"*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, Rensselaer Polytechnic Institute, 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 the Large Language Model (LLM) 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 investigate 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 provide extensive information beyond the recipe, offer customized instructions based on context, and assist them in dynamically planning the task. However, they expect the system to be more adaptive to oral conversation and provide more suggestive responses to keep users 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 several design considerations for future development.

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

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

---

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

Authors' addresses: Szeyi Chan, chan.szey@northeastern.edu, Northeastern University, USA; Jiachen Li, li.jiachen4@northeastern.edu, Northeastern University, USA; Bingsheng Yao, arthuryao33@gmail.com, Rensselaer Polytechnic Institute, USA; Amama Mahmood, amahmo11@jhu.edu, Johns Hopkins University, USA; Chien-Ming Huang, chienming.huang@jhu.edu, Johns Hopkins University, USA; Holly Jimison, h.jimison@northeastern.edu, Northeastern University, USA; Elizabeth D Mynatt, e.mynatt@northeastern.edu, Northeastern University, USA; Dakuo Wang, d.wang@northeastern.edu, Northeastern University, USA.

---

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 [4, 39, 69]. The "hands-free" 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 [56, 72, 92].

Nevertheless, CAs face limitations and challenges when instructing users with hands-on tasks at home. People frequently encounter family-centered problems in daily life, such as experimenting with new recipes for special family dinners, resolving urgent plumbing problems, or collaboratively assembling new furniture together to improve their living spaces. 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 [13, 47]. 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 [26, 58, 59]. Unfortunately, current CAs cannot provide comprehensive, continuous support with these tasks [71, 71]. Existing virtual assistants rely on predefined dialogue logic and often struggle with language comprehension, prohibiting natural back-and-forth conversations for complex tasks [5, 6, 19, 25, 55, 71, 75].

Recent advancements in language models, particularly large language models (LLMs), for example, GPT-3.5/4 [61], LLaMA [74], and PaLM [20], show the ability to overcome the limitations of language models used in current CAs. Existing works have shown LLMs have natural language understanding (NLU) [3, 70] and generation (NLG) [66, 70] capabilities to understand users' lengthy text input and accommodate multi-turn dialogues [88]. Yet, the research community has not explored integrating LLM into CA in real-world settings.

To explore the integration of LLMs into CAs, our research consists of two parts: 1) developing an LLM-based system called *Mango Mango* and 2) conducting a mixed-method study to evaluate users' experiences and establish design guidelines. Using cooking as a case study, *Mango Mango*, powered by the GPT 3.5 Turbo[1] and integrated with the Amazon Alexa Skill Kit[2], was specifically tailored to help individuals cook at home while following recipes. It provides users with step-by-step cooking instructions and responds to cooking-related queries, even when users' hands are occupied. The user study involved a lab-based exploration where participants prepare a salad with assistance from *Mango Mango* in a one-bedroom apartment lab setting. Following the study, we performed both qualitative and quantitative research analyses with semi-structured interviews, surveys, and system logs. We aim to **investigate user perceptions of the system through their successful and unsatisfactory experiences while interacting with the LLM-CA(RQ1)**, and **synthesize design implications for leveraging LLMs' capabilities to meet user perception in their real-world practices(RQ2)**.

The questionnaire results from the study indicate that participants generally have a positive experience using *Mango Mango*. Users' feedback from the interview 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 virtual assistants changed during their interaction with *Mango Mango*. In particular, users shift from perceiving CAs as simply a tool to an assistant that could aid with queries. Following the study, users began to desire virtual assistants that could act as partners, helping them in decision-making processes.

Based on the findings, there is potential for CAs to enhance user expectations toward being partners for users. We discussed design considerations for leveraging LLMs' NLU and NLG capabilities to enhance the effectiveness

---

[1]https://openai.com/

[2]https://www.amazon.com/alexa-skills

and usability of LLM-based CAs in practical applications. These design considerations could potentially cater to users' increasing needs and expectations for future LLM-integrated CAs.

The main contributions of our paper are summarized as follows:

(1) We developed a conversational assistant system that integrates a widely deployed LLM (GPT3.5-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 to meet user expectations on CAs as partners.

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. Throughout this exploration, we acknowledge the challenges faced in the process. 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 [7, 32, 34, 83, 84, 86] 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 [8, 9, 14, 16, 35, 65]. CAs are also used for other scenarios, such as travel [17, 64], music [5], education [22, 28–31, 42, 89, 90], home [10–12, 69], etc., showing promising utilities [40, 72, 77].

However, challenges in developing CAs are identified primarily due to disparities in users' perception of the system's capabilities. Issues like speech detection failure and faulty recognition can occur [57, 63]. The use of heuristics in most existing commercial CAs limits the scope of questions that can be answered. It also constrains the support of basic interaction functionalities such as setting reminders, which can potentially cause users to feel discouraged and lower their expectations of the technology's capabilities [5, 6, 19, 25, 55, 75]. 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 [41, 49].

These 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 [85] and facilitating information retrieval within conversational contexts [52] 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 From LM to LLM

Current language models require substantial amounts of data for training, facing challenges like fine-tuning a system to generate responses with varying tones, such as incorporating emotions or politeness [2]. However, innovative methods and algorithms, such as instructional-finetune [23, 80] and reinforcement learning with human feedback (RLHF) [21, 62] algorithms, have revolutionized the potential of large-language models (LLMs) such as LLaMA [74], FLAN [23, 80], PALM [20], InstructGPT [62], and GPT-4 [61]. Leveraging LLM offers the

advantage of utilizing pre-trained models, reducing the data requirements for effective performance enhancement. These models are fine-tuned on various natural language tasks, enabling them to effortlessly comprehend all instructions and generate high-quality text content [15, 60, 67]. Additionally, LLMs possess an ability to handle lengthy text input (e.g., GPT-4 [61] can take 32, 000 tokens), allowing them to perform tasks that traditional language models cannot handle, such as multi-turn conversations.

As LLM technology advances, researchers are actively exploring its various potential applications [43, 45, 50, 51, 53, 73, 78, 82]. Researchers are particularly interested in harnessing LLM's ability to process inputs through prompt engineering and generate outputs that combine extensive dataset knowledge to make these applications a reality [27]. For instance, these applications could include qualitative analysis with cultural context comprehension [85], connecting LLM to robots for executing complex real-world tasks with task planning [1], co-creation tools for story and sketch generation [24], software engineering tools for code generation [44], and tools for helping mental health awareness [48, 87].

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 [26, 58, 59]. However, individuals new to cooking or attempting to prepare a new recipe often turn to resources like cookbooks and YouTube videos for guidance [46, 54, 79]. At the same time, their hands are occupied during the cooking process, restricting their capability to gather information. Various AI cooking assistants have emerged to address this challenge by using multiple modes of communication, including text, video, and audio, across various devices such as screens, tablets, and computers [18, 68]. For instance, AI-powered cooking assistants like "Cooking Nav" [33], "AskChef" [59], and "MimiCook" [68] 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 of them do not offer hands-free features and are confined to offering guidance based on pre-set recipes. 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 [81, 91]. Our study investigates 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 explain how we developed and designed our LLM-based CA system, including a detailed overview of the system pipeline and prompt design in Section 3.1. In the forthcoming section, we shall elucidate the study protocols and design in detail, specifically regarding Section 3.2. This will include a description of the experimental recipes used in the study and the procedures we followed. Section 3.3 will delineate our approach

toward data collection, encompassing semi-structured interviews, surveys, and system logs. This section will provide an overview of the analysis protocol we will adopt to ensure reliable and valid data collection.

3.1 System Design

In this study, we introduced an LLM-CA system designed to assist users in completing a recipe. We selected Amazon Alexa as our voice-based conversational assistant (CA) because of its flexible functionality and built-in features, particularly the text-to-speech conversion technology. Our platform was developed using the Amazon Alexa Skill Development Kit and Python programming language. 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

Fig. 1. Simplified system diagram of *Mango Mango*. The flow for the system diagram is as follows: (1) Users speak to Alexa as voice input; (2) The text-to-speech process and the transcribed input are saved in the conversational log; (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 and converted into speech to the user to complete the system loop.

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 won't 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. Our *Mango Mango* LLM-based CA is built on the Alexa Skill Kit, in which we have customized the backend LLM support, as discussed in Section 3.1.2. This enables the complete functionalities of the skill with an Amazon Echo smart speaker.

3.1.2 LLM Selection. Our LLM-CA system utilizes OpenAI's GPT-3.5-Turbo LLM in the backend. When choosing an LLM, we take great care to consider various 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.

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.

3.1.3 Prompting Module. After the user's current voice input is correctly captured and recognized by Alexa Skill, it will be converted into text and sent to the back-end prompting module, and this module will organize and re-construct the complete input text and send it to the LLM, GPT-3.5-Turbo in our case, through API to generate a response. Our prompting module is tailored to suit the cooking with recipe scenario and consists of two core parts: Knowledge Resources and Instructions. In Figure 2, we present the complete prompt built based on the salad recipe used in the lab experiment. We will describe each component in detail next.

Knowledge Resources. For the scenario of cooking according to the recipe in our experiment, the Knowledge Resources covers all the necessary information related to the recipe. We further divide it into two major categories: ingredients and steps. Existing work [79] has discovered that people tend to search for recipes on the Internet in real-life scenarios of cooking based on recipes, especially YouTube recipe teaching videos. Therefore, we also chose YouTube cooking teaching videos as recipe source data in our system, which we will elaborate on in Section 3.2 for the user study design.

Specifically, we transcribed the text of the recipe video and then reorganized it according to the recipe steps and ingredients. 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 LLM can more conveniently and accurately locate the sequence of instructions and the details of each item.

It is worth mentioning that the recipes we used in the experiment require additional sauces to be prepared separately during the cooking process. To better distinguish the ingredients of the dish from the ingredients of the sauce, we explored various approaches to presenting the dressing ingredients and eventually decided to compile a list of ingredients for the sauce separately instead of putting all the ingredients together, expecting the LLM to accurately learn from the text and inform the user if being asked for the ingredients needed for the sauce. Based on the fact that our experiment did not include the ingredient preparation part, we omitted the ingredient preparation steps in the recipe.
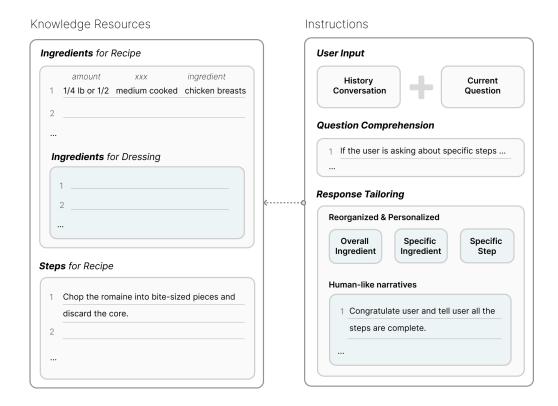
Knowledge Resources

Instructions

Ingredients *for Recipe*

	amount	xxx	ingredient
1	1/4 lb or 1/2	medium cooked	chicken breasts
2			
...			

Ingredients *for Dressing*

1 _____
2 _____
...

Steps *for Recipe*

1 Chop the romaine into bite-sized pieces and discard the core.
2 _____
...

User Input

History Conversation	➕	Current Question

Question Comprehension

1 If the user is asking about specific steps ...
...

Response Tailoring

Reorganized & Personalized

Overall Ingredient	Specific Ingredient	Specific Step

Human-like narratives

1 Congratulate user and tell user all the steps are complete.
...

Fig. 2. Detail Components of *Mango Mango*'s prompting module. The prompting module contains the instructions module (right) and the knowledge resources (left). 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.

Instructions. LLM possesses an exceptional natural language generation ability and has access to an almost boundless wealth of knowledge, enabling it to answer a wide range of questions. However, this also poses the difficulty of limiting the content it produces. To optimize LLM's natural language capabilities for cooking-related inquiries, we have developed a meticulous instruction pipeline in the prompting module. 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 that is more akin to human conversation.

We realize that different questions raised by users when cooking require different levels of detail and methods of response. When a user requests information about necessary ingredients, it can be challenging to provide every single detail, such as names, quantities, and other specifics. Consequently, instead of inundating the user with a plethora of information, it may be more effective to provide them with a complete list of ingredients. If a user wishes to inquire about the specific details of a particular ingredient or step, they are encouraged to ask additional questions. In such situations, the model must be able to provide corresponding and specific responses based on the knowledge available in the knowledge resources module. The model's capabilities should extend beyond merely providing recipes, as it should also be capable of generating appropriate responses to non-recipe-related

inquiries. For instance, users may ask practical questions that are not recipe-specific but rather relate to the cooking process, such as kitchen tool usage or conversion of measurement units.

Therefore, as shown in the right module in Figure 2, we first require the model to understand whether the user's needs are about complete ingredients, specific details or steps, or other knowledge. Based on different user 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 a 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 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.

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.

The 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 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.

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,

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.

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 (CA) **during** the cooking process. Consequently, as previously described, we developed a workflow to translate video content into prompts. 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 [46, 54, 79]. 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 Data collection

3.3.1 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. The interview duration fluctuated according to the interviewees' depth of discussion on various topics. A total of 4 hours and 35 minutes of audio was collected, and all interviews were audio recorded and transcribed. Two authors of this paper conducted an open coding process, identifying themes related to our research questions. These semi-structured interviews provided information for researchers to understand participants' overall experiences with CAs, particularly in the cooking task.

3.3.2 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 current capabilities of the CAs.

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 [93]; the Explainable AI survey (XAI), a six-question scale that evaluates the trustworthiness of explainable AI systems' output from users [38]; the NASA-Task Load Index (NASA-TLX), a six-question scale used to measure participants' subjective workload in six dimensions [36]. 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.

3.3.3 System logs. The system also passively records logs, including users' questions and system replies. These logs were later used for further analysis of user interactions.

3.4 Recruitment process

Participants for this study were recruited via social media platforms and email, where recruitment posters were shared along with a comprehensive description of our research objectives. Recruitment materials included 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 13 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.

3.5 Ethical concern

Our experiment received approval from the university's institutional review board (IRB). 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.

4 RESULTS

In this section, we first reported the demographic information of the participants and presented the survey results. Then, we present the findings for the advantages and challenges encountered when using LLM-based CAs for cooking tasks (RQ1). Finally, we present the design considerations we created based on the results to answer key questions for the future development of the effectiveness and usability of LLM-based CAs. (RQ2)

4.1 Demographic information

We recruited 13 participants in total. However, one data point was removed from the dataset due to a system change, resulting in a final pool of 12 participants.

The sample consisted of 7(58.3%) participants identified as male, 5(41.7%) participants as female, and none as other. Among the sample, 4(33.3%) participants aged between 18 and 24 and 8(66.7%) participants aged between 25 and 34. None of the study participants are allergic to the food used in the study. 8(66.7%) participants cooking in daily basis, while the remaining 4(33.3%) participants cook at least once a week. The majority of participants, 8(66.7%), search for recipes at least once per week, while 2(16.7%) participants search for recipes rarely, one(8.3%) searches at least once per month, and one(8.3%) searches daily.

Among 12 participants, 4(33.3%) use CA daily, 3(25%) use it at least once a week, 3(25%) rarely, and 2(16.7%) use it at least once per month. Participants currently use CA for listening to music (10, 83.3%), checking the weather (9, 75%), and setting the alarm (8, 66.7%) most often, while only one(8.3%) participant uses it for small talk.

For cooking, half of the participants primarily use YouTube to search for recipes, which means the design of our study corresponds to the real scenario of people cooking from a recipe. Among the other half of the participants, 4(33.33%) use recipe-sharing websites, and 2(16.67%) use APP for recipe search. The use of CA for recipe search is infrequent, with 8(66.7%) participants rarely using it and 4(66.7%) never using it for looking up at all. Only one participant uses CA for cooking help at least once per week, while the majority of the participants rarely (7, 58.3%)or never (4, 33.3%) use it for getting cooking help.

The Voice Usability Scale (VUS) evaluates the usability of the voice interface, [93]. In Table 1, the mean values for the first 5 questions related to the positive experience - easy to understand(4.333, SD=0.888), successfully completed task(4.500, SD=0.522), include all the functions needed(4.333, SD=0.985), sufficient response(3.833, SD=1.115) and easy error recovery(4.250, SD=0.866) - are all over 3.8 (with 3.0 indicating 'neutral' and 4.0 indicating 'somewhat agree'). Conversely, for the 6 questions related to the negative experience - irrelevant

Fig. 4. Voice Usability Scale (VUS) results with the bar showing mean and error bar showing standard error, ranging from score 1 ('I disagree strongly') to 5 ('I agree strongly'). The first 5 questions were related to positive experiences, the next 6 were related to negative experiences, and the last was related to the overall experience.

Questions	Mean(SD)
I thought the response from the *Mango Mango* was easy to understand.	4.333(0.888)
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)
I was able to recover easily from errors.	4.250(0.866)
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)
It was easy to lose track of where you were in an interaction with the *Mango Mango*.	2.000(0.853)
I found it difficult to customize the *Mango Mango* according to my needs and preferences.	2.167(1.030)
I found the *Mango Mango* difficult to use.	1.667(0.985)
The *Mango Mango* was unreliable.	1.667(0.985)
Overall, I am satisfied with using the *Mango Mango*.	4.333(0.888)

Table 1. The questions of Voice Usability Scale(VUS) and results in the format of Mean (Standard Deviation)
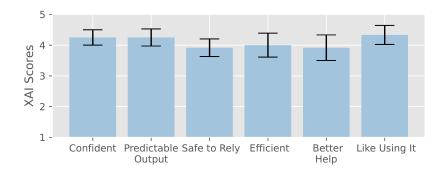
Fig. 5. Explainable AI (XAI) survey results with the bar showing mean and error bar showing standard error, ranging from score 1 ('I disagree strongly') to 5 ('I agree strongly'). All the results of the mean exceeded 3.9, indicating our system was trustworthy.

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

information(2.333, SD=1.371), difficult to understand(2.083, SD=0.515), lost track of interaction(2.000, SD=0.853), difficult to customize(2.167, SD=1.030), difficult to use (1.667, SD=0.985) and unreliable (1.667, SD=0.985) - the mean value are all fall below 2.3 (with 3.0 indicating 'neutral' and 4.0 indicating 'somewhat agree'). Although the results suggest that participants encountered some challenges during the experience, participants overall somewhat agree that they are satisfied with using *Mango Mango* (4.333, SD=0.888) which indicates that our system meets most of our users' needs.

4.2 Survey Result

We requested users to provide subjective evaluations of our system, *Mango Mango*, utilizing four scales to assess various aspects: usability (VUS), trustworthiness (XAI), workload (NASA-TLX), and their perspective of the capability of the current CA (our own questions). In summary, the collective findings across the first three perspectives suggest that participants had successful experiences while interacting with *Mango Mango*. The results from the final questionnaire suggest that people's expectations regarding the capabilities of current CAs have increased after interacting with our system. This aligns with the feedback we gathered during interviews.

Explainable AI (XAI) survey evaluates the trustworthiness of Explainable AI systems' output from the users [38]. Mean values shown in Table 2, related to positive performance, predictability, trustworthiness, efficiency, and superiority over other formats, all exceeded 3.9 (with 3.0 indicating 'neutral' and 4.0 indicating 'somewhat agree'). The overall score for all 12 participants showed our system was trustworthy (mean=4.11, SD=1.026).
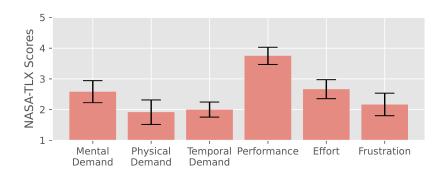
Fig. 6. NASA-TLX results with the bar showing mean and error bar showing standard error, ranging from score 1 ('Very Low') to 5 ('Very High'). Overall workload experienced was below the neutral point.

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

The NASA-TLX is employed to assess users' task workloads [37]. Originally, this scale ranged from 0 to 100 and required participants to rate the significance of six categories: mental demand, physical demand, temporal demand, performance, effort, and frustration. However, due to the exploratory nature of our study and the absence of a control group for comparison, we opted for a five-scale survey (ranging from 1 to 5, with 3 representing a neutral response). To further standardize the scores and accommodate our analysis, we transformed these values to a scale from -50 to 50, resulting in a more manageable and interpretable score range. This approach enables us to make comparisons based on neutrality rather than relying on a control group. Additionally, because of participants' feedback indicating challenges in determining the weight of each category, to alleviate the cognitive burden on participants, we decided to assign equal weight to all categories when calculating the scores, following the approach described by other researchers [36, 76]. In Table 3, all 12 participants yielded a mean workload score of 10.42 (SD=15.442), signifying that the overall workload experienced was below the neutral point. This finding aligns with the insights gathered from participant interviews, as a majority of them reported successful experiences with some encountered challenges, further corroborating the workload assessment.

In addition to the three scales we employed, our study aimed to investigate whether users' perspectives underwent changes following their interaction with *Mango Mango*. To explore this, we included five supplementary questions both before and after the study, as detailed in Table 4. Due to the relatively small sample size and the presence of skewed distributions for some of the questions, we conducted Wilcoxon tests on the results of all five questions. The test results were as follows: Q1: W=0.0, p<0.05; Q2: W=0.0, p<0.05; Q3: W=0.0, p<0.05; Q4: W=3.5, p=0.058; Q5: W=7.5, p=0.135. Our analysis has led us to conclude that users' perspectives on the current

voice assistant differed before and after interacting with *Mango Mango* in terms of its capabilities for engaging in fluent and human-like conversations, remembering past conversations, and allowing follow-up questions. However, there was no significant difference in users' perceptions of its integration into daily life or its ability to collaborate on various tasks. We believe that the possible reason for this difference is the limited time participants had to interact with our system, as well as the controlled and specific lab setting. As a result, participants may have had more confidence in the specific functionalities of *Mango Mango* rather than considering its overall expandability. Despite the acknowledgment of our small sample size (power=0.35, N=12, α =0.05), we believe that our data provides preliminary evidence of a potential trend indicating an increase in users' perspectives
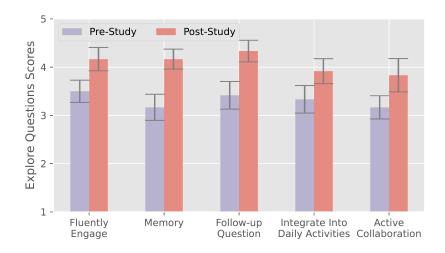

Fig. 7. The result of the five supplementary questions we created with the bar showing mean and error bar showing standard error, ranging from score 1 ('I disagree strongly') to 5 ('I agree strongly'). Overall, the increase result shows a growing in users' perspectives regarding the capabilities of the current CAs.

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

regarding the capabilities of the current CAs following their interaction with our system, specifically in the three aspects described earlier.

In summary, *Mango Mango* demonstrates excellent usability, trustworthiness, and a manageable workload, while also enhancing participants' understanding of current CA capabilities.

4.3 Qualitative Result

To answer RQ1, we categorized users' experiences into two groups: **successful experiences**, where users effectively incorporated the LLM's advanced capabilities into their cooking practices and received valuable assistance, and **unsatisfactory experiences**, where there was a disconnect in users' perceptions of the LLM's capabilities, or their behaviors were negatively impacted because of its advanced capabilities.

4.3.1 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 their usage and experiences they were satisfied with, particularly those related to LLM's 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, P6 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, P9 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. P7 pointed out, "It follows up on your previous question... It sticks to the track, so it's like one continuous flow." Similarly, P6 was impressed by the system's ability to stay on track, stating, "It (*Mango Mango*) can remember 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. P2 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, P11 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 demonstrates 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, P11 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." P11 also highlighted the advantages 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, P5 adopted a similar strategy for task planning, saying, "When I

Theme	Sub-theme	Example
Receive Extensive Information beyond The Recipe	Fundamental Cooking Tips	"Alexa, how to peel avocado?"(P6) "Alexa, tell me that amount of teaspoon if I want one quarter tablespoon"(P2)
	Nutrition Information Related to The Dish	"I asked how many calories are in the in the salad" (P9)
Remember Current Status/History	Current Step	"There was one question I asked which step am I at right now? And he told me on step four." (P6)
Personalize Instructions Based on Context	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" (P2)
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. " (P11) "When I focus on something I just asked you know Mango Mango was the was the next step and then I was cutting stuff and it says the said instruction."(P5)
	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." (P11)
Recover from Errors	Wrong Action	"I added this much salt. Like how much sugar should I add to balance this out?" (P7)
Learn	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."(P11)
A 'Special' System	Congratulation Messages	Q: Do you think this Alexa talks differently? P3: Expressions like enjoy your salad

Table 5. Qualitative code book and description of participants' successful experience and usage with the LLM-based CA.

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 to personalizing the instructions, participants also turned to *Mango Mango* for **assistance when they encountered errors**. For example, P7 explained during the interview that they believed that if they accidentally made a mistake, such as adding too much salt and needing to balance it with sugar, *Mango Mango* could step in to offer assistance. This indicates that users trust the system to provide valid solutions to address cooking incidents that require advanced knowledge and problem-solving capabilities.

Similarly, 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 previously unaware of.

Lastly, an interesting response came from P3 when we asked, "Do you think Alexa talks differently?" They answered, "Expressions like 'enjoy your food'." While this information may not be necessary, it mimics human-like conversation and makes the participant feel like a special CA. This informs us that injecting such human-like redundant information might sometimes change a participant's perception of the system.

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.

4.3.2 *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 LLM's 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, P6 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, P9 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, P8 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. Users often forget to initiate questions with the wake word due to their growing familiarity with natural oral conversation. Participants were

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 our measure the first one. And I forgot about that." (P6)
	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"(P9)
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."(P8)
Forget to Initiate	Redundant Wake Word	"For each time I talk with Alexa, I need to begin with an Alexa. That is very troublesome. For a lot of times, I forgot to say Alexa. If I just want to talk with it, then it will not response with me."(P10)
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."(P5)
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."(P9) "I would like to show me more itself or asked me for confirmation in my case." (P9)
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." (P11)
Uncertain about System Full Capability	Lack of User Guidance on System's Features	"I don't know whether Alexa can help me to control the time or 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." (P10)

Table 6. Qualitative code book and description of participants' unsatisfactory experience and usage with the LLM-based CA.

getting so used to natural interactions that the effort of adhering to system restrictions, such as using the wake word, became more challenging. Similarly, 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 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. P9 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, P9 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 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, P11, 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" 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 P11 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, P10 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. P10 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.

4.4 Design consideration: from tool, assistant to partner (RQ2)

In this section, we will delve into how our insights have informed us about the design of a CA that can harness the full potential of LLM to assist users in completing complex tasks. However, before delving into the design considerations, we recognize that the system's role in users' perception significantly affects the outcome. As a result, we will discuss the design implications of the LLM-based CA in three different phases: the current system, an improved system, and a future system.

*4.4.1 Current system: from successful experience to **design implications**.* Firstly, when examining users' successful experiences while interacting with *Mango Mango*, as described earlier, we realize that these experiences occur

User

Confirmation

Question   *Answer*

LLM-based CA

Question Comprehension

Understanding **narrative**   1

Understanding **goals**   1 2 3

Response Tailoring

Provide **human-like** *narrative*   7 1
Reorganize & **Personalize** *Knowledge*
                                 2 8 9 10 2

General

4   Scenario

Task

Personal Assistant
current

Natural Conversation
improved

Interactive Dialogue
future

Remember & **Retrieval** *knowledge*   5 6

Knowledge Resources

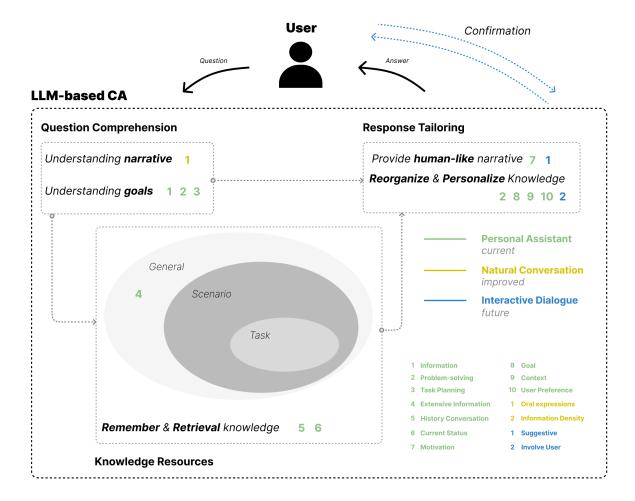| 1 Information | 8 Goal |
2 Problem-solving	9 Context
3 Task Planning	10 User Preference
4 Extensive Information	1 Oral expressions
5 History Conversation	2 Information Density
6 Current Status	1 Suggestive
7 Motivation	2 Involve User

Fig. 8. Design considerations of an LLM-based CA, where we divide the design implications into three phrases: 1) the current system (green), which was established from the successful user experience with our *Mango Mango*, we realize the users' expectations of the system are aligned with its demonstrated capability; 2) an improved system (yellow), where users expect the system to be more oral and be able to comprehend oral expressions better; and 3). a future system (blue), where the LLM-based CA could become a collaborative partner in decision-making and task planning with the users.

when users' expectations of the system's capabilities align with its actual capabilities. Therefore, our initial focus is to establish this aligned perception and then synthesize the underlying design implications of such perception.

We recognized that users view our system as a personal assistant with significantly more knowledge than themselves and advanced abilities to store and retrieve information. This advancement in knowledge is not limited to preset data resources, such as recipe information, that a traditional CA could also possess. It encompasses an extremely broad field of knowledge only LLM can empower the system to tap into. Moreover, users acknowledge its capacity to reorganize and personalize this knowledge even after a brief trial period. This positive feedback largely aligns with the design principles we followed when creating the system and should also be the design consideration for a future LLM-based CA.

*4.4.2 Improved system: be **oral**.* However, in our second analysis, we recognized that not all experiences were successful. Often, this was due to users' expectations of what the system should be capable of not being met by the actual system performance. Many of these unmet needs were related to facilitating a more natural conversation. For example, in understanding user questions, the system needs to treat the conversation as more oral and comprehend oral expressions better. Similarly, in delivering answers, it needs to consider the characteristics of natural conversation, which means providing an appropriate amount of information rather than overwhelming users with everything, which might be more suitable in other contexts, such as reading.

Combining these insights with the system design modules, we structure the takeaways from users' experiences to principles that can inform the design of the system.

An improved LLM-based CA should meet the following **design implications**:

- Precisely comprehend the question.
  - Understand users' goals, including but limited to acquiring information, problem-solving, or task planning.
  - Understand users' narratives, particularly considering oral expressions that may occur in casual conversations.
- Maintain a comprehensive knowledge resource.
  - Have access to extensive information that is not just task-specific but also scenario-specific or even not explicitly related to the task.
  - Be capable of remembering and retrieving knowledge, including but not limited to previous conversations and current status.
- Tailor the response.
  - Capable of reorganizing and personalizing knowledge based on users' goals, preferences, and the current context.
  - Deliver an appropriate amount of information that users can easily comprehend in the context of a natural conversation.
  - Provide human-like narratives, such as oral motivation, in a conversational manner.

*4.4.3 Future system: a **partner** that makes a decision with you.* The previous design considerations ensure that users can benefit from an LLM-based CA system. However, most of these implications were still centered on aligning the system with users' existing mental model of an LLM-based system. Based on the survey and interview results, we have already noticed an enhanced perception of the capabilities of LLM-based CA after interacting with *Mango Mango*. Consequently, we would like to explore participants' suggestions to determine if any of them might necessitate a shift in the system's role in assisting their tasks.

One recurring suggestion from participants was the need for greater integration of user feedback in the decision-making process. As depicted in blue in Figure 8, the system should solicit user confirmation and verification more frequently before providing answers. Instead of issuing commands, it should propose suggestive instructions. When interactive dialogues become more common between users and the system, the LLM-based CA will evolve from being merely an assistant to becoming a collaborative partner in decision-making and task planning with users. This transformation is particularly empowered by the capabilities of LLM, as such conversations are often dynamic and real-time. A traditional rule-based system would not be able to collaborate as closely with users as an LLM-based model could.

In summary, a future LLM-based CA system has the potential to take on a more involved role in complex task completion. To facilitate this, the system should proactively engage users in dialogues.

5 DISCUSSION & LIMITATION

According to our study, the immersive engagement with *Mango Mango* has sparked a profound interest among participants, prompting them to explore LLM-based CA's potential applications in various real-world contexts, transcending the confines of the cooking domain we initially studied.

In this study, we delved into the successful and unsatisfactory experiences of individuals interacting with *Mango Mango*, specifically focusing on the capabilities of LLM. This analysis allowed us to derive design implications for constructing an LLM-based CA system. In addition to the insights discussed in the previous sections, we would also like to offer some inspiration for future work to explore further.

When designing the system, researchers tried to unlock the full potential of LLM to tailor responses to user preferences. However, is this always what users are looking for? An interesting comment from one interviewee mentioned that although they believed everyone's taste buds were different, preferences like these were something they would prefer their family to know rather than the CAs, as they considered it very personal. When considering CAs attempting to provide customized responses, collecting users' personal data is often inevitable, raising concerns about user privacy. Designers should exercise caution regarding the level of personalization offered by CAs, especially considering LLM's extensive capability to reorganize and utilize this information.

Additionally, as mentioned in the earlier section, we are aware that human speech narratives are much more versatile and informal than written text, often featuring incomplete sentences and colloquial expressions. Despite recognizing these gaps, adapting LLMs to the style of casual conversation remains a significant research challenge. Beyond prompt engineering, one potential solution is to explore the training resources of LLMs. We suspect that the current inadequacy of LLMs in understanding colloquial expressions stems from the absence of such data in the model training process. Thus, we propose a future direction for developing LLMs tailored for spoken narratives, involving fine-tuning state-of-the-art LLMs with transcribed colloquial language data. Including more conversational samples from real-world dialogues could be beneficial for LLMs to readily adapt to such scenarios.

Moreover, across these diverse scenarios, *Mango Mango* guides users through tasks and serves as a motivational and instructional tool for enhancing user knowledge. For instance, one participant expressed a desire to store their mother's unique recipes within the CAs so that they could learn from them in the future, highlighting *Mango Mango*'s capacity to customize its responses to new recipes and provide informative and seamless interactions that facilitate user learning.

In the previous section, we discussed how to design the system to align with users' perceptions. Simultaneously, we recognized that users had not fully explored all the functionalities of *Mango Mango*. This points to an issue in effectively communicating the system's full potential to users. LLMs introduce a significant amount of flexibility into human interactions. However, this flexibility also makes it nearly impossible to create user manuals that cover all possible use cases and language choices for communication. 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. Designers must develop mechanisms that make it easier for users to fully appreciate the flexibility introduced by LLM. This could involve the CA encouraging users to explore and discover new functionalities and communication styles or even the system proactively suggesting possible questions based on the previous conversation.

We have already recognized that LLM-based CAs have extensive capabilities in assisting with cooking. However, how do they compare with real human assistance? This comparison becomes crucial for future research because LLM's capabilities have elevated users' perceptions of such systems to a new level. Among our participants, some expressed that *Mango Mango* could replace some of the question-asking efforts they used to direct toward family members who are skilled in cooking. Realizing this potential shift, comparing the different 'affordances' of the two sources of assistance becomes an interesting research topic to explore.

While we used cooking as an example and tried not to over-generalize our findings to other scenarios, we recognized that many of our insights could potentially be applicable elsewhere. 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.

We acknowledge a few limitations in our exploratory lab study. Firstly, our study focused on a salad-making scenario due to safety concerns. This approach falls short of fully simulating real-world cooking scenarios, which often involve different types of equipment, such as ovens and cooktops. However, as described in the methods section, we selected a relatively complex salad recipe that included multiple food preparation and measurement steps. This was done in order to mimic a more comprehensive cooking experience.

Secondly, 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 our results to verify our system's basic usability and general perception. It was not designed to test hypotheses in a rigorous manner. Future work could involve comparative studies to assess the effectiveness of such systems across various dimensions.

6 CONCLUSION

In this study, we investigated 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 that consider the various roles the system could play in assisting users.

REFERENCES

[1] 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]
[2] 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.]).
[3] James Allen. 1995. *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc.
[4] Merav Allouch, Amos Azaria, and Rina Azoulay. 2021. Conversational agents: Goals, technologies, vision and challenges. *Sensors* 21, 24 (2021), 8448.
[5] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. 2019. Music, search, and IoT: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 3 (2019), 1–28.
[6] 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.
[7] 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).
[8] 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.

[9]   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.

[10]   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.

[11]   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

[12]   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

[13]   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.

[14]   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.

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

[16]   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.

[17]   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.

[18]   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.

[19]   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.

[20]   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).

[21]   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).

[22]   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.

[23]   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).

[24]   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.

[25]   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

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

[27]   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]

[28]   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

[29]   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

[30] 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

[31] 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

[32] 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

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

[38] 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).

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

[40] 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

[41] 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).

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

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

[50] 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.

[51] 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).

[52] 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.

[53] 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.

[54] 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.

[55] 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

[56] 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.

[57] 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.

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

[59] 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)*.

[60] 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).

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

[62] 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.

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

[64] 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

[65] 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.

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

[67] 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

[68] 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.

[69] 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

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

[71] 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.

[72] 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

[73] 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.

[74] 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).

[75] 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

[76] Kai Virtanen, Heikki Mansikka, Helmiina Kontio, and Don Harris. 2022. Weight watchers: NASA-TLX weights revisited. *Theoretical Issues in Ergonomics Science* 23, 6 (2022), 725–748.

[77] 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

[78] 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.

[79] 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.

[80] 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).

[81] 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.

[82] 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.

[83] 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

[84] 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

[85] 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.

[86] 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

[87] 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]

[88] 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).

[89] 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

[90] 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

[91] 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.

[92] 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.

[93] 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 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 peper
- 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 7. 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 8. Prompt for *Mango Mango* of Instructions