recipe for Success? Exploring Strategies for Improving Non-Visual Access to Cooking Instructions. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–15.

[80] Qingchuan Li and Yan Luximon. 2020. Older adults' use of mobile device: usability challenges while navigating various interfaces. *Behaviour & Information Technology* 39, 8 (2020), 837–861.

[81] 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).

[82] Q Vera Liao, Werner Geyer, Michael Muller, and Yasaman Khazaen. 2020. Conversational interfaces for information search. *Understanding and Improving Information Search: A Cognitive Approach* (2020), 267–287.

[83] Q Vera Liao and Jennifer Wortman Vaughan. 2023. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941* (2023), 5368–5393.

[84] Tze Wei Liew, Su-Mae Tan, Wei Ming Pang, Mohammad Tariqul Islam Khan, and Si Na Kew. 2023. I am Alexa, your virtual tutor!: The effects of Amazon Alexa's text-to-speech voice enthusiasm in a multimedia learning environment. *Education and information technologies* 28, 2 (2023), 1455–1489.

[85] Xi Lin. 2024. Exploring the role of ChatGPT as a facilitator for motivating self-directed learning among adult learners. *Adult Learning* 35, 3 (2024), 156–166.

[86] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What Makes Good In-Context Examples for GPT-3? *arXiv preprint arXiv:2101.06804* (2021).

[87] Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman. 2022. Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–13.

[88] Leo S Lo. 2023. The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship* 49, 4 (2023), 102720.

[89] 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.

[90] 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.

[91] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[92] Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2024. User interaction patterns and breakdowns in conversing with LLM-powered voice assistants. *International Journal of Human-Computer Studies* (2024), 103406.

[93] Arianna Manzini, Geoff Keeling, Lize Alberts, Shannon Vallor, Meredith Ringel Morris, and Iason Gabriel. 2024. The Code That Binds Us: Navigating the Appropriateness of Human-AI Assistant Relationships. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 943–957.

[94] 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.

[95] Michael McTear. 2018. Conversational modelling for chatbots: current approaches and future directions. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung* (2018), 175–185.

[96] 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.

[97] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. 2017. Privacy expectations and preferences in an {IoT} world. In *Thirteenth symposium on usable privacy and security (SOUPS 2017)*. 399–412.

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

[99] 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)*.

[100] Reham Omar, Omij Mangukiya, Panos Kalnis, and Essam Mansour. 2023. Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. *arXiv preprint arXiv:2302.06466* (2023).

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

[102] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[103] Sunyup Park, Anna Lenhart, Michael Zimmer, and Jessica Vitak. 2023. " Nobody's Happy": Design Insights from {Privacy-Conscious} Smart Home Power Users on Enhancing Data Transparency, Visibility, and Control. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*.

[104] Cathy Pearl. 2016. *Designing voice user interfaces: Principles of conversational experiences*. " O'Reilly Media, Inc.".

[105] Indrasen Poola. 2023. Overcoming ChatGPTs inaccuracies with Pre-Trained AI Prompt Engineering Sequencing Process. (2023).

[106] Shachaf Poran, Gil Amsalem, Amit Beka, and Dmitri Goldenberg. 2022. With One Voice: Composing a Travel Voice Assistant from Repurposed Models. In *Companion Proceedings of the Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 383–387. https://doi.org/10.1145/3487553.3524228

[107] Alisha Pradhan, Amanda Lazar, and Leah Findlater. 2020. Use of Intelligent Voice Assistants by Older Adults with Low Technology Use. *ACM Transactions on Computer-Human Interaction* 27, 4 (Aug. 2020), 1–27.

[108] Harsh Raj, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. 2023. Semantic consistency for assuring reliability of large language models. *arXiv preprint arXiv:2308.09138* (2023).

[109] Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering* 3, 1 (1997), 57–87.

[110] Joshua Robinson and David Wingate. 2023. Leveraging Large Language Models for Multiple Choice Question Answering. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=yKbprarjc5B

[111] 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.

[112] 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* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[113] Paul Semaan. 2012. Natural language generation: an overview. *J Comput Sci Res* 1, 3 (2012), 50–57.

[114] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[115] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548* (2023).

[116] Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979* (2023).

[117] Gina EM Stolwijk and Florian A Kunneman. 2022. Increasing the Coverage of Clarification Responses for a Cooking Assistant. In *International Workshop on Chatbot Research and Design*. Springer, 171–189.

[118] Kevin M. Storer, Tejinder K. Judge, and Stacy M. Branham. 2020. "All in the Same Boat": Tradeoffs of Voice Assistant Ownership for Mixed-Visual-Ability Families. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https:

//doi.org/10.1145/3313831.3376225

[119] Ben Swanson, Kory Mathewson, Ben Pietrzak, Sherol Chen, and Monica Dinalescu. 2021. Story centaur: Large language model few shot learning as a creative writing tool. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. 244–256.

[120] Yuchen Tian, Weixiang Yan, Qian Yang, Qian Chen, Wen Wang, Ziyang Luo, and Lei Ma. 2024. CodeHalu: Code Hallucinations in LLMs Driven by Execution-based Verification. *arXiv preprint arXiv:2405.00253* (2024).

[121] 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).

[122] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376760

[123] Prathiksha Rumale Vishwanath, Simran Tiwari, Tejas Ganesh Naik, Sahil Gupta, Dung Ngoc Thai, Wenlong Zhao, SUNJAE KWON, Victor Ardulov, Karim Tarabishy, Andrew McCallum, et al. 2024. Faithfulness Hallucination Detection in Healthcare AI. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*.

[124] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R. Cowan, and Heinrich Hussmann. 2021. Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 254, 15 pages. https://doi.org/10.1145/3411764.3445536

[125] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W White. 2019. Verse: Bridging screen readers and voice assistants for enhanced eyes-free web search. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 414–426.

[126] 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.

[127] Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, Huawen Hu, et al. 2023. Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670* (2023).

[128] Jiaqi Wang, Zihao Wu, Yiwei Li, Hanqi Jiang, Peng Shu, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, et al. 2024. Large language models for robotics: Opportunities, challenges, and perspectives. *arXiv preprint arXiv:2401.04334* (2024).

[129] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM collaborative annotation through effective verification of LLM labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.

[130] 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.

[131] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).

[132] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2023. Leveraging large language models to power chatbots for collecting user self-reported data. *arXiv preprint arXiv:2301.05843* (2023).

[133] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[134] 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.

[135] 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.

[136] Ziang Xiao, Tiffany Wenting Li, Karrie Karahalios, and Hari Sundaram. 2023. Inform the Uninformed: Improving Online Informed Consent Reading with an AI-Powered Chatbot. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 112, 17 pages. https://doi.org/10.1145/3544548.3581252

[137] Ziang Xiao, Q. Vera Liao, Michelle Zhou, Tyrone Grandison, and Yunyao Li. 2023. Powering an AI Chatbot with Expert Sourcing to Support Credible Health Information Access. In *Proceedings of the 28th International Conference on*

Intelligent User Interfaces (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 2–18. https://doi.org/10.1145/3581641.3584031

[138] Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. 75–78.

[139] Ziang Xiao, Michelle X. Zhou, Q. Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-Ended Questions. *ACM Trans. Comput.-Hum. Interact.* 27, 3, Article 15 (jun 2020), 37 pages. https://doi.org/10.1145/3381804

[140] 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]

[141] 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).

[142] Ying Xu, Kunlei He, Valery Vigil, Santiago Ojeda-Ramirez, Xuechen Liu, Julian Levine, Kelsyann Cervera, and Mark Warschauer. 2023. "Rosita Reads With My Family": Developing A Bilingual Conversational Agent to Support Parent-Child Shared Reading. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (Chicago, IL, USA) *(IDC '23)*. Association for Computing Machinery, New York, NY, USA, 160–172. https://doi.org/10.1145/3585088.3589354

[143] 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.

[144] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[145] Zhiping Zhang, Bingcan Guo, and Tianshi Li. 2024. Can Humans Oversee Agents to Prevent Privacy Leakage? A Study on Privacy Awareness, Preferences, and Trust in Language Model Agents. *arXiv preprint arXiv:2411.01344* (2024).

[146] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery New York, NY, USA, 1–26.

[147] 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* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 218, 21 pages. https://doi.org/10.1145/3491102.3517479

[148] 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.

[149] Tamara Zubatiy, Kayci L Vickers, Niharika Mathur, and Elizabeth D Mynatt. 2021. Empowering dyads of older adults with mild cognitive impairment and their care partners using conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[150] Dilawar Shah Zwakman, Debajyoti Pal, and Chonlameth Arpnikanondt. 2021. Usability evaluation of artificial intelligence-based voice assistants: The case of Amazon Alexa. *SN Computer Science* 2 (2021), 1–16.

A APPENDIX

A.1 Prompt Sample for *Mango Mango*

Prompt for *Mango Mango* (Part 1: Knowledge Resources)
RECIPE = INGREDIENTS FOR CHICKEN AVOCADO MANGO SALAD - 1 1/2 cups or 1/4 head romaine lettuce, rinsed, chopped and spun dry - 1/4 lb or 1/2 medium cooked chicken breasts - 1/4 mango, pitted, peeled and diced - 1/4 avocado, pitted, peeled and diced - 1/8 english cucumber sliced - 1/8 thinly sliced small purple onion - 1/8 cup halved cherry tomatoes - 1/16 cup chopped cilantro chopped STEPS - Step 1: Chop the romaine into bite-sized pieces and discard the core. After rinse and spin dry, place it in a large salad bowl. - Step 2: Slide chicken into bite size strips and place it over the romaine lettuce. - Step 3: Place diced mango in to salad bowl. - Step 4: Peel and dice the advocado, then place it on top of the salad bowl. - Step 5: Place slices cucumber in to salad bowl. - Step 6: Added thinly sliced small purple onion. - Step 7: Cut the cherry tomatoes into half and place it on the salad. - Step 8: Add chopped fresh cilantro. INGREDIENTS FOR HONEY VINAIGRETTE DRESSING - 1/8 cup extra virgin olive oil - 3/4 Tbsp apple cider vinegar - 1/2 tsp dijon mustard - 1/2 tsp honey - 1/4 garlic clove or 1/4 tsp minced garlic - 1/4 tsp sea salt - 1/16 tsp black pepper, or to taste - Step 9: Combine the Honey Vinaigrette Dressing Ingredients in a mason jar, first add olive oil. - Step 10: Add apple cider vinegar, Dijon mustard and honey - Step 11: Add garlic, sea salt and black pepper - Step 12: Cover tightly with lid and shake together until well combined. - Step 13: Drizzle the salad dressing over the chicken mango avocado salad, adding it to taste.

Table 5. Prompt for *Mango Mango* of Knowledge Resources.

Prompt for *Mango Mango* (cont. Part 2: Instructions)

INSTRUCTIONS =
Your main task is to help guiding user to make the chicken avocado mango salad step by step based on the recipe provided delimited by triple backticks.
The recipe is for 1 person.
There are 2 parts of this recipe: the salad part and the dressing part.
Please follow these steps to guide user by answering the customer queries.

1: First decide whether the user is asking a question about a specific ingredients or recipe steps or other. When user ask for next step, assume user is about to perform that step.
Once the dressing steps are finished or all the ingredients are placed, the entire recipe is complete, and no more futher steps since all salad and dressing steps and ingredients covered. Congratulate user and tell user all the steps are complete.

2: If the user is asking about overall ingredients, for example: how to make the dressing. Respond with all the ingredients without measurements, for example: The ingredients for chicken avocado mango salad are romaine lettuce, chicken breasts. Do not respond: The ingredients for chicken avocado mango salad are 1 lb or 2 medium cooked chicken breasts and 6 cups or 1 head romaine lettuce.

3: If the user is asking about one specific ingredients. Identify whether the ingredients is for the salad or the salad dressing, then respond corresponding ingredients with measurement. For example: 1/2 thinly sliced small purple onion is needed for the salad.

4: If the user is asking about specific steps, identify what step of the recipe the user is working on, then respond with short, clear and easy to follow instructions.

5: Respond to user with summarizing the response from steps above in 30 words or less. Please response in complete sentence. Please aim to be as helpful, creative, friendly, and educative as possible in all of your responses.
Do not use any external recipe in your responses.
For question not related to this recipe, try your best to answer it.

Table 6. Prompt for *Mango Mango* of Instructions

A.2 Questionnaire Results

Voice Usability Scale(VUS) Questions	Mean(SD)
Usability	
I thought the information provided by the *Mango Mango* was not relevant to what I asked.	2.333(1.371)
I thought the *Mango Mango* had difficulty in understanding what I asked it to do.	2.083(0.515)
I found the *Mango Mango* difficult to use.	1.667(0.985)
Affective	
I felt the *Mango Mango* enabled me to successfully complete my tasks when I required help.	4.500(0.522)
The *Mango Mango* had all the functions and capabilities that I expected it to have.	4.333(0.985)
I felt the response from the *Mango Mango* was sufficient.	3.833(1.115)
Overall, I am satisfied with using the *Mango Mango*.	4.333(0.888)
Recognizability & Visibility	
I thought the response from the *Mango Mango* was easy to understand.	4.333(0.888)
I found it difficult to customize the *Mango Mango* according to my needs and preferences.	2.167(1.030)

Table 7. The questions of Voice Usability Scale(VUS) and results in the format of Mean (Standard Deviation)

Explainable AI (XAI) Questions	Mean(SD)
I am confident in the *Mango Mango*. I feel that it works well.	4.250(0.866)
The outputs of the *Mango Mango* are very predictable.	4.250(0.965)
I feel safe that when I rely on *Mango Mango* that I will get the right response.	3.917(0.996)
Mango Mango is efficient in that it works very quickly.	4.000(1.348)
Mango Mango can better help me than the recipes in other formats.	3.917(1.443)
I like using *Mango Mango* for cooking instructions.	4.333(1.073)

Table 8. The questions of Explainable AI (XAI) survey and results in the format of Mean (Standard Deviation)

NASA-TLX Questions	Mean(SD)
How mentally demanding was it to interact with *Mango Mango*?	2.583(1.240)
How physically demanding was it to interact with *Mango Mango*?	1.917(1.379)
How hurried or rushed was it to interact with *Mango Mango*?	2.000(0.853)
How successful were you in communicating with *Mango Mango*?	3.750(0.965)
How hard did you have to try to communicate with *Mango Mango*?	2.667(1.073)
How insecure, discouraged, irritated, stressed, and annoyed were you communicating with *Mango Mango*?	2.167(1.267)

Table 9. The questions of NASA-TLX and results in the format of Mean (Standard Deviation)

Exploration Questions	Mean(SD) Pre Study	Mean(SD) Post Study
I thought the current voice assistants could engage in fluent and human-like conversations.	3.500(0.798)	4.167(0.835)
I thought the current voice assistant has the ability to remember and refer back to previous parts of a conversation.	3.167(0.937)	4.167(0.718)
I thought the current voice assistant allowed asking follow-up questions that relate to the ongoing conversation.	3.417(0.996)	4.333(0.778)
I thought the current voice assistant can seamlessly integrate into my daily activities.	3.333(0.985)	3.917(0.900)
I thought the current voice assistant can actively collaborate with me on different tasks.	3.167(0.835)	3.833(1.193)

Table 10. The five supplementary questions we created to understand users' perspectives. Pre-study and post-study results in the format of Mean (Standard Deviation)

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