PrISM-Q&A: Step-Aware Voice Assistant on a Smartwatch Enabled by Multimodal Procedure Tracking and Large Language Models

RIKU ARAKAWA, Carnegie Mellon University, United States
JILL FAIN LEHMAN, Carnegie Mellon University, United States
MAYANK GOEL, Carnegie Mellon University, United States

Voice assistants capable of answering user queries during various physical tasks have shown promise in guiding users through complex procedures. However, users often find it challenging to articulate their queries precisely, especially when unfamiliar with the specific terminologies required for machine-oriented tasks. We introduce *PrISM-Q&A*, a novel question-answering (Q&A) interaction termed *step-aware* Q&A, which enhances the functionality of voice assistants on smartwatches by incorporating Human Activity Recognition (HAR) and providing the system with user context. It continuously monitors user behavior during procedural tasks via audio and motion sensors on the watch and estimates which step the user is performing. When a question is posed, this contextual information is supplied to Large Language Models (LLMs) as part of the context used to generate a response, even in the case of inherently vague questions like "What should I do next with this?" Our studies confirmed that users preferred the convenience of our approach compared to existing voice assistants. Our real-time assistant represents the first Q&A system that provides contextually situated support during tasks without camera use, paving the way for the ubiquitous, intelligent assistant.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; *Interactive systems and tools.*

Additional Key Words and Phrases: context-aware, procedure tracking, task assistance, large language models, question answering

ACM Reference Format:
Riku Arakawa, Jill Fain Lehman, and Mayank Goel. 2024. PrISM-Q&A: Step-Aware Voice Assistant on a Smartwatch Enabled by Multimodal Procedure Tracking and Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 180 (December 2024), 26 pages. https://doi.org/10.1145/3699759

1 Introduction

Tasks or procedures are inherently made up of distinct, sequentially linked steps, and we undertake many such tasks daily, from cooking to using machinery. These tasks are often complex, leading to numerous questions, particularly for those who are untrained or first-time users. Consulting a manual or instruction sheet can help, but research indicates that people often struggle with instruction manuals, thereby attempting tasks without resolving their queries, leading to errors [60]. Such errors can be crucial in many cases, such as using COVID-19 test kits, and research shows that many participants (close to 20% in one study) made critical errors while using these test kits [59]. Consequently, HCI researchers have explored computational methods to address users' queries during task execution to prevent mistakes and enhance user autonomy. Among these, dialogue-based question-answering (referred to as Q&A in the rest of the paper) systems are increasingly popular, where users

Authors' Contact Information: Riku Arakawa, rarakawa@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, United States; Jill Fain Lehman, jfl@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, United States; Mayank Goel, mayankgoel@cmu.edu, Carnegie Mellon University, Pittsburgh, United States.

ACM 2474-9567/2024/12-ART180
https://doi.org/10.1145/3699759

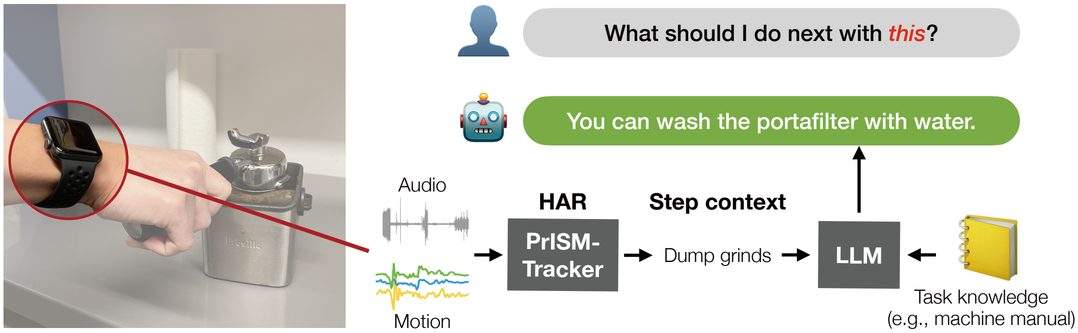


Fig. 1. *PrISM-Q&A* is a novel interaction for question answering (Q&A) for supporting procedural tasks, namely, smartwatch-based *step-aware* Q&A. The system estimates the user context through procedure tracking based on Human Activity Recognition (HAR) and uses the information to augment the Large Language Model (LLM)'s context prior to response. For example, if a user does not know the tool's name and asks a question using a pronoun, the system can resolve the ambiguity based on the step information estimated from the smartwatch's sensors **without using a camera**.

can pose questions and receive answers via speech from hands-free voice assistants without interrupting their primary task [18, 30]. Recently, the advancement of Large Language Models (LLMs) [77, 81] has significantly improved natural language processing to achieve accurate Q&A systems.

However, even voice assistants powered by such models are often inadequate as they rely solely on the information verbalized by the user. Various studies have pointed out that articulating questions to get desired answers can be challenging, and users suffer from nonsensible responses due to assistants' misunderstanding [4, 27, 76]. The lack of context has been identified as the major challenge in supporting users in complex tasks with voice assistants [30, 66]. Lin *et al.* [42] studied users' interactions with voice assistants while they cooked and found users often assumed that the assistant had a contextual understanding and asked questions such as "What is the next step?" and "How long does it last?", which are inherently hard to answer by common voice assistants.

In this paper, we introduce *PrISM-Q&A*, a context-sharing method for dialogue-based Q&A using a smartwatch (Figure 1). The core idea lies in its step awareness, estimating user steps during procedural tasks using multimodal Human Activity Recognition (HAR), information that is then used to enrich the user's verbal questions, helping to clarify their intent and ultimately provide more contextually appropriate responses from a speech-based assistant. For instance, if the system detects that the user has just dumped coffee grounds from a filter when asked, "What should I do with this?", the system can specifically advise, "You can wash the portafilter with water," by resolving the ambiguity in the original question with the estimated step information and retrieving reference from a task-specific knowledge source. This step-aware Q&A interaction can enhance the answer quality while allowing the user to remain somewhat vague in their queries by integrating the user's sensed state in a task while prompting LLMs, thus offering a more intuitive Q&A experience.

Our initial study assessed how users perceived the experience of our proposed step-aware Q&A interaction using a smartwatch. We contrasted our system with existing voice assistants with the existing voice assistants (1) using only what the user asked (like Amazon Alexa) and (2) also using visual information (like visual Q&A models such as Gemini [17] and GPT-4V [54]). Using the Wizard-of-Oz methodology [12], participants ($N = 8$) performed one of two procedural tasks – cooking or latte-making – and responded to survey questions afterward. The study revealed that the step-aware Q&A interaction was preferred for two reasons: the intuitiveness of not having to describe detailed information about the current situation, and the physical comfort of not having to wear a camera.

We then implemented PrISM-Q&A by introducing a step-aware generator with LLMs and a step-aware retriever in conjunction with Retrieval Augmented Generation (RAG) [19] to utilize external knowledge sources, such as machine manuals. We evaluated its performance in three tasks: cooking, latte-making, and skin care. We compared PrISM-Q&A with an LLM-based Q&A pipeline that did not utilize step context and demonstrated the superiority of our step-aware approach, confirmed with multiple state-of-the-art LLMs used as a generator. In the best-performing configuration, when rated on a scale of 1 to 5 (best), the answer quality improved from 2.9 to 4.5, from 3.4 to 4.4, and from 3.9 to 4.5, for the cooking, latte-making, and skin care tasks, respectively. An ablation study suggested the effectiveness of step-aware query translation in the retriever, which increased the contextual precision in the retrieved reference. Moreover, our qualitative analysis illustrated that step information estimated by HAR helped LLMs answer questions related to procedures (*e.g.*, “What is the next step?”) and resolve ambiguities in user queries (*e.g.*, “Where should I place this?”).

Finally, we developed a real-time system using a consumer smartwatch and conducted a user study with the prototype ($N = 10$) in the latte-making task. This study demonstrated that the system helped novice users conduct the task by accurately responding to their questions, which are often ambiguous to answer without the context provided by the activity recognition. Moreover, the participants’ post-hoc comments shed light on important areas to improve the system from the human-AI interaction perspective [1], such as increasing transparency and implementing efficient error recovery. Based on the findings, we implemented an enhanced version of the prototype and discussed future directions to further explore context-aware task assistants.

This paper makes the following contributions:

(1) A novel smartwatch-based Q&A interaction for procedural tasks, *step-aware* Q&A, which offers a preferable user experience to conventional voice assistants by sharing step context between the user and the system.
(2) An approach to achieving the step-aware Q&A that integrates the outputs of multimodal procedure tracking into LLMs by augmenting the user’s question with step context.
(3) A comprehensive study across multiple task datasets to verify the proposed step-aware generator’s and retriever’s effectiveness in increasing factual correctness and overall quality in answers.
(4) A real-time smartwatch system, whose effectiveness in supporting novice users was confirmed through a user study.

We make the implementation and datasets publicly available to support further research in this domain at https://github.com/cmusmashlab/prism. Our dataset will enable researchers to investigate the emerging field of integrating non-visual multimodal sensors with LLMs in an effort to ground AI in the physical world [73].

2 Related Work

We first review the supporting systems for procedural tasks and highlight the need for Q&A systems that answer various user questions. Then, we discuss such dialogue systems, focusing on how the system and user share context to achieve natural and accurate Q&A. Finally, we review existing techniques for Q&A and discuss their potential to support procedural tasks.

2.1 Support Systems for Procedural Tasks

Many HCI systems have been proposed to support users in performing various procedural tasks, including daily activities [23, 26, 52, 67]. For example, Hamada *et al.* [23] introduced *Cooking Navi*, an interface providing multimedia recipe information to aid in cooking. *HoloAssist* [67] is a system where a human observer watches the task performer’s first-person view captured by Augmented Reality (AR) glasses and guides them through verbal instructions. As illustrated by these examples, users often seek information while performing tasks because understanding the task completely by watching instruction videos or manuals is often challenging [42, 61, 79].

Here, voice assistants such as Google Home, Amazon Echo, and Apple Siri are popular methods to provide information so that users do not have to stop the primary task while maintaining agency [27] (*e.g.*, cooking [29], manufacturing [68], medical procedure [13]). Diederich *et al.* [14] conducted an intensive review of research on conversational agents and highlighted the under-explored yet promising role of voice assistants for physical tasks. Qualitative insights were shared by Vtyurina and Fourney [66], who analyzed user questions during a cooking task using the Wizard-of-Oz method and reported the largest class (over 30%) of questions were inquiries about the next steps. In this work, we aim to develop a real-time system designed for procedural tasks that supports the user and answers their queries using the context provided by a common device – a smartwatch.

2.2 Context Awareness in Voice Assistants

The vision of ambient voice assistants has been studied in HCI for a long time and is one of the important subfields of Ubicomp. Multiple studies have concluded that sharing context is crucial in developing dialogue-based human-computer interaction [7, 11, 65], since users often struggle to verbalize queries well enough to get the desired responses [4, 76] and want to rely on indefinite reference like pronouns [21]. For instance, Völkel *et al.* [65] investigated the imaginary "perfect" interaction for voice assistants and suggested that knowledge about the user and the world makes dialogues more effective and natural by creating the impression of shared knowledge and common ground.

In this regard, recent advancements in computer vision and natural language processing now allow dialogue systems to incorporate visual context more easily, with models like Gemini [17] and GPT-4V [54] leading the way. The HCI community has shown that such interactions enhance the naturalness of posing queries about the physical world [22, 24, 40, 46, 47]. For instance, *GazePointAR* [40] is a context-aware voice assistant designed for AR devices. This system utilizes eye gaze and pointing gestures, enabling LLMs to clarify speech queries such as "What is this?", realizing more intuitive Q&A experiences. Such a vision of smart assistants in the physical world has been popularly explored these days, leading to emerging wearable products Ray-Ban Meta Glasses [48]. Therefore, applying these advancements in visual Q&A to procedural task interactions appears promising.

However, the richness of the context shared by cameras also raises concerns about privacy, power consumption, accessibility, and the inconvenience associated with using specific devices that require a camera to maintain a clear view of user actions. The vision-based approaches also often suffer from issues related to the limited field of view, occlusion, and motion blur [51]. Some of these issues around using cameras were highlighted in recent press coverage of now-closed Amazon Go Stores, where a seemingly easy task of monitoring customers required several cameras to get a good view of the users and their actions as well as an army of crowd-workers to label and verify inferences. Moreover, it was perceived as a privacy nightmare [1].

Still, prior research showed that users frequently seek procedural information such as "What is the next step?", "Anything else I haven't done?", and "How long does it last?" [42, 66]. This insight and the challenges of using a camera guided our hypothesis: *It is possible to compensate for the lack of shared visual context with a voice assistant by providing context related to the user's status within the procedure.*

This hypothesis gained further support from various sensing approaches explored in the Ubicomp community, specifically Human Activity Recognition (HAR). For example, smartwatch-based HAR has been widely studied using audio and motion sensors [5, 9, 20, 37–39, 50]. The potential to track user steps using these sensors has been confirmed in various procedural tasks, such as cooking a brownie [63] and fried noodle [36], insulin self-injection [8], and assembly work [72]. Our prior work, *PrISM-Tracker* [3] utilized transition information of procedural tasks for enhanced procedure tracking in latte-making and wound care tasks. This work aims to integrate such multimodal HAR into Q&A systems for procedural tasks as a novel context-sharing approach.

[1] https://www.thedailybeast.com/amazons-just-walk-out-frictionless-checkout-tech-is-a-privacy-nightmare

2.3 Question Answering using LLMs

Recent advancements in Large Language Models (LLMs) have significantly impacted the field of natural language processing, particularly in question-answering (Q&A) tasks [77, 81]. In several studies, LLMs have demonstrated impressive performance in general open-domain Q&A tasks (*e.g.*, Wikipedia) and data-driven response generation [15]. However, research has also shown that LLMs struggle to learn long-tail knowledge [34, 44], resulting in errors [33] that degrade the system performance and fail to meet user expectations. In response, Retrieval Augmented Generation (RAG) is a technique in achieving more reliable Q&A systems [19, 41]. Using RAG, the system refers to a knowledge source in response to a user question, finds relevant information, and generates an answer based on the reference, often using LLMs. Studies have shown that RAG-augmented models can outperform traditional LLMs, especially in domains where precision and factual correctness are paramount [41]. Procedural tasks are typically accompanied by knowledge sources, such as recipes for cooking, instruction sheets for medical kits, and manuals for machinery. Therefore, LLM-based Q&A systems referencing such information when generating answers hold promise. Our work seeks to enhance this approach by integrating sensed user context in the physical world.

3 Research Questions

As outlined in Section 2.2, providing contextual information to Q&A systems is critical, yet current methods mainly rely on users' verbal descriptions or shared visual data. To develop supporting systems for procedural tasks, we introduce a *step-aware* Q&A, where the system uses sensor data to develop an understanding of which step of a procedure the user is performing. This awareness about the step is combined with the fixed description of the procedure itself before it is provided as the context to the language model. The goal of our work is to evaluate the utility of the sensed information as part of the shared context.

Thus, we first aim to understand how the shared context provided by the sensed information influences user interaction with the system:

> **RQ1**: How do users use and perceive smartwatch-based step-aware Q&A compared to conventional voice assistants?

Second, assessing the technical feasibility of step-aware Q&A using a smartwatch is essential. We chose smartwatches for their ubiquity and minimal privacy issues relative to camera-based systems, as well as their ability to monitor various daily activities. Considering recent LLMs' capability in Q&A and HAR studies using a smartwatch, as discussed in Section 2, we consider:

> **RQ2**: Can LLM-Based Q&A pipelines effectively use context estimated from smartwatch's sensor data to achieve step-aware Q&A?

As we aim to assist users in performing procedural tasks more effectively, it is necessary to test the user experience with a real-time prototype, leading to our final research question:

> **RQ3**: Is the real-time step-aware Q&A system helpful in supporting users' needs during procedural tasks?

4 Study 1: Formative Study of Step-Aware Q&A Interactions in Daily Tasks

We first explored how users perceived our proposed step-aware Q&A interaction to address **RQ1**. We employed the Wizard-of-Oz methodology [12], commonly used in dialogue system evaluation, to simulate this new interaction and assess user perceptions.

4.1 Design

This study utilized a within-participant design, comparing three conditions: *voice-only*, *vision-based*, and *sensor-based*. The *voice-only* condition served as a baseline, where the Q&A system only used information verbally provided by the user. The *vision-based* condition incorporated visual data into the Q&A process, an interaction becoming common with emerging smart glasses such as Ray-Ban Meta Glasses [48]. The *sensor-based* condition, our focal condition, leveraged HAR techniques to inform the Q&A system of the user's step within a procedural task using a smartwatch. We assumed that both approaches could track user steps perfectly, which was an unrealistic setting, but for this experiment, we focused on exploring user perception in ideal dialogue scenarios. A more detailed evaluation of response accuracy with actual sensor data will follow in Study 2. The wizard responded to participants' questions according to each condition, which is described in Appendix C.1.

4.2 Task

We used two tasks: cooking and latte-making. The cooking task involved preparing a sunny-side up egg and grilling a sausage, with participants likely asking a variety of questions related to the recipe and cookware use. The latte-making task required using a machine to make a latte, potentially prompting questions about more complex operations detailed in the manual. Each task's detailed procedural steps are outlined in Appendix A. Participants were allowed to modify the order of steps freely.

4.3 Participant

We recruited eight participants (P1–P8, 6 male, 2 female; aged Mean = 39.4, SD = 18.2) via word-of-mouth from our institution and the local community. Half of the participants (P1–P4) were assigned to the cooking task, while the other half undertook the latte-making task (P5–P8). We assessed their familiarity with these tasks by asking how frequently they performed the task, which we used as a measure of their proficiency. They got the cooked meal or cup of latte they made in the session as compensation for their participation.

4.4 Metric

We assessed usability using the System Usability Scale (SUS) [10]. This technology-independent measure is widely used for subjective evaluations of system usability. The SUS consists of 10 items, and participants respond to each item using a 5-point Likert scale where 1 represents 'strongly disagree,' and 5 represents 'strongly agree.' Scores are then calculated on a scale from 0 to 100.

4.5 Procedure

Each participant was initially briefed on the assigned task and the three Q&A conditions they would encounter during the task execution. The task was segmented into three equal parts, each of which was assigned a Q&A condition, with the order in which conditions were experienced randomized. At the start of each segment, we paused the task to inform participants of the upcoming condition. For the *vision-based* condition, participants wore mock AR glasses as part of a simulated setup, while for the *voice-only* and *sensor-based* conditions, they wore a smartwatch. Given the Wizard-of-Oz methodology, these devices were non-functional; however, participants were instructed to imagine these were operational and to internalize their experiences to prepare for a post-task questionnaire. They were also encouraged to ask at least three questions in each segment to facilitate the internalization. Following the task, participants completed the questionnaire, providing ratings for SUS and open-ended feedback about their overall experience, preferred conditions, and the reasons for their preferences.

4.6 Results

4.6.1 Question Types. We noted a pattern in questions participants posed under each experimental condition. In the *voice-only* and *sensor-based* conditions, questions primarily sought factual information likely available in a manual. For instance, during a latte-making task in the *voice-only* condition, a participant asked, "I'm grinding beans. How long should I wait for one-shot beans to be ground?" [P6, experienced user]. Similarly, in the *sensor-based* condition during a cooking task, the question was, "How long should I wait? [while grilling a sausage]" [P2, experienced user]. Conversely, in the *vision-based* condition, participants asked not only factual questions but also those requiring visual information to answer, such as "Is this the right angle to hold the milk jar?" [P8, experienced user]. Notably, in all conditions, the participants often asked about the next step, which is aligned with the findings by Vtyurina and Fourney [66].

Additionally, there was a notable shift in language use. In the *voice-only* condition, participants provided more detailed information to clarify, like, "I have finished grinding beans. What is the next step?" [P8]. This tendency reflects users' efforts to increase the likelihood of receiving the desired response [43, 57]. On the other hand, in the *vision-based* and *sensor-based* conditions, they often used pronouns to refer to objects or actions or omitted the referent, for example, "What should I do next?" [P7, novice user] or "Where should I attach this?" [P7]. Such change is understood given that users desire to communicate to voice assistants using pronouns [21].

From these observations, we hypothesized that although questions demanding visual or world information are still challenging, the *sensor-based* Q&A could facilitate inquiries about factual knowledge by contextualizing the users' state within the procedure. Specifically, we anticipated that the capability might allow it to resolve ambiguous queries like "What should I do next?" or "Where should I attach this?" based on the step context, thus reducing the need for users to elaborate extensively on their situations when posing questions.

4.6.2 User Experience. The SUS scores for the *vision-based* (Mean = 76.3, SD = 13.3) and *sensor-based* (Mean = 81.0, SD = 12.1) conditions were higher than the *voice-only* (Mean = 63.1, SD = 13.4). This improvement was statistically significant, as confirmed by a t-test with Bonferroni correction following a one-way ANOVA. However, there was no significant difference between the *vision-based* and *sensor-based* conditions ($p > 0.05$). By looking at their comments, we found the participants appreciated the context-sharing aspects of the interactions, stating benefits such as, *"Using the Q&A-based task requires a strong learning curve since you have to describe the questions in more detail"*[P8, *voice-only*] [2], *"It was helpful to ask questions about the instruction of the task like what to do next."* [P2, *sensor-based*], and *"It seems very easy to use this kind of system as it is powerful enough to recognize the objects that I am facing and understand my intent."* [P6, *vision-based*].

Moreover, the different modalities and devices used in the two conditions led to varied user experiences. Some participants preferred the convenience of a smartwatch over AR glasses, noting, *"While the glasses would be powerful, I preferred using the smartwatch over AR glasses. Covering my vision is cumbersome"* [P3, experienced user] [3]. Another commented on specific pros and cons: *"From the accuracy of assistance point of view, I would prefer the AR. But putting on a headgear might be a little troublesome compared to Siri."* [P7]. These comments highlighted the trade-offs between device convenience and the types of shared context.

Some participants also elaborated on potential use cases for each context-sharing Q&A interaction. For instance, *"If I want to ask about objects around me, the vision-based Q&A is helpful. At the same time, the sensor-based Q&A seems convenient for questions about the task flow, like asking for instructions based on what I am doing."* [P1, novice user]. This comment suggested that, while both *vision-based* and *sensor-based* interactions could allow questions like "What is the next step?", the smartwatch-based system was preferred for convenience, provided that it could track user steps accurately. Another added, *"The system knowing what I am doing is helpful for me to check what I need to do to finish the task properly. Sometimes, I want to refer to visual information, like to confirm the way I hold*

[2] This notation indicates a response to a question about the specific condition.

[3] This notation indicates a response to a question comparing all conditions.

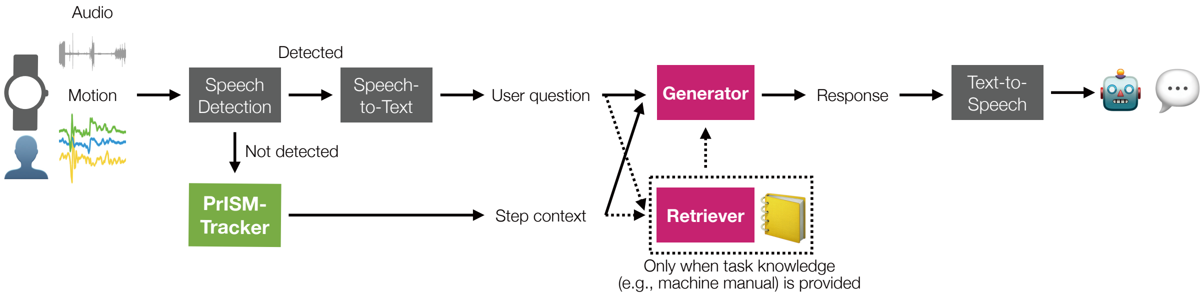


Fig. 2. Overview of PrISM-Q&A's pipeline. The results of PrISM-Tracker [3], multimodal procedure tracking, are used to augment the generator as well as the retriever, which is optionally used to refer to the external knowledge source through the RAG mechanism.

the milk jug is correct." [P5, novice user]. Moreover, one participant suggested a hybrid approach, *"I may rely on the watch first and, if needed, I'd use the camera approach like smartphone."* [P4, experienced user].

4.7 Summary

As an answer to **RQ1**, we concluded that the *sensor-based* Q&A, under the condition of ideal sensing capability, offers a preferable experience for its convenience. Specifically, the asked question types and user comments highlighted its potential to allow users to seek factual information and next steps without fully describing their current status. While incapable of asking questions that require visual information, the handiness of using a smartwatch alone appeared promising to users. Simultaneously, this novel interaction setting (*i.e.*, knowing steps but no access to visual information) affected the language users employed (*e.g.*, pronoun, omission), which would influence the system's Q&A performance. This insight led us directly to **RQ2** concerning the system's ability to accurately respond to user questions posed under the *sensor-based* condition.

5 Proposed Approach for Step-Aware Q&A during Procedural Tasks

This section presents our approach to achieving the step-aware Q&A by connecting multimodal HAR and LLMs.

5.1 Overview

The overview of the proposed approach is shown in Figure 2. First, the sensor data from the smartwatch is processed with PrISM-tracker, introduced in our prior work [3]. This module takes in audio and motion data as well as a graph structure summarizing the transition information, *i.e.*, how long a user is supposed to spend at each step and the transition probability between steps. Sample transition graphs are presented in Figure 6 in Appendix A. The tracker outputs frame-level step prediction, which corresponds with the window size of 0.2 seconds. This information is passed along with the user question to two components of our Q&A pipeline: *step-aware generator* and *step-aware retriever*. The retriever is only used when the system needs to refer to an external knowledge source (*e.g.*, machine manual) to fetch references as a RAG mechanism. The generator is an LLM to synthesize the final response. We describe each component in the following subsections.

5.2 Step-Aware Retriever

A retriever is a common technique used to provide LLMs with reference information based on a vector search using the query embedding and the knowledge source. Query translation is widely explored to rephrase the user query such that their embedding becomes more plausible and diverse to get better references [45]. In this

```
You are an assistant for question answering to support users doing [task name] task. Use the following pieces to answer user questions. If you don't know the answer, just say that you don't know. Use one sentence maximum and keep the answer concise.
User task : [task description]
Users may not follow the step sequence strictly. They ask questions at some point while performing the procedure.
When you refer to a step in your answer, don't mention the step number, as users do not know it. Instead, describe the step content.

As a reference, information on what the user is doing when they ask the question, which is estimated from audio and motion data on their smartwatch, is provided. When you answer the question, imagine the user performing the task according to this information and consider the context of the question. Note that this estimation is not always perfect, and you should not rely on it too much when answering the question.
The user is [probability nuance] at [current step].
If the user is at [current step], the next step is [next step].

Think step by step.
Question: [user question]
Context: [context]
Answer:
```

Fig. 3. Prompt used in the step-aware generator. [context] is used only for the RAG pipeline.

regard, we observed in Study 1 that the participants tended to omit the description of what they were doing in the step-aware interaction. Often, such context is important to identify the relevant document in the knowledge source; for instance, machine manuals are often organized roughly by steps. Therefore, we introduced a query translation based on the HAR results. Specifically, the approach augmented the user question with the estimated step information. For example, if the original question is "What should I do next?" and the system detects the user is washing hands, the translated question is "What should I do next? I'm washing my hands." Notably, we did not assume a particular structure in the knowledge source. If, for instance, the knowledge source has a step-by-step structure that matches with the HAR classes, the retriever could pull up the corresponding documents more accurately than the embedding-based search. Yet, there are many cases where such a structured knowledge source is not available, and PrISM-Q&A was developed to generalize to such scenarios. The translated question was used in the RAG mechanism, for which detailed information is provided in Appendix B.1.

5.3 Step-Aware Generator

A generator is an LLM that synthesizes the final response to the user based on a given prompt. Figure 3 presents the prompt used in our step-aware generator. The [task name] and [task description] are predetermined, and [context] comes from the output of the retriever. Note that if the pipeline does not involve the retrieve, the line with [context] is removed. The [current step] comes from PrISM-Tracker, which is the step having the highest likelihood. The [probability nuance] is based on the confidence level p of the current step, say, washing hands, it would be 'most likely washing hands' ($p \geq 0.85$), 'likely washing hands' ($0.85 > p \geq 0.7$), or 'maybe washing hands' (otherwise). This threshold was determined arbitrarily. Moreover, based on the estimated current step, transition history, and the task's transition graph, it is possible to infer the possible next steps. We provided such information explicitly in LLM's prompt (`[next step]`) as asking about the next steps is a common interaction during procedural tasks, as suggested by prior work [66] and reaffirmed in Study 1.

5.4 Implementation of Real-Time Voice Assistant

We developed a real-time system using an Apple Watch (Series 7), as shown in Figure 2. The system used a laptop (MacBook Pro with 16GB Apple M1 Chip) as the computation server, and we left the self-contained system as future work. In this prototype, the smartwatch plays the role of streaming audio and motion data to the laptop wirelessly, where PrISM-Tracker and the step-aware Q&A pipeline run. The tracker is continuously applied unless a user asks a question. Users can initiate a question with the wake-up word "Hey PrISM," a defacto approach in voice assistants. When a speech is detected, the HAR is stopped so as not to add noise to the tracker, which means the tracker uses the latest tracking output before the user mentions the wake-up keyword. The user question is transcribed using OpenAI Whisper API [53]. The question and the estimated contextual information are then fed into the step-aware Q&A pipeline. The generated answer is converted into audio using OpenAI Text-to-Speech API and played on the laptop, during which the tracker is also stopped.

We did a speed test for each module. The average process time for our step-aware Q&A was 2.8 seconds with RAG and 1.7 seconds without RAG when GPT-4-turbo API was used as an LLM. Additional 1.7 seconds and 1.8 seconds were added for the Whisper and Text-to-Speech API, respectively. Note that the PrISM-Tracker was continuously applied to each frame of 0.2 seconds in parallel. This process was lightweight, taking 0.04 seconds on the laptop.

6 Study 2: Performance Evaluation with Daily-Task Datasets

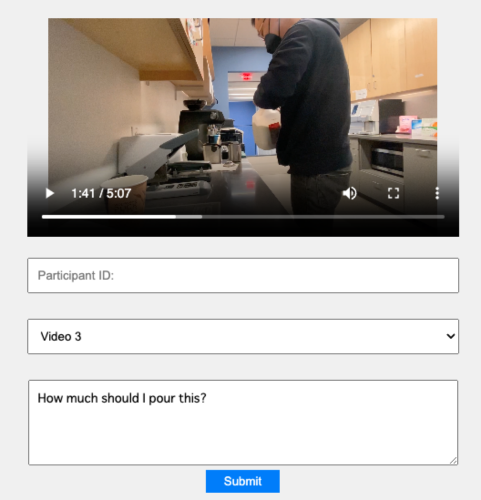
In Study 2, we evaluated the Q&A performance of the proposed approach to investigate **RQ2**. To do this, we first created procedure Q&A datasets in multiple daily tasks where multimodal sensor data from a smartwatch, questions, and their timing were curated. Then, we compared our proposed Q&A approach with baselines.

6.1 Datasets

To test the performance of the proposed step-aware Q&A method, we needed a dataset with questions relevant to specific moments of procedural tasks. Here, to rigorously examine model capability, it is important to collect a large pool of questions users might ask spontaneously. Having participants ask many questions while performing tasks was costly and cognitively highly demanding. Thus, we took an offline question synthesis approach with prerecorded session data; we collected sensor data of users doing the task, and a different set of participants watched task videos later and generated many potential questions as if they were doing the task themselves. This approach, however, might result in a different set of questions from the ones that would occur in actual interactions, and thus, we will evaluate our real-time system's efficacy with novice users later in Study 3.

To explore the performance in various scenarios, we used three procedural tasks: cooking, latte-making, and facial skin care. Cooking is a popular scenario for voice assistant research [30, 66], while latte-making is a complicated machine-use task that involves a long manual document to read. Facial skin care was added to investigate the approach outside the kitchen context and in a situation where wearable cameras (*e.g.*, AR glasses) are not suitable. We used the multimodal sensor dataset introduced by Arakawa *et al.* [3]. The latte-making dataset consisted of 22 sessions done by 15 participants who wore an Apple Watch collecting motion and audio data, including multiple steps of using a latte-maker machine, which has a 28-page long manual book [4]. The cooking dataset consisted of 17 sessions with 8 participants, in which people cooked a sunny-side up and grilled sausage. The skin care dataset consisted of fewer sessions, 5 sessions with 5 participants, in which people washed their faces and did a moisturizing routine. Note that the participants flexibly decided the order of steps to perform

[4] https://assets.breville.com/BES990/BES990_ANZ_IB_G22_FA_LR.pdf



Participant ID:

Video 3

How much should I pour this?

Fig. 4. Interface used in Study 2's data collection to create questions. Users in the videos had worn a smartwatch while doing the task, and the collected sensor data was later synchronized with questions generated by a different set of subjects.

in these procedural tasks, and their familiarity with the tasks differed, resulting in varied behavior in terms of time and transition patterns, as shown in Figure 6 Appendix A. These three tasks encompass a range of procedural complexity, as presented by the different number of branching paths within each task.

The datasets had the video-recorded data synchronized with the watch's sensor data with the participants' permission and appropriate ethics board approval. We used these videos to collect questions that might be asked during the tasks. For each task, we recruited ten volunteers from our institution and the local community through word-of-mouth, allowing participants to overlap across tasks. All participants who made questions for the cooking task had a basic understanding of cooking. Eight of the participants who created questions for the latte-making task had never used the machine before, while two were frequent users. None of the participants who created questions for the skin care task routinely performed it. We had them watch a randomly chosen video of each task individually and come up with potential questions. For this, we developed the simple annotation interface shown in Figure 4. Within the tool, volunteers played and stopped the video at will and posted questions as if they were doing the task. This tool collected their video identifier, posed questions, and corresponding timestamps, later used to synchronize the questions with the watch's sensor data.

Before using the tool, we explained the concept of our step-aware Q&A — the system did not "see" what the users were doing, but it could sense users' actions based on the sensors on the watch. We encouraged participants to ask questions naturally as if they had done the task. We also mentioned that it was possible to ask questions about events that might not be in the video, such as potential error cases (*e.g.*, "The machine is not functioning. What should I do?"). For cooking and latte-making, we asked each volunteer to generate at least five questions for each task video. For skin care, since it is shorter with fewer steps, we asked them to generate at least three questions. As a result, we obtained 54, 68, and 32 questions for the cooking, latte-making, and skin care tasks, respectively.

Two external task designers made a reference answer to each question individually. In doing so, they referred to each instruction (plus the machine manual for the latte-making task) as the knowledge source for each task to develop their answers. Moreover, if a question was related to a transition in the procedure (*e.g.*, "What should

I do next?”), and there could be multiple answers, all possible answers were regarded as the correct answer. After generating each set independently, they discussed and agreed on the final answer, which we treated as the *reference answer* for the later evaluation.

6.2 Compared Pipelines

The three tasks used in Study 2 covered different scenarios: referring to a large external knowledge source (*i.e.*, the machine manual) or not (*i.e.*, the instruction), both of which represented plausible situations where users want to ask questions while performing the task. Whether or not to use the external knowledge source affects the internal process of assistants, as discussed in Section 5. Accordingly, we prepared the *step-aware* pipeline. For the cooking and skin care task, it was a generator only, and for the latte-making task, it was a RAG pipeline where a retriever was incorporated. The sensor data was processed with PrISM-Tracker, and its outputs were utilized within the pipeline. The frame-level tracking accuracy is presented in Figure 7 in Appendix A.

To examine how the addition of step context from HAR helps Q&A performance, we introduced a *baseline* pipeline that does not access sensor data. The baseline retriever used the user question without the translation, while the baseline generator did not incorporate the HAR output in its prompt. Specifically, it used the same prompt in Figure 3 by removing the entire second paragraph starting with “As a reference”. In addition, we introduced a *vanilla* pipeline for the latte-making task where it only accessed the task instruction without the machine manual, serving as a condition to indicate how much existing LLMs can answer user questions without an external knowledge source as a zero-shot. Moreover, we tested our pipelines with multiple state-of-the-art LLMs as generators to examine the generalizability, including GPT-3.5-turbo (`gpt-3.5-turbo-0125`), GPT-4-turbo (`gpt-4-turbo-2024-04-09`), and Llama-3 (8B).

6.3 Metrics

We followed a common approach to evaluate the Q&A performance when the reference answer is available [32]. The metrics that can be automatically computed include answer semantic similarity and factual correctness [16]. Factual correctness quantifies the factual overlap between the generated answer and the reference answer, taking a value of 0 − 1. Answer semantic similarity assesses the semantic resemblance between the generated answer and the reference using a cross-encoder model and cosine similarity, taking a value of 0 − 1. We used `ragas-score` [5] implementation for these two metrics. Each metric was calculated per question, and the scores were averaged for all the questions in each task.

In addition, since these automated evaluations, while correlating, do not perfectly reflect human preference [78], we asked the task designers who made the reference answers to rate the outputs with a score from 1 to 5, where 1 indicates a ‘completely bad answer’, and 5 indicates a ‘completely good answer’. The annotators were not aware of the conditions of each output. They were provided with questions, their reference answers, and the generated responses. We took their average value as the score of human evaluation. To avoid degradation in their annotation quality due to the high workload, we had them annotate the outputs of the pipelines with the best LLM in the automated evaluation.

6.4 Results

6.4.1 Effectiveness of the Step Awareness. The evaluation result in the three datasets is shown in Table 1, summarizing the averaged score of each metric. We found the GPT-4-turbo performed the best among the three LLMs. Moreover, within each LLM, the step-aware pipeline worked outperformed other pipelines in all tasks. When we looked at factual correctness using a paired t-test, there were significant differences in all but one case ($p < 0.05$): Llama-3 (8B) in the latte-making task. Interestingly, when using GPT-3.5-turbo and Llama-3

[5] https://docs.ragas.io/en/latest/concepts/metrics/index.html

Table 1. Performance evaluation in the three procedural tasks (cooking, latte-making, and skin care), comparing three pipelines (vanilla, baseline, and step-aware). Scores range from 0 to 1, and the higher is better. FC and SS stand for factual correctness and semantic similarity metrics, respectively. The step-aware pipeline outperforms other pipelines in both metrics.

	Task	Cooking		Latte-Making		Skin Care	
LLM	Pipeline	FC	SS	FC	SS	FC	SS
GPT-3.5-turbo	Vanilla	-	-	0.32	0.85	-	-
	Baseline	0.25	0.84	0.26	0.85	0.46	0.90
	Step-Aware	0.49	0.89	0.48	0.89	0.50	0.90
GPT-4-turbo	Vanilla	-	-	0.27	0.82	-	-
	Baseline	0.24	0.81	0.36	0.85	0.46	0.88
	Step-Aware	**0.52**	**0.89**	**0.57**	**0.89**	**0.62**	**0.92**
Llama-3 (8B)	Vanilla	-	-	0.24	0.83	-	-
	Baseline	0.24	0.83	0.16	0.78	0.40	0.88
	Step-Aware	0.36	0.86	0.15	0.78	0.43	0.88

Table 2. Results of the human evaluation in the three tasks. Scores range from 1 (‘completely bad answer’) to 5 (‘completely good answer’). The step-aware pipeline outperforms the baseline significantly ($p < 0.05$).

Task	Cooking	Latte-Making	Skin Care
Baseline	2.9 ± 1.7	3.4 ± 1.6	3.9 ± 1.3
Step-Aware	**4.5** ± 0.94	**4.4** ± 1.0	**4.5** ± 1.1

(8B) in the latte-making task, the performance dropped in the baseline RAG pipeline compared to the vanilla (zero-shot) pipeline. This implied the difficulty in referring to external knowledge sources with less powerful LLMs, especially when the retriever is not perfectly fetching the desired knowledge source, as suggested by BehnamGhader *et al.* [6]. Simultaneously, the performance increased in the proposed step-aware pipeline with GPT-3.5-turbo and GPT-4-turbo, which implied the contribution of the enhanced retriever in addition to the generator.

Given this, we obtained the human annotation on the baseline and step-aware pipelines with GPT-4-turbo. As summarized in Table 2, the human evaluation scores were also significantly higher in the step-aware pipeline than in the baseline. The inter-annotator agreement of the human evaluation was 0.67, 0.43, and 0.62 for the cooking, latte-making, and skin care tasks, indicating a good agreement. The high scores suggested the outputs of the proposed approach were mostly perceived as good answers. Note that the number of tokens used in the API call of GPT-4-turbo per question was 887.7 (SD = 10.6), 2453.9 (SD = 409.3), and 769.1 (SD = 10.9) for the cooking, latte-making, and skin care task, respectively in the step-aware pipelines while it was 765.8 (SD = 9.1), 2203.9 (SD = 425.1), and 642.4 (SD = 9.5) in the baseline pipelines.

By looking at the outputs, we found that the proposed step-aware approach could resolve inherently ambiguous user questions, such as “What should I do next?” It was also effective in answering questions seeking factual knowledge; for instance, “Where should I pour milk?” was posed when the user took milk from the fridge. The step-aware pipeline outputted, “Pour the milk into a jug, ensuring it fills to just below the spout position, with enough milk to cover the steam wand seal,” while the baseline outputted, “Pour the milk directly into the espresso in the cup after swirling the jug to polish and reintegrate the texture.” The step-aware pipeline successfully utilized the context information inferred from PrISM-Tracker to guide the user in preparing for the next step of

Table 3. Results of the ablation study in the latte-making Q&A dataset. Scores range from 0 to 1, and the higher is better.

Retriever	Factual Correctness	Semantic Similarity	Context Precision
with query translation	0.57	0.89	0.41
without query translation	0.52	0.89	0.37

steaming milk, while the baseline provided misinformation by not understanding the user's step, which could have led to a negative consequence. Conversely, when HAR prediction was wrong, the step-aware pipeline could not resolve the ambiguity appropriately to questions like "What's next?", which happened when tracking was not reliable (*e.g.*, Steps 7 and 8 in the cooking task, as shown in Figure 7 in Appendix A). This is an important limitation of the current approach, and we discuss ways to address it in Section 8.

6.4.2 Effectiveness of Step-Aware Query Translation in the Retriever. We conducted an ablation study to examine the effect of the step-aware query translation in the retriever using the latte-making task dataset. The results are shown in Table 3. Here, we added a metric – context precision, which evaluates whether the chunks in the retrieved documents are relevant to the reference answer, which was also calculated with the `ragas-score` implementation. The results indicated the improvement of the retriever's performance (from 0.37 to 0.41 in context precision), which led to the enhanced answer of the generator (from 0.52 to 0.57 in factual correctness). The difference was not significant according to the paired t-test. Regarding this, we observed instances where the step awareness in the retriever did not enhance the context precision because the estimated step was wrong, thereby fetching irrelevant information and adding noise to the generator.

6.5 Summary

We demonstrated the proposed step-aware pipeline improved the Q&A performance significantly in all three datasets by resolving question ambiguity and complementing the context with HAR. Based on the high score in the human evaluation, we concluded that the step-aware Q&A using a smartwatch is feasible as an answer to **RQ2**. The effectiveness of the step awareness was confirmed with different state-of-the-art LLMs, which suggested the generalizability of the approach. At the same time, the failure cases were coupled with the error in HAR. This remaining imperfection in our pipeline, as well as the fact that questions used in this study might differ from those that would occur in actual situations, motivated us to explore **RQ3** with a real-time voice assistant.

7 Study 3: User Study

Finally, we evaluated the usability of the step-aware Q&A interaction through a user study to answer **RQ3**.

7.1 Task

We tested our prototype in the latte-making task as an example scenario where users need to know detailed information to use the machine properly. The tracker was trained with the dataset used in Study 2 prior to the experiment, the details of which were presented in Appendix C.2.

7.2 Participant

We invited 10 participants (P1–P10, 8 male, 2 female; aged Mean = 30.9, SD = 14.0) who were unfamiliar with the task through word-of-mouth from our institution and the local community. Five of them (P3–P6, P8) did not know how to make a latte, three (P1, P9, P10) roughly knew the procedure but had never used a machine, and two (P2, P7) had used a different machine to make a latte. One of them (P3) had never used a voice assistant, seven (P1, P2, P4–P6, P8, P9) had used one a few times but not frequently, and two (P7, P10) were frequent users. All of them were right-handed. None of them had participated in either Study 1 or Study 2. They were compensated with $10 USD for their participation.

7.3 Procedure

After consent, the participants read a brief list of steps and watched a video in which another person made a latte with each step labeled. Here, for safety, we explained that the group head and milk wand of the machine would get hot. We also emphasized that users could switch step orders within a reasonable range. In addition, we told them not to do more than one step simultaneously. Then, we explained the prototype system and the concept of the step-aware Q&A. We had a short practice session where the participants familiarized themselves with the system (*i.e.*, wake-up keyword, response latency).

Then, the participant wore the watch on their right wrist [6] and started the task. During the task, the participant could freely ask questions to the system. Here, for simplicity, an experimenter pressed a button on the laptop to initiate the question phase when the user uttered the keyword "Hey, PrISM" before asking a question. Then, 7 seconds were allocated for the user to ask the question. On the rare occasion that the participant got stuck during the task due to nonsensible responses from the system or unexpected events, the experimenter would offer guidance. After the task, the participant filled out the SUS [10] questionnaire and answered open-ended questions: "How did the voice assistant help you complete the task?", "How did you perceive the reliability of the voice assistant?", "How did you perceive the latency in the response?", "Are there any particular questions and responses and your feelings you would like to share with us?", "Do you have any scenarios other than latte-making where the voice assistant can be helpful?", and "Tell us any comments about the experience." The entire study took approximately 30 minutes.

7.4 Results

7.4.1 Question Types. Participants asked 69 questions in total (*i.e.*, Mean = 6.9, SD = 4.6 per session). In six cases, participants either could not complete their questions within the 7 seconds allotted or the speech recognition model failed significantly, resulting in losing the original question's meaning. These instances were excluded from further analysis. The experimenter checked the remaining questions and categorized them according to non-exclusive phenomena. A total of 29 questions involved indefinite reference or omission (*e.g.*, "What's next?", "Should I do manual or automatic?", *etc*), 33 were about steps (*e.g.*, "What should I do first?", "Did I forget any step?", *etc*), 24 were about factual knowledge within each step (*e.g.*, "How much milk should I pour?", "I can't get out the ground bean from the filter", *etc*), and 5 were not answerable with the given context description or manual, requiring either visual or environment-specific information (*e.g.*, "Where are beans?", *etc*).

7.4.2 Accuracy. The experimenter manually classified each generated answer as sensible or nonsensible. As a result, 46 questions were marked as sensible (73.0%). Out of the nonsensible 17 cases, 9 were due to the errors in HAR, 3 failed in the RAG-based Q&A even though the HAR was correct, and 5 were cases where the system did not respond "I don't know" even though the questions were not answerable using the given context and manual.

Out of the nonsensible cases, eight were related to the steps (*e.g.*, "What's next?") and the error was made by the incorrect HAR. In all of these cases, the users re-asked the question by clarifying the intention. For example, P9 asked, "What should I do next?" and the system responded, "Next, you should throw away the towel" and P9 asked again, "What should I do after throwing away the towel?" and got the desired response. In the other nonsensible cases, the participants ignored the response and proceeded by guessing themselves. For instance, P2 wanted to check if the basket was properly attached to the portafilter and asked, "Is this attaching ok?" and the system responded, "Yes, the attachment is okay if you have selected the correct filter basket for the number

[6] This setting matched with the dataset.

of coffees or strength you are making and ensured it is properly inserted into the portafilter." The question essentially required visual information to answer, and the response was not ideal, but P2 continued the task after pushing the basket a few times. These results imply the importance of transparency in the dialogue (*i.e.*, why the assistant said what it said), as users may not have a clear idea about the system's sensing capability.

7.4.3 User Experience. The SUS score indicated good usability (Mean = 79.6, SD = 9.1). To further understand the user experience, we analyzed their answers to the open-ended questions using thematic analysis. This involved manual coding of the responses to identify and categorize recurring themes and patterns within the data.

All participants found the Q&A function helpful, "It answered my questions that I did not come up with in the beginning but later encountered while doing it" [P2]. They appreciated the assistant resolved their step-related questions, "It helped me confirm where to start in the process, what the next step was, and whether I missed any steps. I wasn't familiar with this exact machine so the assistant helped me confirm I am following the correct process" [P7]. Five participants (P3, P5, P7, P8, P10) found the step-awareness particularly helpful, "It was more reliable than I thought as it indeed gave me the answer I was looking for without mentioning the details of what I was doing" [P5].

At the same time, three participants (P6, P7, P9) mentioned the error in the step awareness, "There was one time when the step was not correct based on the assistant's answer, and I knew it was wrong, so I skipped one step. That made me think a little bit that the assistant was not very reliable" [P6]. This comment suggested that offering transparency and error recovery is critical in developing context-aware voice assistants. This is even more remarkable given that the majority of errors (9 out of 17 failure cases) were due to imperfections in the smartwatch-based HAR, as discussed in Section 7.4.2. In their session, P9 understood the error in the step awareness and re-asked questions with less ambiguity, saying, "Sometimes it misunderstood my intention, but it was very easy to know that it was misunderstanding and how it misunderstood. So I could just ask another question to avoid the misunderstanding easily." Regarding transparency, two participants suggested ways to improve it, "I wanted to know if it knew my step correctly before I asked the question. Maybe showing their estimation on the watch's interface will help" [P2] and "For example, if you're doing something something now, next is bra bra bra" [P8]. Increasing the shared context in the dialogue is also suggested to be key by Jaber *et al.* [30]. In addition, there was a certain latency in the prototype mainly due to the heuristic of allocating 7 seconds for the user question as the minimal implementation, and four participants (P1, P2, P4, P6) requested faster interaction, mentioning "It seemed a bit slow while I guess it is faster than I manually read the manual" [P1]. Based on these comments, we developed an enhanced implementation in the next section.

Additionally, the participants suggested further interaction possibilities. P4 mentioned a desire to combine visual information to *see* if the milk amount is appropriate. This is an interesting direction, also suggested in the formative study, as a hybrid of *sensor-based* and *vision-based* systems. Specifically, the assistant would prompt users to use a camera to visually check the task quality only when necessary, based on the recognized context. On the other hand, P6 and P8 mentioned proactive dialogue from the assistant, mentioning, "I think it's good to have the feature that the assistant talks more even if the user does not ask questions" [P8]. This is also an interesting research opportunity to investigate such mixed initiative [25] in task support assistance by involving people with varied proficiency and needs.

7.5 Summary

From the high SUS score and the participants' positive comments, we concluded that the real-time system of PrISM-Q&A was helpful for novice users to conduct the latte-making task as an answer to **RQ3**. The results reconfirmed the benefit of step-awareness in resolving ambiguity, especially for procedure-related questions like "What should I do next?" which constituted roughly half of all questions. Simultaneously, the comments suggested (1) room for improvements in the current prototype in terms of transparency and speed and (2) possibilities for

{{same prompt as in the previous step-aware generator}}
If the question includes ambiguity (*e.g.*, “this” and “next”) and you use the currently estimated step to resolve it, mention the currently estimated step in your answer for transparency, for instance, "After `[current step]`, do `[next step]`" to answer "What is next?". In this case, do not include the step number in your response.
Question: `[user question]`
Context: `[context]`
Answer:

Fig. 5. Prompt used in the generator of the enhanced prototype.

future task-support interactions. In response, we present an enhanced system in Section 8 based on (1) and discuss implications based on (2) in Section 9.3. Additionally, it is important future work to examine the broader effect of the prototype, for instance, on the completion time of the task in comparison with existing voice assistants, by involving more participants.

8 Enhanced Real-Time System Implementation

Based on the feedback received in Study 3, we developed an enhanced prototype (See Video Figure). While user studies with this enhanced prototype have yet to be conducted, we anticipate that these improvements will further enhance the user experience. Here, we describe key implementations.

8.1 Minimized Latency

First, we implemented a real-time voice activity detection to find the end of the user utterance, instead of waiting for a fixed duration. This is a common technique in voice assistants, and our implementation detail is described in Appendix B.2. Next, by using Whisper.cpp [7] locally instead of the Whisper API, the time for Speech-to-Text was reduced to 0.4 seconds (from 1.7 seconds) on average. Also, we replaced Text-to-Speech API with the SAY command in the MacOS, which reduced the network latency. As a result, the latency from the moment a user finishes a question and the moment a user starts hearing the first token of the response is roughly 4 seconds and 3 seconds with and without the RAG mechanism, respectively, while it was about 14 seconds in the previous prototype. While future work, such as adopting faster local LLMs like Phi-3 [49] and real-time speech interaction models like OpenAI’s advanced voice mode, is promising to further reduce the latency, the enhanced prototype offers a reasonable speed.

8.2 Increased Transparency

Secondly, we prompted the model to mention its estimated step context in answering questions that use the information, such as “What is the next step?” and “What should I do with this?” This design was motivated by the fact that most failure cases were caused by HAR errors. As suggested by the participants in Study 3, conveying the estimated step will increase the transparency in the dialogue. This approach allows users to trust the response and correct it if necessary. For instance, instead of “The next step is to take out milk from the fridge”, the response is now “After brewing coffee, take out milk from the fridge.” Our finalized prompt is shown in Figure 5.

9 Discussion

Our studies exploring research questions successfully demonstrated the efficacy of our proposed novel interaction, step-aware Q&A. Lastly, we discuss limitations and future work.

[7] https://github.com/ggerganov/whisper.cpp

9.1 Task Scope

The system's context awareness is based on HAR using a smartwatch. As a result, the applicability of our approach is limited by the capabilities of the HAR. Thanks to the advancements in ubiquitous computing research as discussed in Section 2.2, there are several tasks that can benefit, including cooking, machine use, and medical procedures. While Study 2 demonstrated the effectiveness in three scenarios, it is important to keep exploring the applicability.

In particular, medical procedures are an important domain where LLMs struggle to answer various questions patients may ask due to issues like erroneous responses [70], in which RAG-based approaches will be helpful. We are currently working with skin cancer patients who need to perform post-surgical self-care procedures and evaluating the effectiveness of our assistants [64]. Considering that specific populations, such as the elderly or individuals with limited technological literacy, often face challenges when using voice assistants [31, 35, 58], we believe that PrISM-Q&A's ability to understand user context and complement their queries will be beneficial.

Additionally, given the variety of tasks in our daily lives, it can be helpful if the system detects high-level activity automatically in advance (*e.g.*, cooking, gardening, *etc*) and respond to questions accordingly. Techniques for hierarchical HARs such as proposed by Imoto *et al.* [28] will be beneficial for PrISM-Q&A to support a broader range of scenarios, including multitasking.

9.2 Task Scalability

The current approach requires data collection to train the HAR module. In our study, there were 22, 17, and 5 sessions for the latte-making, cooking, and skin care tasks, respectively. We expect system designers to collect the data before deploying the assistant, and end-users will not need to do this themselves. To facilitate the process, we release our code, which includes the data collection app and a detailed procedure. In the future, it could be possible to use large-scale pre-trained models like *CLAP* [71] to do a zero-shot HAR. Moreover, the assistant should be able to adapt to each end user's behavior and environment after deployment through interactions. Post-deployment learning is actively explored in HAR research [69], and we envision successful human-AI collaboration in this domain.

9.3 Broader Task-Support Interaction Design

In Study 3, we gathered insights for broader interactions beyond the proposed step-aware Q&A, such as proactive intervention to prevent errors (*e.g.*, forgetting a step). *PrISM-Observer* [2] demonstrated such an interaction. It is important to explore the optimal balance between such proactive interactions and the Q&A interaction proposed in this work, as part of the mixed-initiative design [25].

On the other hand, the participants suggested the hybrid use of the *sensor-based* and *vision-based* approaches. While they favored the *sensor-based* approach for its convenience and minimal privacy concerns, there are questions that require a camera for accurate answers, as discussed in Section 7.4.3. We suppose using a camera ad-hoc instead of always-on will be a plausible approach that meets the needs while maintaining the benefits of using a smartwatch. Parikh *et al.* [56] recently proposed a similar approach using low-power active acoustic sensing to guide head-mounted cameras to capture egocentric videos to track eating behaviors. We will explore the versatile potential of task-support assistants in our future work.

9.4 Improving Response Accuracy

While Study 2 confirmed the efficacy of the step-aware pipeline, there are potential ways to improve it by integrating the recent advancement around LLMs and RAG techniques. We believe our datasets will help investigate these potentials to ground LLMs' Q&A capability in the physical world.

Firstly, the current pipeline treats all questions with the same, single pipeline, but the *agent-based approach* [62, 75], where different processes or tools were chosen based on the type of questions, is promising, given there are some patterns in questions participants ask during procedural tasks. For instance, the *ReAct* [75] framework allows LLMs to selectively use reasoning and actions (*e.g.*, referring to knowledge source) to solve various tasks. In our case, since users often ask about the next step during procedural tasks [66], designing a specific tool to answer such questions based on the task's transition graph will be promising, instead of providing the information in the generator's prompt as in our current approach. Such a tool could help reliably answer questions like "Did I miss any steps?" Additionally, as Yang *et al.* [74] did in Q&A within instructional videos, judging the answerability of questions given the context information could mitigate incorrect responses.

Moreover, the retriever's performance is crucial in the RAG mechanism in general, which was also confirmed in Study 2. The current retriever uses the similarity between the embeddings of the user query and the source document. While Study 2 showed efficacy, further improvement is expected by introducing structure into the knowledge source, for example, using knowledge graph [55]. For example, Zhou *et al.* [80] proposed an approach to constructing an open-domain hierarchical knowledge base of procedures. Combining the step-awareness of PrISM-Q&A with such approaches will help deploy assistants to various tasks.

10 Conclusion

We proposed *PrISM-Q&A*, a step-aware question-answering (Q&A) interaction designed for procedural tasks. It enhances the capability of existing voice assistants by enabling them to comprehend user context within a procedure, as inferred from multimodal Human Activity Recognition (HAR) using a smartwatch. Specifically, the output of HAR is integrated into an LLM-based generator and retriever to resolve ambiguity in user queries and synthesize more accurate answers. Our series of studies involving three daily tasks – cooking, latte-making, and skin care – confirmed the improved quality of the generated answers and the real-time system's effectiveness in aiding novice users. This research represents an initial effort to anchor LLMs in physical tasks using sensors and to design supportive interactions. Future work will explore the assistant's effectiveness across diverse scenarios to promote successful human-AI interactions.

ACKNOWLEDGMENTS

This work was partially supported by the NSF Award SCH-2406099 and the Longitude Prize on Dementia. We also acknowledge the support of the Center for Machine Learning and Health at Carnegie Mellon University.

References

[1] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3. https://doi.org/10.1145/3290605.3300233

[2] Riku Arakawa, Hiromu Yakura, and Mayank Goel. 2024. PrISM-Observer: Intervention Agent to Help Users Perform Everyday Procedures Sensed using a Smartwatch. In *UIST '24: The 37th Annual ACM Symposium on User Interface Software and Technology, Pittsburgh, USA, October 13-16, 2024*. ACM, New York, NY, 1–16. https://doi.org/10.1145/3654777.3676350

[3] Riku Arakawa, Hiromu Yakura, Vimal Mollyn, Suzanne Nie, Emma Russell, Dustin P. DeMeo, Haarika A. Reddy, Alexander K. Maytin, Bryan T. Carroll, Jill Fain Lehman, and Mayank Goel. 2022. PrISM-Tracker: A Framework for Multimodal Procedure Tracking Using Wearable Sensors and State Transition Information with User-Driven Handling of Errors and Uncertainty. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4 (2022), 156:1–156:27. https://doi.org/10.1145/3569504

[4] Ahmed Hassan Awadallah, Ranjitha Gurunath Kulkarni, Umut Ozertem, and Rosie Jones. 2015. Characterizing and Predicting Voice Query Reformulation. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*. ACM, New York, NY, 543–552. https://doi.org/10.1145/2806416.2806491

[5] Vincent Becker, Linus Fessler, and Gábor Sörös. 2019. GestEar: combining audio and motion sensing for gesture recognition on smartwatches. In *Proceedings of the 23rd International Symposium on Wearable Computers, UbiComp 2019, London, UK, September 09-13, 2019*. ACM, New York, NY, 10–19. https://doi.org/10.1145/3341163.3347735

[6] Parishad BehnamGhader, Santiago Miret, and Siva Reddy. 2023. Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 15492–15509. https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.1036

[7] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle M. Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3 (2018), 91:1–91:24. https://doi.org/10.1145/3264901

[8] Edgar A. Bernal, Xitong Yang, Qun Li, Jayant Kumar, Sriganesh Madhvanath, Palghat Ramesh, and Raja Bala. 2018. Deep temporal multimodal fusion for medical procedure monitoring using wearable sensors. *IEEE Transactions on Multimedia* 20, 1 (2018), 107–118. https://doi.org/10.1109/TMM.2017.2726187

[9] Sarnab Bhattacharya, Rebecca Adaimi, and Edison Thomaz. 2022. Leveraging Sound and Wrist Motion to Detect Activities of Daily Living with Commodity Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2 (2022), 42:1–42:28. https://doi.org/10.1145/3534582

[10] John Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale. In *Usability Evaluation In Industry*, Patrick W. Jordan, B. Thomas, Ian Lyall McClelland, and Bernard Weerdmeester (Eds.). CRC Press, London, UK, 207–212.

[11] Yuanyuan Chen, Zhengjie Liu, and Juhani Vainio. 2013. Activity-Based Context-Aware Model. In *Design, User Experience, and Usability. Design Philosophy, Methods, and Tools - Second International Conference, DUXU 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 8012)*. Springer, New York, NY, 479–487. https://doi.org/10.1007/978-3-642-39229-0_51

[12] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st International Workshop on Intelligent User Interfaces, IUI 1993, Orlando, Florida, USA, January 4-7, 1993*. ACM, 193–200. https://doi.org/10.1145/169891.169968

[13] Praveen Damacharla, Parashar Dhakal, Sebastian Stumbo, Ahmad Y. Javaid, Subhashini Ganapathy, David A. Malek, Douglas C. Hodge, and Vijay Kumar Devabhaktuni. 2019. Effects of Voice-Based Synthetic Assistant on Performance of Emergency Care Provider in Training. *Int. J. Artif. Intell. Educ.* 29, 1 (2019), 122–143. https://doi.org/10.1007/S40593-018-0166-3

[14] Stephan Diederich, Alfred Benedikt Brendel, Stefan Morana, and Lutz M. Kolbe. 2022. On the Design of and Interaction with Conversational Agents: An Organizing and Assessing Review of Human-Computer Interaction Research. *J. Assoc. Inf. Syst.* 23, 1 (2022), 9. https://aisel.aisnet.org/jais/vol23/iss1/9

[15] Dat Duong and Benjamin D. Solomon. 2023. Analysis of large-language model versus human performance for genetics questions. *European Journal of Human Genetics* 32, 4 (May 2023), 466–468. https://doi.org/10.1038/s41431-023-01396-8

[16] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217* (2023).

[17] Rohan Anil et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *CoRR* abs/2312.11805 (2023). https://doi.org/10.48550/ARXIV.2312.11805

[18] Alexander Frummet, Alessandro Speggiorin, David Elsweiler, Anton Leuski, and Jeff Dalton. 2024. Cooking with Conversation: Enhancing User Engagement and Learning with a Knowledge-Enhancing Assistant. *ACM Transactions on Information Systems* (March 2024). https://doi.org/10.1145/3649500

[19] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *CoRR* abs/2312.10997 (2023). https://doi.org/10.48550/ARXIV.2312.10997

[20] Yu Guan and Thomas Plötz. 2017. Ensembles of Deep LSTM Learners for Activity Recognition using Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2 (2017), 11:1–11:28. https://doi.org/10.1145/3090076

[21] Raymonde Guindon, Kelly Shuldberg, and Joyce Conner. 1987. Grammatical and Ungrammatical Structures in User-Adviser Dialogues= Evidence for Sufficiency of Restricted Languages in Natural Language Interfaces to Advisory Systems. In *25th Annual Meeting of the Association for Computational Linguistics, Stanford University, Stanford, California, USA, July 6-9, 1987*. ACL, 41–44. https://doi.org/10.3115/981175.981181

[22] Nancie Gunson, Daniel Hernández García, Weronika Sieinska, Angus Addlesee, Christian Dondrup, Oliver Lemon, Jose L. Part, and Yanchao Yu. 2022. A Visually-Aware Conversational Robot Receptionist. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2022, Edinburgh, UK, 07-09 September 2022*. Association for Computational Linguistics, 645–648. https://doi.org/10.18653/V1/2022.SIGDIAL-1.61

[23] 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 ACM International Conference on Multimedia, Singapore, November 6-11, 2005*. ACM, 371–374. https://doi.org/10.1145/1101149.1101228

[24] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. 2018. Interactively picking real-world objects with unconstrained spoken language instructions. In *Proceedings of the 2018 IEEE International Conference on Robotics and Automation*. IEEE, New York, NY, 3774–3781. https://doi.org/10.1109/ICRA.2018.8460699

[25] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceeding of the 1999 ACM SIGCHI Conference on Human Factors in Computing Systems*, Marian G. Williams and Mark W. Altom (Eds.). ACM, New York, NY, 159–166. https://doi.org/10.1145/302979.303030

[26] Gaoping Huang, Xun Qian, Tianyi Wang, Fagun Patel, Maitreya Sreeram, Yuanzhi Cao, Karthik Ramani, and Alexander J. Quinn. 2021. AdapTutAR: An adaptive tutoring system for machine tasks in augmented reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 417:1–417:15. https://doi.org/10.1145/3411764.3445283

[27] Alyssa Hwang, Natasha Oza, Chris Callison-Burch, and Andrew Head. 2023. Rewriting the Script: Adapting Text Instructions for Voice Interaction. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference, DIS 2023, Pittsburgh, PA, USA, July 10-14, 2023*. ACM, New York, NY, 2233–2248. https://doi.org/10.1145/3563657.3596059

[28] Keisuke Imoto and Suehiro Shimauchi. 2016. Acoustic Scene Analysis Based on Hierarchical Generative Model of Acoustic Event Sequence. *IEICE Trans. Inf. Syst.* 99-D, 10 (2016), 2539–2549. https://doi.org/10.1587/TRANSINF.2016SLP0004

[29] Takahiko Ito, Shintaro Inuzuka, Yoshiaki Yamada, and Jun Harashima. 2019. Real World Voice Assistant System for Cooking. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*. Association for Computational Linguistics, 508–509. https://doi.org/10.18653/V1/W19-8663

[30] Razan Jaber, Sabrina Zhong, Sanna Kuoppamäki, Aida Hosseini, Iona Gessinger, Duncan P Brumby, Benjamin R. Cowan, and Donald Mcmillan. 2024. Cooking With Agents: Designing Context-aware Voice Interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. https://doi.org/10.1145/3613904.3642183

[31] Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q. Vera Liao, Khai N. Truong, and Shwetak N. Patel. 2018. FarmChat: A Conversational Agent to Answer Farmer Queries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4 (2018), 170:1–170:22. https://doi.org/10.1145/3287048

[32] Jane Huang. 2024. Evaluating Large Language Model (LLM) systems: Metrics, challenges, and best practices. https://medium.com/data-science-at-microsoft/evaluating-llm-systems-metrics-challenges-and-best-practices-664ac25be7e5

[33] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12 (2023), 248:1–248:38. https://doi.org/10.1145/3571730

[34] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large Language Models Struggle to Learn Long-Tail Knowledge. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 15696–15707.

[35] Sunyoung Kim and Abhishek Choudhury. 2021. Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study. *Comput. Hum. Behav.* 124 (2021), 106914. https://doi.org/10.1016/J.CHB.2021.106914

[36] Yusaku Korematsu, Daisuke Saito, and Nobuaki Minematsu. 2019. Cooking state recognition based on acoustic event detection. In *Proceedings of the 11th Workshop on Multimedia for Cooking and Eating Activities*. ACM, New York, NY, 41–44. https://doi.org/10.1145/3326458.3326932

[37] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-play acoustic activity recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, 213–224. https://doi.org/10.1145/3242587.3242609

[38] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. ViBand: High-fidelity bio-acoustic sensing using commodity smartwatch accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, New York, NY, 321–333. https://doi.org/10.1145/2984511.2984582

[39] Chi-Jung Lee, Ruidong Zhang, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda, Oliver Lopez, James Kim, Sicheng Yin, Boao Dong, Ke Li, Mose Sakashita, François Guimbretière, and Cheng Zhang. 2024. EchoWrist: Continuous Hand Pose Tracking and Hand-Object Interaction Recognition Using Low-Power Active Acoustic Sensing On a Wristband. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*. ACM, 403:1–403:21. https://doi.org/10.1145/3613904.3642910

[40] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY.

[41] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[42] Georgianna Lin, Jin Yi Li, Afsaneh Fazly, Vladimir Pavlovic, and Khai N. Truong. 2023. Identifying Multimodal Context Awareness Requirements for Supporting User Interaction with Procedural Videos. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*. ACM, 761:1–761:17. https://doi.org/10.1145/3544548.3581006

[43] 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, CA, USA, May 7-12, 2016*. ACM, 5286–5297. https://doi.org/10.1145/2858036.2858288

[44] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 9802–9822. https://doi.org/10.18653/V1/2023.ACL-LONG.546

[45] Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-Augmented Retrieval for Open-Domain Question Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Association for Computational Linguistics, 4089–4100. https://doi.org/10.18653/V1/2021.ACL-LONG.316

[46] Fabrice Matulic, Riku Arakawa, Brian K. Vogel, and Daniel Vogel. 2020. PenSight: Enhanced Interaction with a Pen-Top Camera. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*. ACM, 1–14. https://doi.org/10.1145/3313831.3376147

[47] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing Mobile Voice Assistants with WorldGaze. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*. ACM, 1–10. https://doi.org/10.1145/3313831.3376479

[48] Meta Platforms. 2023. Smart glasses for living all in. https://www.meta.com/smart-glasses/

[49] Microsoft. 2024. Tiny but mighty: The Phi-3 small language models with big potential. https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/

[50] Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. 2022. SAMoSA: Sensing Activities with Motion and Subsampled Audio. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3 (2022), 132:1–132:19. https://doi.org/10.1145/3550284

[51] Thi-Hoa-Cuc Nguyen, Jean-Christophe Nebel, and Francisco Flórez-Revuelta. 2016. Recognition of Activities of Daily Living with Egocentric Vision: A Review. *Sensors* 16, 1 (2016), 72. https://doi.org/10.3390/S16010072

[52] Jennifer Ockerman and Amy Pritchett. 2000. A review and reappraisal of task guidance: Aiding workers in procedure following. *International Journal of Cognitive Ergonomics* 4, 3 (2000), 191–212. https://doi.org/10.1207/s15327566ijce0403_2

[53] OpenAI. 2022. Introducing Whisper. https://openai.com/research/whisper

[54] OpenAI. 2023. GPT-4 Technical Report. *CoRR* abs/2303.08774 (2023). https://doi.org/10.48550/ARXIV.2303.08774 arXiv:2303.08774

[55] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *CoRR* abs/2306.08302 (2023). https://doi.org/10.48550/ARXIV.2306.08302

[56] Vineet Parikh, Saif Mahmud, Devansh Agarwal, Ke Li, François Guimbretière, and Cheng Zhang. 2024. EchoGuide: Active Acoustic Guidance for LLM-Based Eating Event Analysis from Egocentric Videos. *CoRR* abs/2406.10750 (2024). https://doi.org/10.48550/ARXIV.2406.10750

[57] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*. ACM, 640. https://doi.org/10.1145/3173574.3174214

[58] Alisha Pradhan, Leah Findlater, and Amanda Lazar. 2019. "Phantom Friend" or "Just a Box with Information": Personification and Ontological Categorization of Smart Speaker-based Voice Assistants by Older Adults. *Proc. ACM Hum. Comput. Interact.* 3, CSCW (2019), 214:1–214:21. https://doi.org/10.1145/3359316

[59] Meghana Ratna Pydi, Petra Stankard, Neha Parikh, Purnima Ranawat, Ravneet Kaur, AG Shankar, Angela Chaudhuri, Sonjelle Shilton, Aditi Srinivasan, Joyita Chowdhury, and Elena Ivanova Reipold. 2023. Assessment of the Usability of SARS-CoV-2 Self Tests in a Peer-Assisted Model among Factory Workers in Bengaluru, India. (Nov. 2023). https://doi.org/10.1101/2023.11.20.23298784

[60] A. RAOUF and S. ARORA. 1980. Effect of informational load, index of difficulty direction and plane angles of discrete moves in a combined manual and decision task. *International Journal of Production Research* 18, 1 (Jan. 1980), 117–128. https://doi.org/10.1080/00207548008919653

[61] Jorge Rodrguez, Teresa Gutirrez, Emilio J., Sara Casado, and Iker Aguinag. 2012. *Training of Procedural Tasks Through the Use of Virtual Reality and Direct Aids*. InTech. https://doi.org/10.5772/36650

[62] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

[63] Ekaterina H. Spriggs, Fernando De la Torre, and Martial Hebert. 2009. Temporal segmentation and activity classification from first-person sensing. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, 17–24. https://doi.org/10.1109/CVPRW.2009.5204354

[64] Annalise Vaccarello, Alexander K. Maytin, Yash Kumar, Toluwalashe Onamusi, Haarika A. Reddy, Mayank Goel, Riku Arakawa, Jill Fain Lehman, and Bryan T. Carroll. 2024. Barriers to use of digital assistance for postoperative wound care: a single-center survey of dermatologic surgery patients. *Archives of Dermatological Research* 316, 7 (June 2024). https://doi.org/10.1007/s00403-024-03025-w

[65] 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 *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*. ACM, 254:1–254:15. https://doi.org/10.1145/3411764.3445536

[66] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*. ACM, 208. https://doi.org/10.1145/3173574.3173782

[67] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. 2023. HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 20213–20224. https://doi.org/10.1109/ICCV51070.2023.01854

[68] Stefan Wellsandta, Zoltan Rusak, Santiago Ruiz Arenas, Doris Aschenbrenner, Karl A. Hribernik, and Klaus-Dieter Thoben. 2020. Concept of a Voice-Enabled Digital Assistant for Predictive Maintenance in Manufacturing. *SSRN Electronic Journal* (2020). https://doi.org/10.2139/ssrn.3718008

[69] Jason Wu, Chris Harrison, Jeffrey P. Bigham, and Gierad Laput. 2020. Automated Class Discovery and One-Shot Interactions for Acoustic Activity Recognition. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*. ACM, 1–14. https://doi.org/10.1145/3313831.3376875

[70] Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi Riantawan, Daniel E. Ho, and James Zou. 2024. How well do LLMs cite relevant medical references? An evaluation framework and analyses. *CoRR* abs/2402.02008 (2024). https://doi.org/10.48550/ARXIV.2402.02008

[71] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095969

[72] Qingxin Xia, Atsushi Wada, Joseph Korpela, Takuya Maekawa, and Yasuo Namioka. 2019. Unsupervised factory activity recognition with wearable sensors using process instruction information. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 60:1–60:23. https://doi.org/10.1145/3328931

[73] Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. 2024. Penetrative AI: Making LLMs Comprehend the Physical World. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications, HOTMOBILE 2024, San Diego, CA, USA, February 28-29, 2024*. ACM, New York, NY, 1–7. https://doi.org/10.1145/3638550.3641130

[74] Saelyne Yang, Sunghyun Park, Yunseok Jang, and Moontae Lee. 2024. YTCommentQA: Video Question Answerability in Instructional Videos. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*. AAAI Press, 19359–19367. https://doi.org/10.1609/AAAI.V38I17.29906

[75] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

[76] Ye Yuan, Stryker Thompson, Kathleen Watson, Alice Chase, Ashwin Senthilkumar, A. J. Bernheim Brush, and Svetlana Yarosh. 2019. Speech interface reformulations and voice assistant personification preferences of children and parents. *Int. J. Child Comput. Interact.* 21 (2019), 77–88. https://doi.org/10.1016/J.IJCCI.2019.04.005

[77] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. *CoRR* abs/2303.18223 (2023). https://doi.org/10.48550/ARXIV.2303.18223

[78] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

[79] Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. 2023. Procedure-Aware Pretraining for Instructional Video Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 10727–10738. https://doi.org/10.1109/CVPR52729.2023.01033

[80] Shuyan Zhou, Li Zhang, Yue Yang, Qing Lyu, Pengcheng Yin, Chris Callison-Burch, and Graham Neubig. 2022. Show Me More Details: Discovering Hierarchies of Procedures from Semi-structured Web Data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*. Association for Computational Linguistics, 2998–3012. https://doi.org/10.18653/V1/2022.ACL-LONG.214

[81] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering. *CoRR* abs/2101.00774 (2021).

Appendix A Details of the Used Procedural Tasks

Our studies involved three procedural tasks: cooking, latte-making, and skin care. The three procedural tasks used in this work are shown in Figure 6.

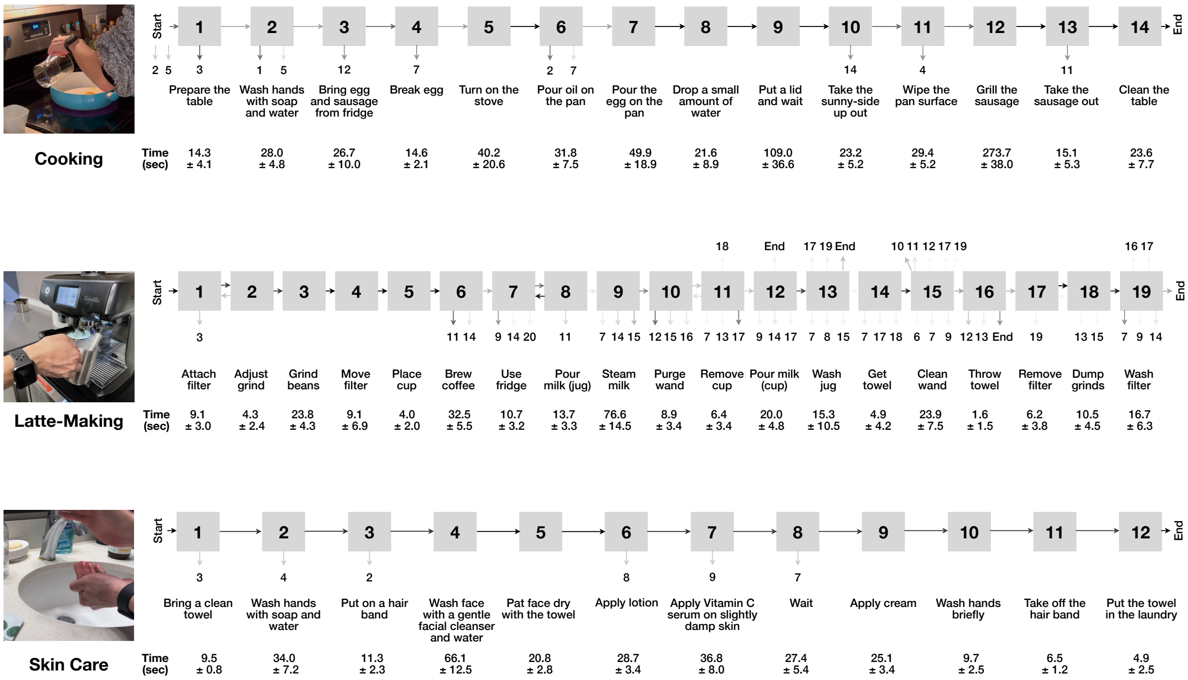

Cooking
Start
End
1 Prepare the table 14.3 ± 4.1
2 Wash hands with soap and water 28.0 ± 4.8
3 Bring egg and sausage from fridge 26.7 ± 10.0
4 Break egg 14.6 ± 2.1
5 Turn on the stove 40.2 ± 20.6
6 Pour oil on the pan 31.8 ± 7.5
7 Pour the egg on the pan 49.9 ± 18.9
8 Drop a small amount of water 21.6 ± 8.9
9 Put a lid and wait 109.0 ± 36.6
10 Take the sunny-side up out 23.2 ± 5.2
11 Wipe the pan surface 29.4 ± 5.2
12 Grill the sausage 273.7 ± 38.0
13 Take the sausage out 15.1 ± 5.3
14 Clean the table 23.6 ± 7.7
2 5
3
1 5
12
7
2 7
14
4
11
Time (sec)
Latte-Making
Start
End
1 Attach filter 9.1 ± 3.0
2 Adjust grind 4.3 ± 2.4
3 Grind beans 23.8 ± 4.3
4 Move filter 9.1 ± 6.9
5 Place cup 4.0 ± 2.0
6 Brew coffee 32.5 ± 5.5
7 Use fridge 10.7 ± 3.2
8 Pour milk (jug) 13.7 ± 3.3
9 Steam milk 76.6 ± 14.5
10 Purge wand 8.9 ± 3.4
11 Remove cup 6.4 ± 3.4
12 Pour milk (cup) 20.0 ± 4.8
13 Wash jug 15.3 ± 10.5
14 Get towel 4.9 ± 4.2
15 Clean wand 23.9 ± 7.5
16 Throw towel 1.6 ± 1.5
17 Remove filter 6.2 ± 3.8
18 Dump grinds 10.5 ± 4.5
19 Wash filter 16.7 ± 6.3
18
End
17 19 End
10 11 12 17 19
16 17
3
11 14
9 14 20
11
7 14 15
12 15 16
7 13 17
9 14 17
7 8 15
7 17 18
6 7 9
12 13 End
19
13 15
7 9 14
Time (sec)
Skin Care
Start
End
1 Bring a clean towel 9.5 ± 0.8
2 Wash hands with soap and water 34.0 ± 7.2
3 Put on a hair band 11.3 ± 2.3
4 Wash face with a gentle facial cleanser and water 66.1 ± 12.5
5 Pat face dry with the towel 20.8 ± 2.8
6 Apply lotion 28.7 ± 3.4
7 Apply Vitamin C serum on slightly damp skin 36.8 ± 8.0
8 Wait 27.4 ± 5.4
9 Apply cream 25.1 ± 3.4
10 Wash hands briefly 9.7 ± 2.5
11 Take off the hair band 6.5 ± 1.2
12 Put the towel in the laundry 4.9 ± 2.5
3
4
2
8
9
7
Time (sec)

Fig. 6. Transition graph of the three procedural tasks used in this work: cooking, latte-making, and skin care. Each task has a different level of graph complexity, *i.e.*, the number of branches. The opacity of the arrows represents the probability of the transition. In other words, the sum of the transitions of arrows from a single step is 1.0.

We used PrISM-Tracker [3] as a multimodal HAR module that provides step information to the Q&A module. The tracker uses Viterbi correction with a transition graph as a post-process to a frame-level HAR. Figure 7 shows the frame-level HAR confusion matrix with and without the Viterbi correction. One frame corresponds to 0.2 seconds.

Appendix B Details of Implementation

B.1 Retrieval Augmented Generation (RAG)

In the latte-making task, we employed the retrieval augmented generation (RAG) framework for the system to refer to knowledge in a long manual [8]. To implement this, we first parsed text data in the manual document using `PyPDFLoader` and divided text into smaller chunks by recursively splitting it based on characters. We used 1000 tokens as the chunk size. Then we constructed a vector store using `Chroma` by using OpenAI's text embedding (`text-embedding-3-large`). The retriever takes in the contextually translated query as described in Section 5.2, embeds it with the same embedding model, and searches the vector store by similarity. The text of the top one

[8] https://assets.breville.com/BES990/BES990_ANZ_IB_G22_FA_LR.pdf

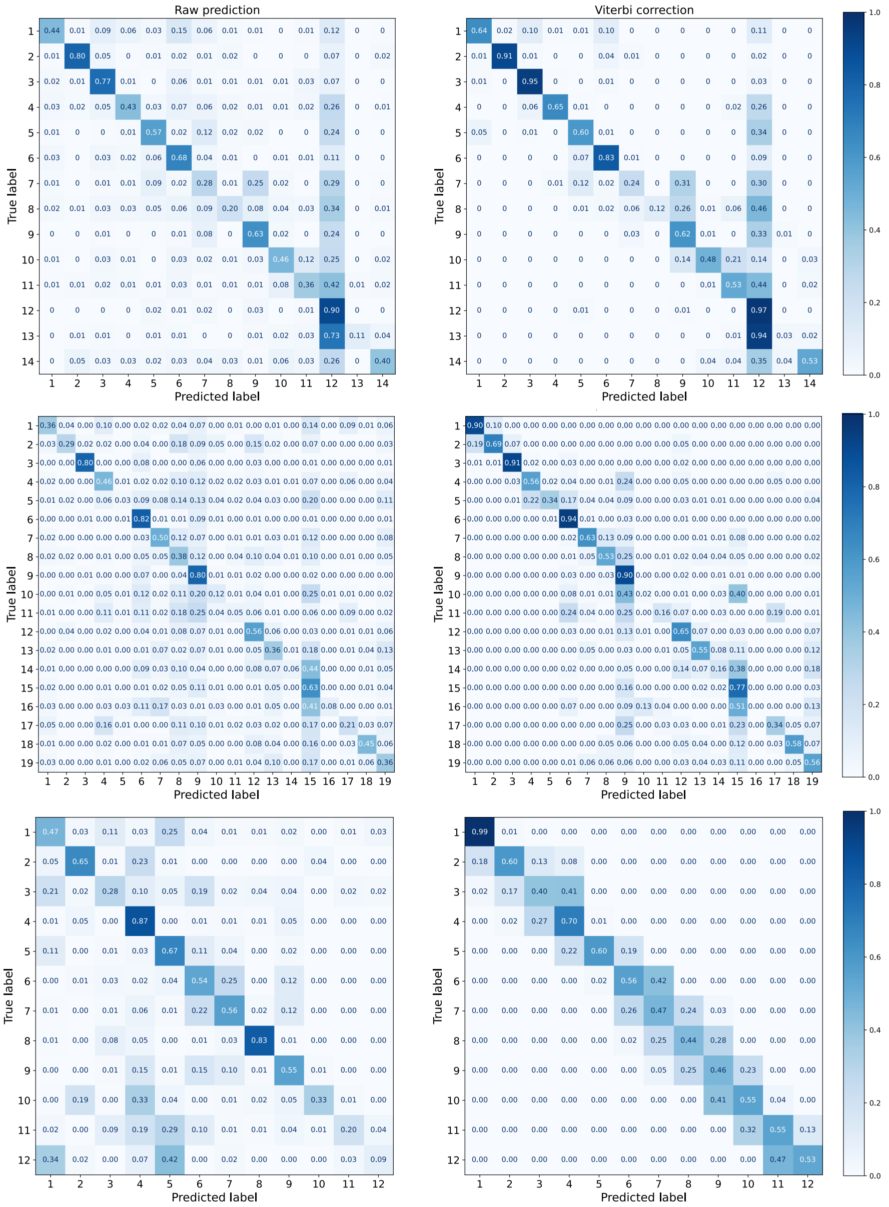


Fig. 7. Frame-level confusion matrix on the three tasks. (left) raw HAR (right) PrISM-Tracker [3]. (from top to bottom) cooking, latte-making, and skin care.

chunk is provided to the generator as `[context]` in Figure 3. We used the `LangChain` framework [9] to implement the above.

[9] https://www.langchain.com/

B.2 Real-Time Voice Activity Detection

The implemented Voice Activity Detection in Section 8.1 employed a combination of frequency-based filtering and amplitude-based thresholding to distinguish speech from non-speech segments in an audio signal. In the frequency-based filtering stage, the audio signal is processed using a 4th-order Butterworth filter with a passband from 300 Hz to 1500 Hz. This range captures the majority of the speech signal while attenuating lower and higher-frequency noises in our test. In the amplitude-based thresholding stage, the filtered audio signal is divided into short, overlapping frames of 25 ms with a 10 ms overlap. For each frame, the short-term energy is computed, and an amplitude threshold is then applied: if the average energy in a frame exceeds this threshold, the frame is classified as containing speech; otherwise, it is classified as non-speech. We regarded consecutive 1.0 seconds of the non-speech segments as the end moment of the user speech to start running the Q&A pipeline.

Appendix C Details of Study

C.1 Wizard Script in Study 1

To answer **RQ1**, we analyzed differences in user experience across three conditions: *voice-only*, *vision-based*, and *sensor-based*. Participants were prompted to ask questions under each condition. The experimenter, acting as a wizard, responded according to the stipulated conditions and said, "I don't know. Please elaborate more," when the question was too ambiguous to answer under each condition. We acknowledge potential discrepancies between the wizard's responses and those an actual system might generate (*e.g.*, *vision-based* system required user actions to happen within the field of camera view), but the primary goal of Study 1 was to gauge user perceptions of the different Q&A interactions rather than the accuracy of the responses, a point emphasized in the participant instructions. The wizard used the following rules based on our initial classification of question types.

C.1.1 Questions Requiring the Current Step Information. If the question required the wizard to know the user's current step, like, "What should I do next?", it responded, "I don't know", in the *voice-only* condition. In contrast, in the *vision-based* or *sensor-based* condition, the wizard could provide the next step based on the user's real-time step.

C.1.2 Questions Requiring Visual Information. If the question required visual information to answer, like, "What is in front of me?" or "Is the way I am holding the filter right now appropriate?", only the *vision-based* wizard gave the correct answer.

C.1.3 Questions Containing Ambiguity in the Language. If the question was ambiguous but could be completed by using visual information, like, "Where should I attach this?", the *vision-based* wizard could give the correct answer. Also, if the ambiguity could be complemented by accessing the current step information, then the *sensor-based* wizard also gave the correct answer, such as, "How much should I pour?" [Step 8 in the latte-making task – pouring milk].

C.2 Procedure Tracking Module in Study 3

PrISM-Tracker uses a transition graph (See Figure 6) as a post-process to frame-by-frame HAR. The graph includes the probability between steps and how long a user typically spends at each step. While their latte-making dataset included participants with various proficiency (seven regular users and eight first-time users), our Study 3 focused on novice users. In our pilot study, we noticed that these users spent a longer time at each step due to unfamiliarity. Therefore, we used a subset of their dataset (the eight first-time users) to create the transition graph, while we used all participants' data to train the frame-by-frame HAR model.