Cooking with Conversation: Enhancing User Engagement and Learning with a Knowledge-enhancing Assistant

ALEXANDER FRUMMET, University of Regensburg, Regensburg, Germany
ALESSANDRO SPEGGIORIN, University of Glasgow, Glasgow, United Kingdom
DAVID ELSWEILER, University of Regensburg, Regensburg, Germany
ANTON LEUSKI, University of Southern California, Los Angeles, USA
JEFF DALTON, University of Edinburgh, Edinburgh, United Kingdom

We present two empirical studies to investigate users' expectations and behaviours when using digital assistants, such as Alexa and Google Home, in a kitchen context: First, a survey (N = 200) queries participants on their expectations for the kinds of information that such systems should be able to provide. While consensus exists on expecting information about cooking steps and processes, younger participants who enjoy cooking express a higher likelihood of expecting details on food history or the science of cooking. In a follow-up Wizard-of-Oz study (N = 48), users were guided through the steps of a recipe either by an *active* wizard that alerted participants to information it could provide or a *passive* wizard who only answered questions that were provided by the user. The *active* policy led to almost double the number of conversational utterances and 1.5 times more knowledge-related user questions compared to the *passive* policy. Also, it resulted in 1.7 times more knowledge communicated than the *passive* policy. We discuss the findings in the context of related work and reveal implications for the design and use of such assistants for cooking and other purposes such as DIY and craft tasks, as well as the lessons we learned for evaluating such systems.

CCS Concepts: • **Information systems** → *Search interfaces*; *Collaborative search*; • **Human-centered computing** → *Empirical studies in interaction design;*

Additional Key Words and Phrases: Conversational agents, interactive search, wizard-of-oz, conversational search

ACM Reference Format:
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 Trans. Inf. Syst.* 42, 5, Article 122 (April 2024), 29 pages. https://doi.org/10.1145/3649500

Anton Leuski's work was supported in part by the U.S. Army; statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred. Jeff Dalton and Alessandro Speggiorin's work was supported by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/V025708/1 and a 2021 Alexa Prize grant from Amazon.
Authors' addresses: A. Frummet and D. Elsweiler, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Bavaria, Germany; e-mails: alexander.frummet@ur.de, david.elsweiler@ur.de; A. Speggiorin, University of Glasgow, 18 Lilybank Gardens, Glasgow G12 8RZ, Scotland, United Kingdom; e-mail: alessandro@dcs.gla.ac.uk; A. Leuski, University of Southern California, Institute for Creative Technologies, 12015 Waterfron Drive, Playa Vista, LA, CA, USA; e-mail: leuski@ict.usc.edu; J. Dalton, University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB, Scotland, United Kingdom; e-mail: jeff.dalton@ed.ac.uk.

ACM 1046-8188/2024/04-ART122
https://doi.org/10.1145/3649500

1 INTRODUCTION

The growing usage of intelligent speakers and smart home devices, facilitated by digital assistants such as Alexa, Siri, and Google Assistant, allows for voice-activated and natural language-based interactions to complete various tasks. Beyond simple, smart device control, a growing use case is assisting people with cooking and food preparation tasks. It is increasingly popular in the research literature (e.g., References [4, 19, 30, 48]) as well as one of the core domains in the recent Amazon Alexa Taskbot Challenge. Despite promising early research efforts in this area that simulate user interaction [4, 19, 45] and developed prototypes [30], we still know little about what people need and want from an assistant that provides support to people while cooking.

This work adds to understanding about the information needs that occur in such contexts, how these can be resolved via question answering, as well as how users interact with systems to achieve their kitchen-oriented goals. The presented research addresses the following two research questions:

— RQ1: What information interactions do users expect from a digital cooking assistant?
— RQ2: How does the interaction policy of the assistant influence the kinds of questions users ask and the amount of knowledge communicated?

We address RQ1 using a user survey to ask about desired types of information interactions participants expect from an assistant, for example, to be able to discuss the history, science, and rationales of key parts of the recipe. Unlike previous work, we focus on the role of knowledge and types of information. We find that although users expect the system to be able to discuss the key elements of the cooking process (e.g., time, temperatures, equipment), they are also interested in the ability of the assistant to answer knowledge-based questions (i.e., questions that seek information on the science, history, and so on, of a recipe or cooking step).

To address RQ2, we focus on the role of the assistant to be able to engage with the user to proactively (system initiative) offer and suggest relevant information at the appropriate time. We perform a Wizard-of-Oz study using the open-source TaskMAD framework [41], which supports multimodal output that includes images and videos. A key element we study is the role of the wizard in communicating knowledge with one key experimental variable: the wizard policy that is either *passive* (control) and that responds to user questions and the *active* policy (experimental), modelled on a teacher/teaching process—asking questions to have the user be more mindful and engaged in not just what they are doing, but why they are doing it. Here, the wizard prompts the user and proactively indicates that it is in possession of relevant knowledge. The passive setup is typical for most current assistant systems. Although others have looked at clarifying questions, this is the first work, to our knowledge, that studies initiative in the cooking domain from the perspective of knowledge, in the form of scientific and historical content, and the first to explicitly study how relevant knowledge can be communicated by a conversational agent. We wish to facilitate and study conversations where users engage with questions intentionally to enhance their knowledge, rather than in scenarios focused on quickly preparing meals or where there are distractions from children or friends vying for the cook's attention.

Our in-depth study and analyses extend previous work and augment existing information intent taxonomies with new knowledge categories. We use a nugget-based approach to study the differences in the amount and types of knowledge in the conversations based on different policies. According to our findings, the number of utterances in conversations for the *active* policy was almost double compared to the other policy. Additionally, we observe that in the *active* condition, user questions were 1.5 times more likely to be related to knowledge rather than cooking steps. Moreover, an *active* policy resulted in over 1.7 times more knowledge being communicated compared to the conversations in the *passive* condition.

Overall, our study provides new and innovative findings for knowledge-grounded conversational information assistants. It suggests that users are interested in an agent capable of providing background knowledge. We also find that an active policy is an important design factor for an assistant, leading to longer conversations and significantly increasing the amount of knowledge transferred.

This research benefits developers and researchers in conversational agents, offering insights into user expectations and the impact of interaction policies in the context of cooking assistance. Smart device manufacturers and designers of cooking assistants can find valuable guidance for enhancing user experience and knowledge transfer in this domain. Additionally, academics in natural language processing and human-computer interaction can leverage the study's methodology and findings for further research.

2 RELATED WORK

The presented work builds on and is influenced by a growing body of related research. We summarise this in two main subsections. The first reviews how humans interact with conversational assistants and what is known to influence their behaviour; the second focuses on the kitchen domain by reviewing research contributions offering assistance in this context.

2.1 Interacting with Conversational Assistants

Extensive research has been conducted by numerous scholars to explore the dynamics of human interaction with conversational assistants, as well as the multitude of factors that shape and influence these interactions. Numerous investigations have been conducted in diverse contexts to explore the interaction between users and conversational assistants. These investigations span various tasks, including practical ones such as booking flights [11] or holidays [40], as well as informational tasks such as finding accounts of heroism or evaluating the pros and cons of medical treatments [42]. A crucial aspect of successful user-assistant interaction is the accurate comprehension of the user's intentions. In the literature, intent prediction has been extensively studied, with researchers such as Qu et al. [36] and Ghosh et al. [20] focusing on predicting dialogue acts and speech acts to understand user intent from a linguistic standpoint. Additionally, there are studies that concentrate on task-specific intents and information requirements [19]. Some approaches combine both linguistic and task-specific intents, as demonstrated by Shiga and colleagues in their modelling efforts [40].

The way people converse with an agent to complete a task has been shown to vary based on numerous variables, including the complexity [44] and difficulty [42] of the task. The characteristics of the agent are also important. Users tend to exhibit preference for agents whose conversational style matches their own [39]. Thomas et al., who investigated the effects of an agent's style in detail, did not find any single "best" style but reported several effects on aspects such as perceived effort, engagement and the feeling of being understood by the agent [42]. When agents make reference to previous utterances, this leads to greater user satisfaction and a lower cognitive load [13]. Thus, the way an agent communicates can have a strong influence on the user experience.

Several scholars have examined the effect of the agent strategy or initiative (i.e., the interlocutor driving the conversation). Researchers refer to a mixed-initiative spectrum where an active or passive agent influences the characteristics of conversations and variables such as workload, user satisfaction, and learning. Active agents tend to result in more but shorter conversational turns, including more follow-up questions [16], whereas passive agents have fewer turns that are often longer [16]. Active assistants can improve task performance [11] and be considered to be more

engaging and truthful by users, especially in goal-oriented tasks [14], but also in social-bot contexts active strategies have been shown to foster engagement [22]. In some settings, such as tutoring, active systems tend to facilitate the learning process [12]. Passive agents, however, have been suggested to be better suited to simpler tasks, where users do not need assistance in describing what they want [37, 44].

Not only does it seem that different tasks are suited to different interaction modes, but the evidence seems to suggest that, depending on the initiative strategy employed, users will need divergent support [1, 3]. Moreover, these opposing interaction modes come with their own challenges. Whereas passive agents need to decipher both needs and context from a user utterance [19], active agents are required to say the right thing at the right time—even a few seconds of silence after submitting a query to the agent can be perceived as an indicator for errors [35] and people may avoid using an active agent when an intervention is mistimed [2].

2.2 Assistance in the Kitchen

Assistance in a kitchen can take many forms. It is well documented, for example, that conversational assistants have functionalities that can be useful for cooking, such as setting timers or adding items to shopping lists [18, 21]. Assistance can be provided via recommendations for meals that a user may like to cook [17]. Accomplishing this goal can involve leveraging the preferences of users who have similar profiles to the target user [24], as well as considering the inherent properties of the food itself [15]. Furthermore, such personalisation can be adapted to address specific dietary needs, such as weight loss objectives [43]. The generation of recommendations can be carried out through conversational interactions, with minimal differences in the interaction patterns regardless of whether the user interacts with the system by typing or speaking [4]. Other systems aim to provide assistance during the cooking process. For example, the AskChef system [30] provides a recommendation for every step in a recipe utilising either a smart-speaker or a laptop screen, depending on the context and support needed.

A notable body of work has focused on how users interact to gain assistance, for example, when interacting with a human mimicking the "perfect conversational assistant" [19] or via WoZ studies [4, 47]. Frummet provides a taxonomy of information needs that people might ask an assistant while cooking. This is an important step in understanding user needs, however, the empirical setup restricted the sample to a small and relatively homogeneous group. Inspired by other conversational search research, Vtyurina and Fourney explored initiative in the cooking domain, discovering that unrestricted environments of communication exhibit many signals that could be processed by future assistants [47]. For example, implicit cues, such as "okay," express the intent to move to the next step, something that is not captured well by current systems [48]. This finding has since been confirmed in further studies [19, 33] where users make extensive use of such cues when they communicate with an assistant. Studies in this space have employed varying initiative strategies. The human agent in Reference [19] acted as a passive collaborator, whereas the wizard in Reference [4] played a pro-active role by explicitly prompting for information and asking clarifying questions.

Two primary points can be extracted from the related work: (1) a broad variety of assistance is possible, but we do not yet know what people need or expect from a digital kitchen assistant. To date, only small-scale studies with homogeneous samples exist [19]. Therefore, we aim to learn about the information users expect to attain from digital assistants in a cooking context (RQ1). (2) While we know that agent interaction strategy influences user behaviour, we do not yet know what impact this has on the kinds of questions asked in the cooking context and the assistance received (or knowledge transferred) as a result (RQ2).

3 SURVEY: WHAT DO USERS EXPECT FROM DIGITAL ASSISTANTS IN A COOKING CONTEXT?

To build a picture of what people expect from conversational cooking assistants and answer RQ1, we performed a survey. During this survey, participants provided feedback on the level of support they anticipate for various features from such an assistant. The questions posed were primarily drawn from existing literature, with a majority based on the information needs categories established by Frummet and colleagues [19]. These include questions such as whether they think assistants should be able to recommend recipes, guide them through the recipe steps, and help them learn the cooking techniques required by the recipe. In addition, we included "knowledge-based" needs. These needs, while not present in Frummet et al.'s study, are deemed significant in the conversational search literature [41, 49], and they are plausible in a kitchen setting. For example, in Reference [48], the authors found that 2.8% of all requests asked for the definitions of ingredients or cooking procedures, such as "What is an ear of corn?", "What do you mean to taste? How much should I put?", "How high is medium-high heat?" We presented each of the survey questions with an example query (user utterance to request information) and arranged the questions in random order. The questions were tested and refined in a preliminary study involving a convenient sample of students, colleagues, and friends of the authors. Some participants gave qualitative feedback on potential misinterpretations of the questions and suggested alternative phrasings, which were accounted for in the final version. Finally, the participants provided demographic information regarding gender, age, education, employment status, and cooking habits. A complete list of the questions answered and their formulations can be found in Appendix A.1.

Two hundred participants were recruited via the crowd-sourcing platform Prolific. We restricted our sample to native English speakers from the US or UK to make our findings as comparable as possible with the second study, which required familiarity with the English language ingredients and cooking terminology. One-hundred twenty-three participants (61.5%) were female, 75 (37.5%) were male, and 2 participants (1.0%) did not attribute themselves to either the male or female gender.

Most people in this survey are aged between 25 and 34 years (n = 60 (30%)), followed by 52 participants (= 26%) being between 35 and 44 years old, 29 participants (= 14.5%) are aged between 45 and 54, 28 (= 14%) are between 18 and 24 years old, 22 (= 11%) are aged between 55 and 64, and 9 participants (4.5%) are older than 65. The sample included individuals with diverse employment backgrounds ranging from home workers and stay-at-home partners to students, freelancers, and retired people. The sample also included business owners and others employed by businesses. The sample included both unemployed and full- and part-time workers. The frequency of cooking and the pleasure derived from cooking also varied ($\overline{x}_{cooking\_enjoyment} = 5.13, min_{cooking\_enjoyment} = 1$, $max_{cooking\_enjoyment} = 7, SD = 1.60$). Most of the participants reported cooking 2 to 3 times (n=61 (30.5%)) or 4 to 6 times a week (n = 52 (26%)). Thirty-six participants tend to cook on a daily basis (=18%), 25 once (=12.5%), and 23 less than once a week (=12.5%). Three people indicated that they do not cook at all (=1.5%). When categorising participants by age (older: > 45 vs. younger: ≤ 45)[1] and cooking frequency (often, i.e., at least two to three times a week vs. less often), it becomes evident that older participants in our sample cook significantly more frequently compared to their younger counterparts ($\chi^2 = 4.22, df = 1, p < .05$). However, employing the same approach reveals no significant differences in the usage of smart assistants between these two age groups. This suggests that, in our sample, the older participants exhibit greater technical proficiency than the general population [32], potentially fitting the description of what is commonly referred to as

[1]Forty-five was taken as a cutoff, as it was a central value in our distribution and easily delineated based on our scale.

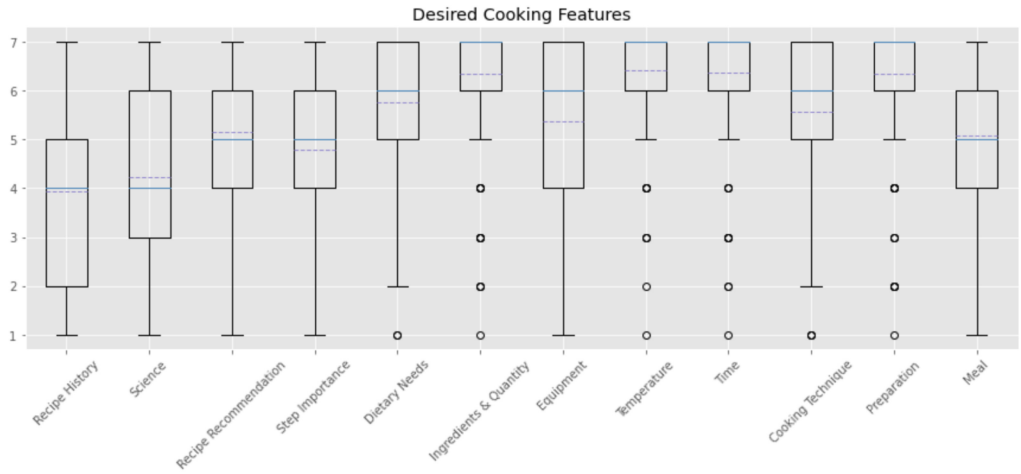

Fig. 1. Desired cooking features by participants.

"silver surfers" [8]. We believe this is appropriate, given our research aims, since we would expect that only technically proficient or indeed technically curious older adults would use such a system in practice.

3.1 Results

The distribution of responses to each question is shown in Figure 1. The main takeaway from the graphic is that the participants can envisage all of the suggested functionality provided by conversational kitchen assistants. The median response for all of the questions is higher than the mid-point of the Likert scale. A second observation is that the median responses are higher for some features than others. For example, features revolving around the cooking process (e.g., ingredients and their quantities, cooking temperature, cooking time) were scored consistently high, whereas the knowledge-grounded features, which did not appear in Frummet's study, were less so. The knowledge-grounded questions also exhibited the largest variance in responses.

To dig deeper into this, we examined the data to determine whether demographic information influenced how participants responded. The findings suggest that different demographic groups have varying expectations regarding the types of assistance that should be provided. Visualising the data revealed that older participants (>45, $n = 60$) rated knowledge-grounded features lower than younger participants ($\leq$45, $n = 140$). More precisely, younger people rated recipe history higher ($M = 4.28; SD = 1.80; MED = 4.0; IQR = 3 - 6$) than older people ($M = 3.15; SD = 1.63; MED = 3.0; IQR = 2 - 4$). Also, science questions were seen as more important by younger ($M = 4.44; SD = 1.73; MED = 5.0; IQR = 3 - 6$) and less important by older participants ($M = 3.77; SD = 1.84; MED = 4.0; IQR = 2 - 5$). This is confirmed statistically.[2] Younger participants are significantly more interested in recipe history ($U = 5679.0; p < .001$) and science questions ($U = 5061.5; p = .020$) compared to those in the older participant group. Slight correlations were observed between reporting enjoying cooking and wanting to learn more about recipe history ($r = .13; p = .062$) and science ($r = .17; p = .014$).

[2]We utilised non-parametric tests when the conditions for conducting a parametric test, such as normal distribution, were not satisfied, as indicated by a Shapiro–Wilks test. In the reported correlation tests, we applied Pearson's r.

4 METHODOLOGY: THE WIZARD-OF-OZ STUDY

The expectation for an assistant to not just handle procedural steps but also handle knowledge-related questions is the motivation for focusing on this previously unexplored aspect in the second study. We perform a user study simulating a cooking scenario where participants work through the steps of a recipe and are encouraged to converse with an agent and ask knowledge-grounded questions that come up along the way. It is important to note that, for practical reasons, participants did not physically cook the recipes during the study. The logistics of controlling this process would have been highly intricate, potentially prolonging the experiments and complicating an already challenging recruitment process. As detailed later, we made efforts to simulate the cooking process to closely resemble a naturalistic experience. We address the implications of this in the Limitations section.

Inspired by the literature, participants were randomly assigned to one of two conditions: *passive*, where the wizard simply responded to messages and questions from the participant; and *active*, where the wizard proactively interacted with participants to indicate that it possessed knowledge about the cooking steps and asking participants if they were interested to learn. Participants in both conditions were free to ask whatever questions they wanted, but there was no explicit need for them to ask questions.

The experimental condition (active vs. passive) was balanced such that each of the six recipes used featured in each condition with one of two wizards. That is, the two wizards had the same coverage in terms of recipes and conditions totalling 24 conversations for each wizard. Participants were unaware they were interacting with a human until the post-experiment debrief. In the following subsections, we explain the methodology in detail, outlining the procedure (Section 4.1), participants (Section 4.2), and the justifications for studying the recipes we did (Section 4.3). We then continue to detail the Wizard-of-Oz setup (Section 4.4), including an extensive piloting process used to establish guidelines for wizard behaviour (Section 6.5). Finally, in Section 4.6, we outline the measure we used to quantify the characteristics of conversations, as well as the knowledge transferred from wizard to participant.

4.1 Procedure

Each experiment comprised the following steps:

(1) Participants read the informed consent form.
(2) If they agreed to the study conditions, then we provided them with specific task instructions and a short tutorial on how to use the chat interface.
(3) After the tutorial, the participants opened the chat interface to receive a welcome message from the cooking assistant (wizard) and start the experiment.[3]
(4) After completing the task (i.e., working through the recipe), participants were thanked and debriefed before their participation was confirmed on the crowd-sourcing platform.

Participants could take as much or as little time as they wished for the study. The time taken ranged from 30 to 62 minutes.

4.2 Participants

As with the survey, participants were recruited via Prolific based on our experience that the portal provides access to heterogeneous and motivated participants that deliver high-quality data. We applied the same restrictions in our sampling process as the survey, accepting only native English

[3]Welcome message and task instructions can be found here: https://github.com/AlexFrummet/cooking-with-conversations/tree/main/woz_study_instructions

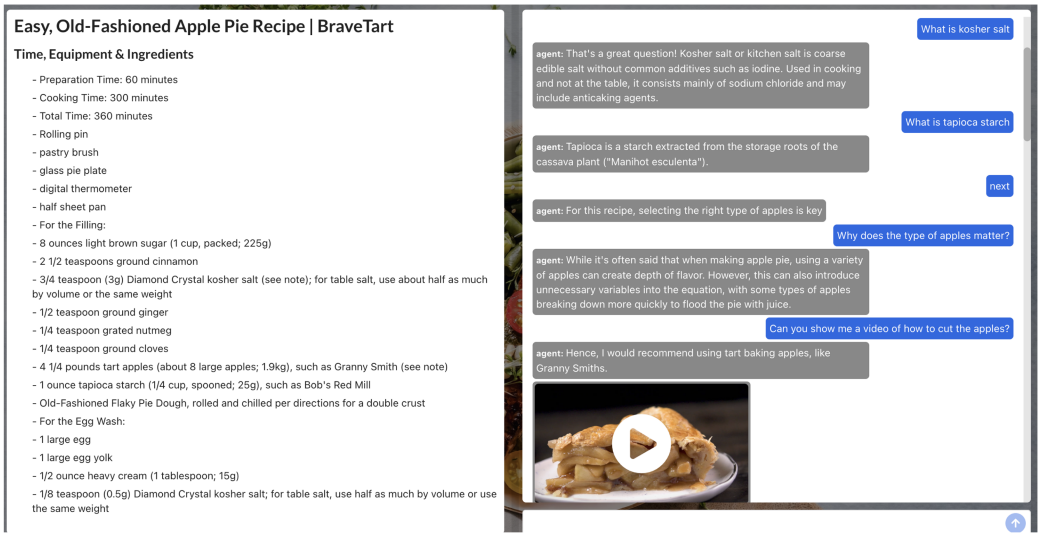

Fig. 2. User chat interface.

speakers from the US or UK. This makes it possible to compare the findings and was necessary, since the recipes contain complex cooking procedures in English. Participants received between 9 and 11 GBP as compensation for their time. This ensured that participants received at least the minimum remuneration as defined by Prolific, and the majority received considerably more.

Forty-eight participants were recruited, 65% of which identified as female, 33% male, and 2% as non-binary, similar to the distribution in the survey. Most participants were aged between 45 and 54 ($n = 16(33.3\%)$), followed by 10 (20.8%) between 25 and 34, 9 participants (18.9 %) were aged between 35 and 44, 5 (10.4%) between 18 and 24, and 5 (10.4%) between 55, and 64. 3 participants (6.3%) were aged between 65 and 74 years old. Unsurprisingly, given the recruitment procedure and the subject matter and type of study, the sample was biased towards individuals who cook or have an interest in cooking.

4.3 Recipe Selection

The sessions were based around the steps involved in cooking six recipes from the SeriousEats[4] website. These are ideal for our experiment, since they feature interesting ingredients and cooking procedures and are complemented by associated how-to resources including the history behind the recipe and the science behind its methods. The recipes were chosen such that they contained a minimum of five steps and had at least one picture illustrating the steps. We assumed pictures would help with the simulation of the cooking process by allowing participants to imagine what the outcome of the steps would have looked like. The recipes were diverse in that they included both main meals and desserts, as well as omnivorous and plant-based dishes. We wanted recipes to be appealing to a broad set of participants.

4.4 Wizard-of-Oz setup

We utilised TaskMAD [41] as a platform for our experiments. The interface, used by the participants, provided a simple chat interface as well as contextual recipe information presented as sequential steps (see Figure 2). The wizard interface (Figure 3) presented wizards with information

[4]A list of the recipes used can be found in Table 5 in Appendix A.2.

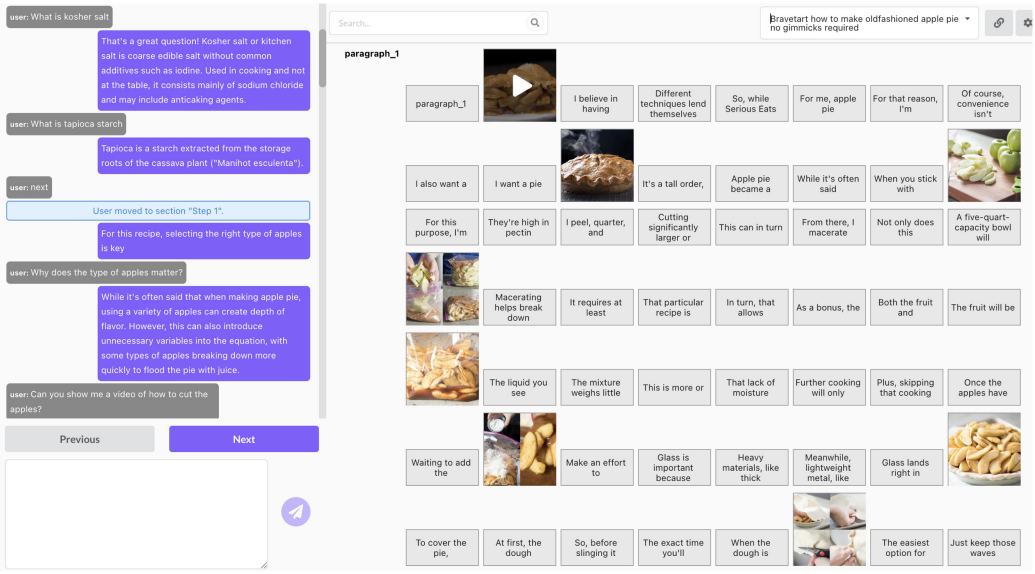

Fig. 3. Wizard interface.

about the recipe and how-to in a structured manner. The interface also provided the ability to perform federated searches over custom external data sources. The resources included the SeriousEats website (how-tos), StackExchange Cooking[5] (questions and answers), and Wikipedia (KILT) [34]. Pilot sessions demonstrated that these datasets would be sufficient to answer most questions.

To share the workload and minimise personality and learning effects, the experiments were conducted by two wizards. Extensive piloting (28 experiments and over 21 hours of conversation) familiarised the wizards with the interface and allowed behavioural guidelines to be established. After every test run, wizards discussed their experiences and how they reacted, leading to a consistent response framework for both conditions. Numerous pilot sessions were necessary to ensure consistent interaction strategies and establish saturation regarding the types of questions participants might ask and how to respond to them consistently. In the following section, we outline how the pilot sessions informed the final study.

4.5 Lessons Learned from Pilot Conversations

The pilot sessions led to formalised procedures for the active condition that reflected the experiences and lessons learned in terms of how and when wizard interventions should be formulated. The active agent condition, which has the aim of provoking curiosity during a task-based context, blends a task-oriented setting [23, 25, 31] with more social chatbot settings where the aim is to promote user initiative and engagement [5, 22]. There are differences between our setting and a socialbot. In socialbot contexts, the questions focus on eliciting a users' preference or topic interest. Whereas, here, the task and topic are fixed and questions aim to invoke users' interest.

The wizards took into account the factors that had sparked interest and interaction from the participants during the pilot conversations when formulating the guidelines. Specifically, they deliberated on how to create effective prompts and when it was suitable to present them. This resulted in two main strategies that encourage participants to reflect on the background knowledge

[5]https://cooking.stackexchange.com/

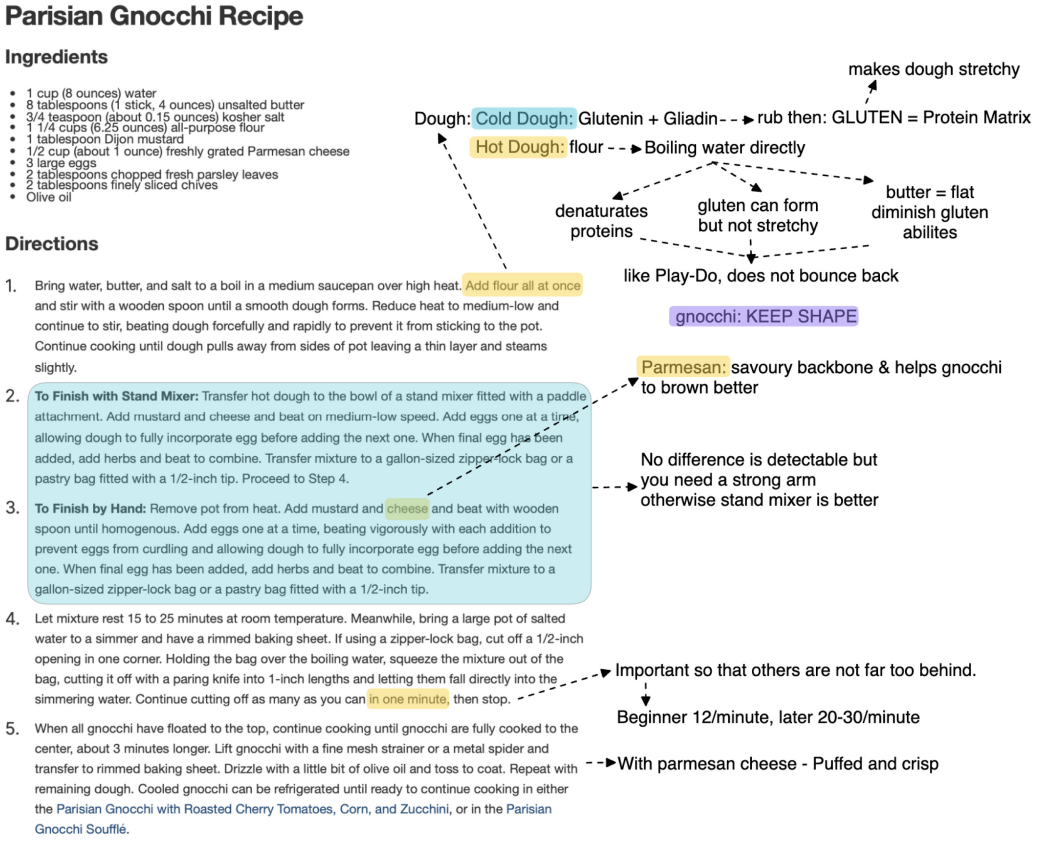

Fig. 4. Example of a mapping used by the wizards in the experiments.

necessary for each recipe step. The first was to **make statements** emphasising the importance of an aspect of the process, e.g., *Don't forget to add X as this is crucial to make the perfect Y* or *Notice that X is important here*. The second strategy was to **formulate a question** that relates to required knowledge, e.g., *Why do you think apples should be put in a gallon-sized zip-top bag?* Both of these strategies are similar to those that have been shown to be effective in increasing user initiative in social chatbot settings [22].

In terms of when to actively intervene, the wizards agreed that an appropriate moment was when the participant transferred to the next recipe step. It was determined that an intervention would be especially beneficial when the recipe step description provided limited information in comparison to the corresponding how-to instructions for that particular action. This was determined based on a mapping illustrated in Figure 4. Wizards informed the participants about this additional knowledge in the following ways:

> "Having a hot dough has significant effects on the shape of the parisian gnocchi." (conv. 2, active cond.)

or

> "Before going next, be aware of the impact cheese has on the gnocchi." (conv. 2, active cond.)

In the passive condition, the primary guideline was that wizards were only to react when explicitly asked a question, as was the case in Reference [19]. The wizards behaved intentionally personable in both conditions. According to existing research, for a bot to engage in social communication effectively, it needs to possess qualities such as empathy, supportiveness, and a genuine interest in the thoughts and ideas of its human interlocutor [29, 38]. This included being friendly, positive, and polite. In the active condition, when participants correctly answered a question, this was intentionally praised to build confidence and encourage further questions to be asked. See, for example,

> "Oh wow, that's correct!! You really are an expert! [...]" (conv. 7, active cond.) or "Wow, seems that you're already an expert :-) You're right: [the wizard continued to provide more extensive details]" (conv. 14, active cond.)

The wizards were also encouraged to express their own personal interest in the recipe and relevant facts like in

> "The zip-top bag trick is great, right?" (conv. 15, active cond.) or "[the wizard provides background knowledge] So this is why you need to add flour directly all at once to the saucepan. Interesting, right? :-)" (conv. 14, active cond.)

This approach pertains to personal disclosure and the concept of the "disclosure-reciprocity effect" [9], which has been observed in chatbots to result in users sharing more information than they would typically [28]. We anticipated a comparable transfer of curiosity in our study.

By the time the full study started, both wizards were intimately familiar with the recipes and were prepared for potential questions about techniques, chemical processes, and related information associated with recipe steps.

4.6 Measuring Conversations

To understand the impact of wizard behaviour on the resulting conversations, it was imperative to establish robust quantitative metrics for the conversational characteristics. In this section, we outline the metrics and justify choices. Following the definition in Zamani et al. [49, p. 4], we define an utterance as a message that has been sent by either the wizard or the participant. An utterance can consist of multiple sentences. The resulting conversations comprise 1,396 utterances in total. To examine the kinds of questions asked by participants, we annotate utterances using an appropriate information needs taxonomy and derived a process to quantify the knowledge transferred by the agent as a result. The following subsections describe the annotation processes in detail:

4.6.1 Annotating Questions. To establish the kinds of questions users tended to ask, the utterances were annotated using the information needs taxonomy as defined by Frummet et al. [19]. This taxonomy is useful, as it offers a means to describe the information needs that occur during real-life cooking sessions. Only utterances determined to be questions were classified. Each utterance (question) received one information need label from level 1 of the taxonomy, out of the 12 labels that distinguish between questions on ingredients, their quantities, the cooking process, and so on. The annotation was conducted by the lead author of this article who derived the taxonomy in prior work. He is an information scientist, experienced in annotating both spoken and written conversational data using qualitative methods. We did not feel it necessary to annotate with more than one coder, since the reliability of the coding scheme and the coder's consistency in its application with other annotators were assessed in prior research. In that study, 10% of a

comparable, albeit naturalistic, dataset was annotated by the same coder and another annotator. Together, they attained a Cohen's κ score of 0.75 [19]. To ensure consistency and reliable annotation in the current dataset, however, a subset of 50 randomly selected utterances was relabelled by the same annotator resulting in a Cohen's κ score of 0.87, which is considered an almost perfect agreement according to Landis & Koch [27]. It became clear very early in the annotation process that the categories at level 1 of the taxonomy were appropriate for our purpose, since the questions in our dataset and the phraseology employed by participants were very similar to those reported in the previous work. For transparency, we disclose that the annotator served as one of the wizards. The annotation was conducted without knowledge of which experimental condition corresponded to the transcripts. The process was based purely on the user utterances alone and took place several weeks after data collection. Consequently, we do not believe this introduces any bias to the process or the findings.

The labels illustrated in Table 2 focus on two types of labels—*Process* and *Knowledge*. *Process* questions relate to the actions participants needed to take and, in our data, were typically phrased as *what* questions, for example:

> "Ok, **what's** the first step?" (part. 0) or "**What** do I do after I've made the pesto?" (part. 3)

We did observe other kinds of formulations for process questions such as

> "Can I use a knife?" (part. 20, passive) and "Does it [strainer] have to be a metal one?" (part. 21, passive)

Knowledge questions, in contrast, sought knowledge in the form of explanations. These were mostly phrased as *why* or *how* questions. For example:

> "**Why** [should I use] Dijon mustard?" (part. 1) or "Agent: Egg yolks play an important role in this step. Part.: **How?**" (part. 8).

To better understand what kind of *Knowledge* questions were asked, we further annotated these questions with an additional label not present in the original taxonomy, which describes the type of knowledge being sought. These labels were *Science* when participants wanted to know about the underlying mechanisms. For example:

> "What does blanching do?" (part. 5) or "How does reducing the temperature affect the duck instead of keeping it at a normal temperature and cooking for a shorter time?" (part. 13)

These questions prompt the agent to provide knowledge that explains underlying scientific processes. *History* questions were different, for example:

> "What's the origin of a soufflé?" (part. 8) or "Is [the dish] French?" (part. 5)

These questions triggered information about origins of recipes and meals. Questions about the *Step Importance* were phrased, for example, as follows:

> "Why should I put cream in my egg wash?" (part. 15).

When this kind of question was asked, participants expected suitable explanations from the agent, which, in some cases, also involved providing scientific knowledge. The distinguishing criteria

Table 1. Example Questions from Knowledge Information Need Subtypes *History*, *Step Importance*, and *Science*

History	Step Importance	Science
Where is the dish from?	Why do we use cream of tartar?	What does blanching do?
Where is cayenne pepper from?	Why do I flip the bag?	Will frying a second time change the nutritional value of the recipe?
What other types of soufflé are there?	Why did you add tapioca starch?	What does macerate mean?
Can you tell me the history of this chicken recipe.	Why coat the chicken in flour mixture?	What does putting 3 tablespoons of marinade into the flour do?
When was bechamel sauce invented?	Why seasoning lightly with salt before it goes into the oven?	Why does soufflé deflate?

between *Science* and *Step Importance* questions are that in *Step Importance* questions, participants explicitly ask **why** there is a need to perform a certain action, for example:

> "**Why should I put** cream in my egg wash?"

This is not the case for *Science* questions where participants do not ask for the reason a specific step or action needs to be performed. To enhance the reader's understanding of the types of questions that were asked and how these were phrased, further illustrative examples can be found in Table 1.

4.6.2 Quantity of Knowledge Communicated. To establish the quantity of knowledge communicated by agents in the conversations, we counted what we refer to as *information nuggets* in wizard utterances. An information nugget is an atomic piece of information that the assessor considered useful or interesting and that had not previously featured in the same utterance [46]. To avoid bias, the utterances were annotated independently (i.e., free from conversational context), in a random order, and without any indication of the experimental condition.

The following example illustrates that this can be relatively straightforward with all of the annotators (we employed three) agreeing that the utterance contains two nuggets of information:

That's a great question! *Chilling ensures the dough is cold to start*, *which keeps the crust flaky and light*.

Nugget 1 Nugget 2

As a general rule, the annotators agreed to treat multiple adjectives with a separator, as in nugget 2, as a single nugget. Other utterances were less straightforward to define. In the next example, all three annotators counted differently:

Tapioca is a starch, *extracted from the storage roots of the cassava plant* ("*Manihot esculenta*").

Annotator 1 Annotator 1 Annotator 1

Annotator 2 Annotator 2

Annotator 3

As shown in the example, Annotator 1 counted three information nuggets, Annotator 2 counted two, and Annotator 3 counted only one piece of information.

To establish the consistency of annotation, the annotators each labelled 638 agent utterances with the number of information nuggets they believed each utterance to contain. Despite the differences in annotating granularity, the mean pairwise (pairs of annotators) Pearson's correlation between the counts of the three annotators was very high ($\bar{r} = 91.89\%, r_{max} = 94.33\%, r_{min} = 89.31$). To determine if it is possible to replicate the human annotations and automate the process, such that we can scale to the entire dataset, we provided a suitable prompt to GPT-3[6] This resulted in an average correlation of $\bar{r} = 78.83\%$ with the human applied counts and a mean absolute error of 0.71, $SD = 1.37$ when the GPT-3 counts are compared to the average of the three human annotators. We conclude that this is a sufficient signal to use GPT-3 counts to test the amount of knowledge conveyed in utterances in the experimental conditions. Consequently, we used the counts provided for each utterance by GPT-3 on the full dataset as the basis for our analyses below.

5 EXPERIMENTAL RESULTS

This section addresses RQ2. Our objective is to investigate how participants interacted with the wizard and whether the wizard's mode of interaction influenced user behaviour and the information transferred. We present our findings in four sections: In Section 5.1, we outline the general conversation statistics. Section 5.2 explores the types of questions that were asked during the conversations. Subsequently, we evaluate the wizard's responses and the extent of knowledge shared with the participants in Section 5.3. Finally, we examine the knowledge sources utilised to answer participant questions in Section 5.4.

5.1 Conversation Characteristics

In the first step, we focus on evaluating the overall characteristics of the conversations. Specifically, we analyse whether there were differences between the two interaction modes concerning the quantity and length of utterances, as well as the interactions between the user and the wizard.

The active condition yielded more utterances compared to the passive one. Specifically, the agent and participant issued 1,005 and 923 utterances, respectively. In the passive condition, however, 989 utterances were gathered: 409 by the agent and 580 by the participant. In both conditions, agent utterances were significantly (active: $U = 742{,}794.5, p < .0001$; passive: $U = 182{,}621.0, p < .0001$) longer (active: $\bar{x}_{words} = 23.59$; passive: $\bar{x}_{words} = 25.60$) than user utterances (active: $\bar{x}_{words} = 5.25$; passive: $\bar{x}_{words} = 5.40$). Furthermore, the number of utterances in the active condition ($\bar{x}_{utterances} = 80.33, SD = 25.40$) was significantly ($U = 529.0, p < .0001$) higher compared to the passive condition ($\bar{x}_{utterances} = 41.21, SD = 13.31$). Thus, the active condition led to longer conversations overall, however, the number of user utterances compared to agent utterances and the length of user utterances remained similar regardless of the condition.

The transition graphs in Figures 5 and 6 visually depict the disparities in user and wizard interactions across the various conditions. Modelling the dialogues in this way illustrates how the conversations moved from one type of user/agent utterance to another. The active condition is generally more connected with several transitions (shown in red) in this condition that were not present in the passive condition.

[6]Davinci model 3 using the Open AI API. The full prompt can be found here: https://github.com/AlexFrummet/cooking-with-conversations/blob/main/data/gpt-3_prompt.txt. We experimented with newer GPT models but were not able to improve on this performance for this task.

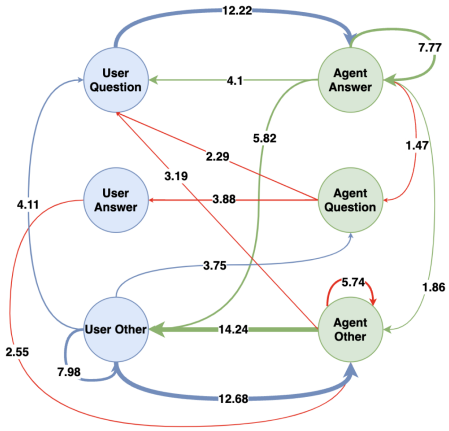

Fig. 5. Transition graph active condition with transition probabilities as percentages.

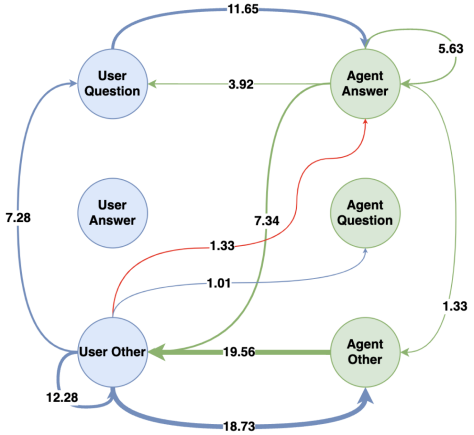

Fig. 6. Transition graph passive condition with transition probabilities as percentages.

Table 2. Information Need Distribution in the Active and Passive Conditions

Info. Need Type	Active		Passive		Examples
	n	%	n	%	
Knowledge	117	**37.86**	53	**25.12**	Why whisk constantly?
Process	192	62.14	158	74.88	Can I use salted butter?
Knowledge Needs					
History	20	15.63	21	25.93	Where are pinenuts from?
Science	40	31.25	8	9.88	Why is one-third of volume lost?
Step Importance	41	32.03	19	23.46	Why do I brine the chicken pieces?
Other	27	21.09	33	40.74	Why is a glass plate better?

5.2 Distribution of Information Need Types

Here, we examine the annotations based on Frummet et al.'s taxonomy [19]. As Table 2 shows, participants asked questions in both the active and passive conditions. Overall, however, 1.5 times more questions were asked in the active condition [309 vs. 211, $\chi^2 = 13.95$, $df = 1$, $p < .001$], and information needs were 1.5 times more likely to be knowledge-related in the active compared to the passive condition [37.86% vs. 25.12%,$\chi^2 = 78.59$, $df = 3$, $p < .0001$].

These findings suggest that when the agent hints that it possesses additional information, it effectively prompts users to ask questions. However, we wish to emphasise that agent interventions were far from guaranteed to result in questions. This is demonstrated by cases where the participants were clearly not interested in discovering what the agent knows and replied that they wish to move to the next recipe step:

> "Agent: There are three possible ways of making [the sauce]. Part.: Next" (part. 5) or "Agent: The way in which you cut apples here is a crucial step when making apple pie. Part.: Next" (part. 12)

Since *Knowledge* questions were the most commonly asked question type, we examined these more closely to establish the types of knowledge-related questions participants asked. The bottom part of Table 2 illustrates that, in both conditions, most of the questions asked were related to *science*, *step explanations* (= step importance), and *history* aspects. Whereas questions about the *history* of the recipe or ingredients were balanced across conditions, *science* questions were nearly five times and *step explanations* nearly two times more common in the *active* condition.

It is important to note that participants did not need to explicitly ask *knowledge* questions to receive knowledge from the wizard. Process-related questions also led to knowledge transfer. Examples of this in our data include:

> "Part.: How small should the additions of vinegar be? – Agent: Very small additions. It will boil and bubble violently, so take your time to avoid a boil-over. (amount, part. 13)"

5.3 Knowledge Communicated in Conversations

In this section, we analyse the quantity of information imparted in conversations to assess if this varied across conditions. A total of 309 out of 1,415 agent utterances included information nuggets. On average, conversations in the active condition contained 15.88 information nuggets ($\tilde{x}_{info\_nuggets} = 14.50; IQR = 10 - 21$), which was significantly higher ($U = 144.5; p = .002$) than the passive condition, where a conversation contained an average of 9.13 ($\tilde{x}_{info\_nuggets} = 9.5; IQR = 3.75 - 12.25$).

From the analyses in Section 5.2, we know that knowledge was not only transferred in response to *Knowledge* questions, but also via answers to *Process* questions, including those classified as *Ingredient*, *Equipment*, *Time*, or *Preparation*. To assess the quantity of knowledge conveyed in the agent's responses, we calculated and compared the amount of knowledge transferred across various information needs.

In the context of *Knowledge* and *Process* questions, it was observed that the wizard displayed a higher tendency to provide more extensive knowledge-based responses compared to process-oriented inquiries. On average, a response to a *Knowledge* question included 1.58 ($SD = 1.78$) pieces of information, more than double than for *process* questions, which was 0.79 ($SD = 1.39$) pieces of information. This difference is also significant ($U = 19{,}396.5, p < .0001$).

Our results indicate that, on average, a greater amount of knowledge is conveyed in a conversation when in the active condition. This suggests that conversational assistants that are actively engaged can assist users in acquiring more background information while they are cooking.

5.4 Sources of Knowledge Used

As described in Section 4.4, wizards had access to various sources of information, including recipes and accompanying how-tos from SeriousEats, StackExchange Cooking, and Wikipedia. They also had access to other pages from Serious Eats, not directly related to the recipe used in our study. Table 3 displays the frequency with which each knowledge source was utilised to respond to specific types of questions, grouped by the corresponding information need.

While Wikipedia, recipe directions, and associated recipe how-tos were the most frequently utilised sources overall, the findings emphasise that different sources were employed depending on the type of question asked. The information in a recipe was primarily utilised to address questions related to the *process*, although, for such needs, how-tos, Wikipedia, and other sources were more frequently employed.

How-tos of the recipe served as the primary source of information for answering *knowledge-based* queries, particularly for queries pertaining to the significance of steps and scientific aspects.

Table 3. Information Source Types Employed by the Agent Based on Information Need Type

Info. Need Type	Recipe		How-to		Wikipedia		Other sources	
	n	%	n	%	n	%	n	%
Knowledge	7	3	147	61	60	25	26	11
History	0	0	12	20	46	77	2	3
Step Importance	6	7	66	79	2	2	10	12
Science	1	1	52	73	7	10	11	16
Process	47	18	110	42	55	21	52	20
All Needs	54	11	257	51	115	23	78	16

In contrast, queries about historical context were largely resolved by referring to Wikipedia as a source of information.

These results offer valuable insights for replicating similar systems in practice. Depending on the particular information need a user has, different knowledge sources should be utilised by the agents, and this could be used to weight silos in an information retrieval setup.

5.5 Understanding the Influence of Intervention

As described in Section 6.5, wizards applied two different kinds of intervention in the active condition: *statements* and *questions*. There was no guideline as to when and how these should be used, and this was left up to the wizards themselves to decide. Here, we analyse if there is any evidence post-experiment that the tactic employed influenced outcomes.

We counted and analysed questions and information nuggets, which occurred in the conversation between a wizard intervention and either the subsequent intervention or the user proceeding to the following stage in the recipe. This is illustrated in more detail in Figure 7. Conversation (a) shows which turns we examined for our analysis between two wizard interventions. Conversation (b) shows the case where a wizard intervention occurs prior to proceeding to the next step in the recipe. Here, we used the turns between the wizard intervention and the user's next step message for our analysis.

In the case where questions were used as an intervention ($N = 49$), this resulted in an average of 2.67 questions ($\tilde{x}_{questions} = 2.0$, $IQR = 2 - 4$) being asked, which is more than for statement interventions ($N = 71$, $\bar{x}_{questions} = 2$, $\tilde{x}_{questions} = 2.0$, $IQR = 1 - 3$, $U = 1275.0$, $p < 0.006$). We did not find any significant differences in the distribution of kinds of questions asked. The fact that more questions were asked did not, however, result in more knowledge being transferred. We found no significant difference between the counted nuggets ($\bar{x}_{nuggets} = 1.8$ nuggets after statements ($IQR = 0 - 3$), $\bar{x}_{nuggets} = 2.0$ after questions ($IQR = 0 - 3$)).

6 DISCUSSION

This section integrates the main findings from the survey, the differences in outcomes observed between the active and passive conditions in the WoZ study, and the insights gained from the 28 pilot experiments that led to the strategy guidelines for the wizards.

The key outcomes of the work are:

— The expansion of an existing taxonomy from Frummet and colleagues to include more precise and detailed descriptors for knowledge-based information needs. This not only has theoretical significance but also has practical implications, as it enabled us (and would others) to conduct a more detailed analysis of the types of questions that were asked.

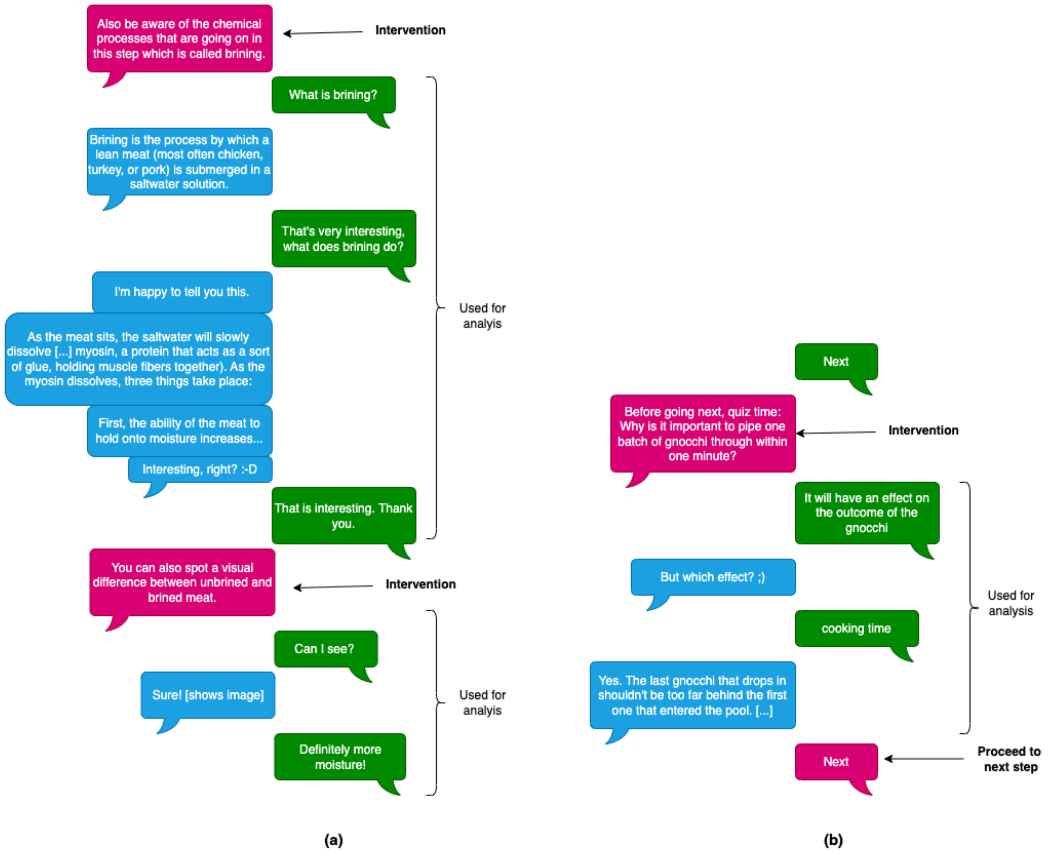

Fig. 7. Intervention analysis. Utterances on the left are from the wizard, utterances on the right are from the user.

— Through our survey, we learned what users want from conversational assistants in a kitchen context. We discovered that individuals are interested in gaining knowledge about the food they cook, such as the science and history of the dish, the ingredients used, and how they are prepared. The survey results also revealed demographic patterns, with younger individuals who are more passionate about cooking showing a greater expectation for conversational assistants to offer such information.
— Despite a sample with similar characteristics, we did not find these demographic trends in the WoZ study. There were no differences in the number or type of questions asked, nor in the amount of knowledge imparted across groups.
— The WoZ study revealed that agents implementing an active policy resulted in increased interactivity, a higher number of questions asked, especially knowledge-related questions, and a greater amount of knowledge being conveyed.
— The pilot studies enabled the wizards to test and establish guidelines for the active condition, which have implications for the design of future systems and introduce new research challenges.
— The conversations are available for researchers to further analyse and experiment with. We anticipate that the datasets will be valuable for testing retrieval and question-answering

algorithms within the context of conversational interactions. What sets our dataset apart is not only its focus on knowledge-based questions and answers in the domain of cooking, but that the information needs associated with the questions are grounded within the context of completing a particular task. The answers to these questions are located across different information silos. Our dataset comprises user-generated questions, contextual information about the associated recipe and recipe step, responses formulated by the wizard, and comprehensive provenance information indicating the sources and combinations of the answers. This makes our data different from existing resources such as *CookDial* [26], *TREC CAsT* [10], and *Wizard of Tasks* [7]. The *CookDial* [26] corpus consists of questions that are based on the recipe document without incorporating external knowledge resources. However, *TREC CAsT* [10] dataset offers comprehensive provenance information but lacks representation of task-oriented dialogues. In the *Wizard of Tasks* [7] dataset, conversations were generated through crowdsourcing and revolve around questions and answers tied to the recipe (or DIY) document. Wizards were allowed to use external knowledge sources and provide URL links as references. However, the specific passages that served as the grounding for these external resources are not easily traceable.

In the upcoming sections, we discuss the limitations of our studies (Section 6.1) before continuing to interpret the findings in relation to the existing literature (Section 6.2) and discuss their implications for the practical design of conversational cooking assistants (Section 6.3) and beyond (Sections 6.4 and 6.5).

6.1 Limitations

Before interpreting what the contributions mean for the conversational assistance literature and the design of these systems, it is important to acknowledge the limitations of our approach and how it was implemented.

In our simulated study, participants did not physically engage in cooking the meals; rather, they followed the outlined steps and posed questions that naturally arose as they envisioned carrying out each step. The choice to conduct a simulated, or "hypothetical," cooking process raises important implications and limitations that warrant discussion. We did not ask participants to physically cook the recipes due to numerous reasons. First, recruitment for this study was already challenging and was compounded by the need for English-speaking participants, in various demographics, which was particularly challenging, since some of the authors are situated in non-English-speaking countries. Using Prolific enabled us to access participants.

Mandating participants to physically engage in cooking would have added significant complexity and expense to the experiments. Verifying whether participants were genuinely cooking would have posed a challenge. Additionally, this approach would likely introduce numerous additional variables related to the environment, equipment, and other factors, making it impractical to maintain control. Moreover, it would have made experiments much more time-consuming and involved covering the costs of ingredients, cleaning, and additional travel. Consequently, our experiments cannot be considered naturalistic. However, there is precedence for this approach in two of the primary datasets for digital assistants in a cooking scenario [7, 26]. To increase the simulation's validity, we provided participants with images depicting the completed steps as a means of helping them imagine the process.

We acknowledge that our study assumes an optimal cooking process, devoid of external factors such as something burning or the presence of children or other distractions that could potentially hinder the success of the cooking endeavor or hamper curiosity. While this approach allows for a more controlled examination of the conversational assistance, it also limits the generalisability of

our findings to real-world, unpredictable cooking scenarios. That being said, the learning context that we envisage would be one where users have time and curiosity to ask questions as a deliberate means to further their knowledge rather than a situation whereby the aim is to cook a meal as quickly as possible or where children or friends are competing for the cook's attention. We argue that this makes our "distraction-free" scenario suitable for our research aims.

Similarly, we must acknowledge the limitations of our participant sample. Although we applied the same rules within Prolific to define the sampling strategy, the samples drawn had slightly different characteristics. We believe that the differences in the types of study, such as the interactive nature, the time taken to participate, and the scheduling of experiments, led to unavoidable sampling bias as, it was less appealing to some groups of participants. This may go some way toward explaining why the differences in expectations between younger and older participants in the survey were not reflected in the behaviours exhibited in the WoZ study. We discuss this further below.

Both samples, however, exhibited a bias towards younger individuals and older adults with a greater affinity for technology. We contend that this sample selection aligns with our research objectives. It is reasonable to anticipate that only older adults who are either technologically proficient or genuinely curious about technology would utilise such a system in practical scenarios. We can only speculate as to why more younger participants reported a higher expectation for systems to provide information on science and history. We tend to believe it has less to do with desire for learning and more to do with understanding and expectation of the technology, but this would require additional investigations to test.

Finally, we wish to acknowledge that the behaviour of the participants in our studies, both the survey and the WoZ study, will have been shaped by past experience of conversational systems, which currently do not support the answering of knowledge-related questions. This may have naturally prevented knowledge questions being asked. We argue that, regardless of past experience, this does not detract from the findings that in the active condition significantly more knowledge questions were asked and more knowledge was communicated by the agent, as we discuss in greater detail below.

6.2 Interpreting the Findings

Our goal in experimenting with initiative was to create a more human-like and engaging experience that would foster curiosity and make the participants feel comfortable asking questions. All of the metrics we studied (number of utterances, number and types of questions asked, and amount of knowledge conveyed) were increased in the active wizard condition. These findings suggest that when users are made aware that an agent has more information to share, they tend to proactively ask questions to obtain that knowledge instead of remaining passive. These findings are consistent with the survey results, which suggest that a significant number of users desire agents to offer this information.

We see parallels with the literature regarding social-bots, which have experimented with similar kinds of initiative strategies (making statements and asking questions) to make conversations more human-like and encourage users to take control [22] and share information [28]. The findings are also consistent with previous research indicating that active tutoring systems can enhance the learning process for users [12]. Furthermore, these findings help explain the limited number of knowledge-oriented inquiries (1.89% of all inquiries) in Frummet and colleagues' naturalistic investigation. Given our results, this is reasonable, because the human-agent employed a passive approach in their research. When considering the totality of the evidence—which includes Frummet and colleagues' observation of low knowledge needs, the subsequent increase in those needs under an active condition, and the survey results—it appears that users typically do not

Table 4. Naturalistic Knowledge Need Examples from Frummet et al. [19]

Knowledge Information Need	Example
Standard Amounts	"Are there always just 150 gram in a crème fraîche cup?"
Chemical Elements	"[What are] bitter substances?"
Taste	"How does bulgur taste?"
Cooking Technique	"Can you take a look at 'steaming' and what that actually means?"
Ingredient	"What is Bulgur actually?"
Consistency	"The question is what is 'slightly mushy'?"

spontaneously ask questions related to knowledge acquisition when not explicitly prompted to do so. However, they do demonstrate an interest in acquiring knowledge when they are prompted.

To put our results in context of the literature, Table 4 presents examples of knowledge information needs identified in Frummet et al.'s naturalistic study. These examples highlight the variations in knowledge information needs within a naturalistic setting. Except for a few inquiries about chemical elements, which have a scientific focus, the majority of questions centered around practical and general knowledge and "did not relate to the implementation of [the current] cooking step" [19, p. 12]. The acquired knowledge, such as standardised ingredient quantities and their desired characteristics, can be applied and utilised in future cooking sessions. In contrast to Frummet et al.'s naturalistic study, the knowledge information needs uncovered in our study delve into more specialised and detailed historical and scientific aspects of cooking. Nonetheless, the information needs related to the importance of cooking steps maintain a strong practical relevance.

6.3 Design Implications for Conversational Cooking Assistants

The findings of our study yield several design implications for conversational assistants in the context of a kitchen. We outline these implications in three sections: The first section focuses on the initiative strategy, the second section discusses the implementation of an active strategy, and the final section highlights the technical challenges associated with implementing these strategies, specifically from a retrieval perspective.

Initiative Strategy: Assistants should adapt the degree of initiative to the user.

Those users wishing to acquire background knowledge may choose for the system to perform more actively. This background knowledge proves particularly beneficial for individuals seeking to enhance their cooking abilities and expand their culinary knowledge. The presence of an active tutoring assistant aids the learning process, as highlighted by Dubiel et al. [12]. Specifically, questions relating to the significance of specific steps, such as the purpose of coating chicken in a flour mixture or the scientific principles underlying phenomena like soufflé deflation, contribute to users' comprehension of fundamental cooking processes. This understanding can then be applied in subsequent cooking sessions, facilitating further improvement in their culinary skills.

There are scenarios, however, where this kind of active strategy would be less appropriate. The survey results indicate that some users do not believe that assistants should provide such

knowledge. Moreover, users who wish to cook quickly or minimise interventions (e.g., they are in a rush or are entertaining guests or children) may prefer to interact with an agent using a passive strategy, which our results show lead to more streamlined conversations with fewer utterances by both the user and the agent.

Providing a means for users to switch between initiative modes depending on their preference and context may be a desirable feature. There are various potential approaches for implementing this feature. One option is to allow users to initiate the learning mode by instructing the system with a command like "Enter learn mode." This may be communicated to the user as follows: "Research indicates that I share more cooking knowledge when I actively engage in suggesting information at different steps. Would you prefer me to adopt this approach, or should I stick to simply answering the questions you ask, which would lead to a faster cooking process?" Alternatively, the system could proactively inquire about the user's preference for the mode. Further research can explore and compare these approaches to determine their effectiveness in terms of user experience.

Implementing an Active Strategy: The WoZ guidelines provide insight into the functionality that needs to be implemented to successfully recreate the wizard behaviour automatically.

First, on the basis of wizard experience, we recommend that interventions need to be personable, enthusiastic, and empathetic. As illustrated in Figure 7(a), participants appreciated empathetic statements such as "I'm happy to tell you this." or "Interesting, right? :-D" by saying "That's interesting. Thank you." This aligns with previous research finding that users appreciate engaging assistants [42].

Second, interventions can be in the form of questions derived from knowledge associated necessary to complete the step or in the form of statements. Our findings showed that users ask slightly (but significantly) more questions when the system prompts with a question, but there was no evidence of more knowledge being communicated as a result. Therefore, systems should be designed to use either approach in an active strategy.

Finally, in our extensive piloting where we tested various strategies, moving to the next step in a recipe was found to be an appropriate time to make interventions. This would be one way of getting around the difficulty of timing interventions as reported in the literature [2]. The wizards determined timing to be especially appropriate when extensive knowledge is available for a step compared to that communicated in the recipe instructions. However, determining this would require systems to automatically derive the mappings (see Figure 4) used by wizards, which we curated by hand. This is a further open problem for future research.

Federated Search Problem: To successfully answer questions, an agent must generate utterances using various sources of knowledge, regardless of the initiative strategy they employ. We imagine a testing framework for retrieving and formulating these utterances that would exhibit a similar setup to the one used CAsT [10]. The data collected in this study could form the basis of experimentation of this sort, as it includes user questions, conversational context, and answers provided by the wizard, as well as from which silo they were sourced. For researchers interested in performing such experiments, we have made the data available.[7]

Through our research, we have discovered that different types of questions are typically associated with distinct categories of information. This insight could potentially impact the weighting of different silos in retrieval experiments, which in turn relates to the classification of information needs, such as that proposed by Frummet et al. and other similar systems.

[7] https://github.com/AlexFrummet/cooking-with-conversations

6.4 Beyond the Cooking Domain

While we exercise caution in avoiding over-interpretation of our findings, which are derived from a study specifically focused on supporting knowledge acquisition in a cooking scenario, we find no compelling reason why many of the insights uncovered in our research would not extend to other task-based conversational assistance contexts. For instance, it seems plausible that employing an active strategy could facilitate the communication of more knowledge in analogous task-based scenarios, such as DIY projects or the repair of items such as coffee machines, bikes, or cars.

The evidence indicates that users engaged in these types of tasks are inclined to broaden their knowledge rather than solely focusing on accomplishing the primary goal of task completion. Choi et al. [6] investigated procedural tasks, including activities such as cooking, DIY projects, and learning new skills. Their findings revealed that, while the majority of individuals sought step-by-step instructions, approximately 10% expressed a desire for additional background knowledge that was not strictly necessary for completing the task. We find it plausible that if users want such supplementary information and agents make it apparent via an active strategy that they can provide it, then users are more likely to seek and engage with it, as was the case in our cooking investigation. Additional factors reinforcing our confidence in the generalisability of our findings include parallels with other learning contexts, such as tutoring, where active strategies have demonstrated support for the learning process as well as insights from the social bots literature indicating that personal disclosure, coupled with the "disclosure-reciprocity effect" [9], leads to users sharing more information than they normally would [28]. These aspects taken with the evidence that active assistants can enhance task performance in search tasks like booking flights [11] seem to paint a consistent picture. Our interests are focused on the cooking domain, but we would encourage other scholars to verify our suspicions empirically. It is clear, however, that, regardless of the domain, there are situations where people are more open to learning and others where users want to focus on completing the task at hand. We, as a research community, know little about this and the decisive factors, and this represents a challenging but important research direction for the future.

6.5 Studying Conversational Interaction—Lessons Learned

Through our studies and extensive piloting, we gleaned valuable insights that have implications for the future of **Conversational User Interface (CUI)** testing. While the concept of **Wizard of Oz (WoZ)** testing is not new, our approach introduced novel aspects. We utilised Prolific to recruit participants for offline experiments, employed TaskMAD to facilitate and regulate wizard interaction, and incorporated images to enhance the simulation. Our takeaways from this process could prove useful to other scholars, extending beyond the realm of cooking studies. We summarise these below.

— Employing two wizards proved beneficial, since it shared the workload and facilitated discussion and shared understanding between the wizards. This necessitated a clear strategy not only for intervention methods but also for determining optimal timing. Conducting numerous pilot experiments was essential to derive robust wizard behaviour and shorten response times.

— The use of TaskMAD as a platform offered advantages. Its button-based interface streamlined social interaction, reducing wizard response times and promoting smoother engagement. The search interface allowed quick access to background knowledge from multiple silos and, since frequently used information could be represented by buttons, this reduced response time significantly.

— It is our impression that the use of images and video is a great means to both improve interaction and realism in chat simulations. We are unaware of any other study that has done this.
— This study represented our first experience of using Prolific to recruit for a study of this type, i.e., where interaction with the experimenter is necessary, and we are unaware of any other study that has done this. Prolific is certainly not designed for this purpose, but it offered us access to a heterogeneous pool of suitable participants. Participant recruitment and retention was far lower than for previous Prolific studies, which led to data collection over several weeks.

Despite these insights, running Wizard-of-Oz studies remains difficult, time-consuming, and requires expert task knowledge. It may be interesting to consider how LLMs with task knowledge might be used to augment or support wizards in future studies.

7 CONCLUSIONS AND FUTURE WORK

In this article, we have presented the results of two empirical studies aimed at shedding light on how users interact with digital assistants in a kitchen context. Our first study, a survey of 200 participants, revealed that users generally expect assistants to provide information on cooking steps and processes. However, we found that younger participants who enjoy cooking were more likely to expect assistants to provide information on the history of food or the science behind cooking processes.

Our second study was a follow-up Wizard-of-Oz experiment with 48 participants, in which we compared the effectiveness of an active wizard policy versus a passive wizard policy. We found that the active policy led to almost double the number of conversational utterances, 1.5 times more knowledge-related user questions, and 1.7 times more knowledge communicated than the passive policy. These findings suggest that providing users with proactive guidance and information can lead to a more engaging and productive interaction with digital assistants in the kitchen.

Overall, our results have important implications for the design and use of digital assistants in a kitchen context. Specifically, our findings suggest that assistants should be designed to offer proactive guidance and information to users, especially younger users who are more interested in the science and history of food. We believe the data collected in our study provide a solid basis for future work. In a first step, we plan to study how existing QA and passage retrieval approaches are able to recreate the answers given by the wizards. We hope to build on these baselines by developing new approaches based on the insights from our experience as wizards. Moreover, we hope to examine ways of automating the interventions of wizards in the active condition, which includes the creation of suitable questions and statements, as well as the automated creation of how-to mappings that were valuable to determine suitable timing of interventions.

A APPENDIX

A.1 Survey

What information should a smartcooking assistant provide?
Imagine a digital assistant (e.g., Siri, Alexa, or other) that can provide you with information while you are cooking. We are interested in learning what information would be desirable to you in such a situation.

A digital assistant in the kitchen should ...

— Recommend recipes (e.g., “Ok Alexa, what should I cook tonight?”)
Strongly disagree 1 2 3 4 5 6 7 Strongly agree

— Help learn about the origin of the recipe and its development (e.g., "Where does Duck à l'Orange originate?")
Strongly disagree 1 2 3 4 5 6 7 Strongly agree
— Help learn about the science behind cooking processes (e.g., "What happens to sour cream when it is heated?")
Strongly disagree 1 2 3 4 5 6 7 Strongly agree
— Help to adapt a recipe to my (dietary) needs and preferences
Strongly disagree 1 2 3 4 5 6 7 Strongly agree
— Explain how and why a step in the recipe is important
Strongly disagree 1 2 3 4 5 6 7 Strongly agree
— Inform about ingredients and quantities needed for a recipe (e.g., "Which ingredients do I need?" "How many potatoes should I use?")
Strongly disagree 1 2 3 4 5 6 7 Strongly agree
— Inform about the equipment/cooking utensils to use (e.g., "Can I use a pot for this?")
Strongly disagree 1 2 3 4 5 6 7 Strongly agree
— Inform about the temperature at which ingredients/meals should be cooked (e.g., "At which temperature?" "Do I need to preheat the oven?")
Strongly disagree 1 2 3 4 5 6 7 Strongly agree
— Inform about the time required until the meal is prepared (e.g., "How long does it take? 10 minutes or 20 minutes?")
Strongly disagree 1 2 3 4 5 6 7 Strongly agree
— Help learn the cooking techniques required by the recipe (e.g., "OK how do you prepare potatoes properly?")
Strongly disagree 1 2 3 4 5 6 7 Strongly agree
— Guide through the process of preparing the recipe (e.g., "What should I do next?")
Strongly disagree 1 2 3 4 5 6 7 Strongly agree
— Provide suggestions about complementary dishes (e.g., "Which desserts go with chili?")
Strongly disagree 1 2 3 4 5 6 7 Strongly agree

A few questions to help us understand who has answered our survey

— How confident do you feel about being able to cook from raw or basic ingredients?
Extremely Confident 1 2 3 4 5 6 7 Not confident at all
— How confident do you feel about following a simple recipe?
Extremely Confident 1 2 3 4 5 6 7 Not confident at all
— How confident do you feel about preparing and cooking new foods and recipes?
Extremely Confident 1 2 3 4 5 6 7 Not confident at all
— How often do you prepare and cook a main meal using raw ingredients (for example, cooking soup using fresh vegetables, or cooking chili using raw meat and fresh vegetables)?
Daily 4–6 times a week 2–3 times a week Once a week Less than once a week Never

— *To which extent do you agree with the following statement?*
I enjoy cooking.
Strongly disagree 1 2 3 4 5 6 7 Strongly agree

— Gender
Male Female Other
— Age (dropdown w/ age steps)
18–24 25–34 35–44 45–54 55–64

— How would you describe your current employment status? XXXX
— What is the highest degree or level of education you have completed?
Less than high school High school graduate (includes equivalency) Bachelor's degree Master's degree Ph.D. or higher Vocational Education
— How often do you use a smart assistant such as Alexa, Siri, Google Home, or other?
Daily 4–6 times a week 2–3 times a week Once a week Less than once a week Never

A.2 Recipes

Table 5. List of Recipes Used in Our Experiments

Recipe Name	URL
Parisian Gnocchi	https://www.seriouseats.com/parisian-gnocchi-recipe
Buttermilk-brined Southern Fried Chicken	https://www.seriouseats.com/the-food-lab-southern-fried-chicken-recipe
Duck à l'orange	https://www.seriouseats.com/duck-a-lorange
Savory Cheese Soufflé	https://www.seriouseats.com/savory-cheese-souffle
Pesto alla Genovese	https://www.seriouseats.com/best-pesto-recipe
Old-fashioned Apple Pie	https://www.seriouseats.com/bravetart-easy-apple-pie-recipe

REFERENCES

[1] Mohammad Aliannejadi, Leif Azzopardi, Hamed Zamani, Evangelos Kanoulas, Paul Thomas, and Nick Craswell. 2021. Analysing Mixed Initiatives and Search Strategies during Conversational Search. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM'21)*. Association for Computing Machinery, New York, NY, USA, 16–26. https://doi.org/10.1145/3459637.3482231

[2] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. 2018. SearchBots: User engagement with chatbots during collaborative search. In *Proceedings of the Conference on Human Information Interaction & Retrieval (CHIIR'18)*. Association for Computing Machinery, New York, NY, 52–61. DOI: https://doi.org/10.1145/3176349.3176380

[3] Leif Azzopardi, Mohammad Aliannejadi, and Evangelos Kanoulas. 2022. Towards building economic models of conversational search. In *Advances in Information Retrieval*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, 31–38.

[4] Sabrina Barko-Sherif, David Elsweiler, and Morgan Harvey. 2020. Conversational agents for recipe recommendation. In *Proceedings of the Conference on Human Information Interaction and Retrieval (CHIIR'20)*. Association for Computing Machinery, New York, NY, 73–82. DOI: https://doi.org/10.1145/3343413.3377967

[5] Ethan A. Chi, Ashwin Paranjape, Abigail See, Caleb Chiam, Trenton Chang, Kathleen Kenealy, Swee Kiat Lim, Amelia Hardy, Chetanya Rastogi, Haojun Li, Alexander Iyabor, Yutong He, Hari Sowrirajan, Peng Qi, Kaushik Ram Sadagopan, Nguyet Minh Phu, Dilara Soylu, Jillian Tang, Avanika Narayan, Giovanni Campagna, and Christopher Manning. 2022. Neural generation meets real people: building a social, informative open-domain dialogue agent. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Association for Computational Linguistics, Edinburgh, 376–395. DOI: https://doi.org/10.18653/v1/2022.sigdial-1.37

[6] Bogeum Choi, Jaime Arguello, and Robert Capra. 2023. Understanding procedural search tasks "in the Wild." In *Proceedings of the Conference on Human Information Interaction and Retrieval (CHIIR'23)*. Association for Computing Machinery, New York, NY, 24–33. DOI: https://doi.org/10.1145/3576840.3578302

[7] Jason Ingyu Choi, Saar Kuzi, Nikhita Vedula, Jie Zhao, Giuseppe Castellucci, Marcus Collins, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. 2022. Wizard of tasks: A novel conversational dataset for solving real-world tasks in

conversational settings. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, 3514–3529. Retrieved from https://aclanthology.org/2022.coling-1.310

[8] Michael J. Cody, Deborah Dunn, Shari Hoppin, and Pamela Wendt. 1999. Silver surfers: Training and evaluating Internet use among older adult learners. *Commun. Educ.* 48, 4 (1999), 269–286.

[9] Nancy L. Collins and Lynn Carol Miller. 1994. Self-disclosure and liking: A meta-analytic review. *Psychol. Bull.* 116, 3 (1994), 457.

[10] Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. CAsT-19: A dataset for conversational information seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[11] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, Damien Anderson, and Sylvain Daronnat. 2020. Conversational strategies: Impact on search performance in a goal-oriented task. In *Proceedings of the 3rd International Workshop on Conversational Approaches to Information Retrieval*.

[12] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, Matthew Aylett, Miriam Wester, and David A. Braude. 2018. Improving conversational dynamics with reactive speech synthesis. In *Proceedings of the Voice-based Conversational UX Studies and Design Workshop*.

[13] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, and Sylvain Daronnat. 2018. Investigating how conversational search agents affect user's behaviour, performance and search experience. Retrieved from https://pureportal.strath.ac.uk/files/81774688/Dubiel_etal_CAIR_2018

[14] Mateusz Dubiel, Martin Halvey, Pilar Oplustil Gallegos, and Simon King. 2020. Persuasive synthetic speech: Voice perception and user behaviour. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI'20)*. Association for Computing Machinery, New York, NY. DOI: https://doi.org/10.1145/3405755.3406120

[15] M. A. El-Dosuky, Magdi Zakria Rashad, T. T. Hamza, and A. H. El-Bassiouny. 2012. Food recommendation using ontology and heuristics. In *Proceedings of the International Conference on Advanced Machine Learning Technologies and Applications*. Springer, 423–429.

[16] David Elsweiler, Alexander Frummet, and Morgan Harvey. 2020. Comparing Wizard of Oz & observational studies for conversational IR evaluation. *Datenbank-Spektrum* 20, 1 (2020), 37–41. DOI: https://doi.org/10.1007/s13222-020-00333-z

[17] David Elsweiler, Hanna Hauptmann, and Christoph Trattner. 2022. Food recommender systems. In *Recommender Systems Handbook*. Springer, 871–925.

[18] Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2019. Detecting domain-specific information needs in conversational search dialogues. In *Proceedings of the 3rd Workshop on Natural Language for Artificial Intelligence*.

[19] Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2022. "What can I cook with these Ingredients?"—Understanding cooking-related information needs in conversational search. *ACM Trans. Inf. Syst.* 40, 4, Article 81 (Jan. 2022), 32 pages. DOI: https://doi.org/10.1145/3498330

[20] Souvick Ghosh, Satanu Ghosh, and Chirag Shah. 2023. Toward Connecting Speech Acts and Search Actions in Conversational Search Tasks. (2023). arXiv:cs.HC/2305.04858

[21] David Graus, Paul N. Bennett, Ryen W. White, and Eric Horvitz. 2016. Analyzing and predicting task reminders. In *Proceedings of the Conference on User Modeling Adaptation and Personalization*. 7–15.

[22] Amelia Hardy, Ashwin Paranjape, and Christopher D. Manning. 2021. Effective social chatbot strategies for increasing user initiative. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 99–110.

[23] Jan-Gerrit Harms, Pavel Kucherbaev, Alessandro Bozzon, and Geert-Jan Houben. 2018. Approaches for dialog management in conversational agents. *IEEE Internet Comput.* 23, 2 (2018), 13–22.

[24] Morgan Harvey, Bernd Ludwig, and David Elsweiler. 2013. You are what you eat: Learning user tastes for rating prediction. In *Proceedings of the International Symposium on String Processing and Information Retrieval*. Springer, 153–164.

[25] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 159–166.

[26] Yiwei Jiang, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2023. CookDial: A dataset for task-oriented dialogs grounded in procedural documents. *Appl. Intell.* 53, 4 (2023), 4748–4766. DOI: https://doi.org/10.1007/s10489-022-03692-0

[27] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174. Retrieved from http://www.jstor.org/stable/2529310

[28] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I hear you, I feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[29] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957* (2017).

[30] Elnaz Nouri, Robert Sim, Adam Fourney, and Ryen W. White. 2020. Step-wise recommendation for complex task support. In *Proceedings of the Conference on Human Information Interaction and Retrieval (CHIIR'20)*. Association for Computing Machinery, New York, NY, 203–212. DOI: https://doi.org/10.1145/3343413.3377964

[31] David G. Novick and Stephen Sutton. 1997. What is mixed-initiative interaction. In *Proceedings of the AAAI Spring Symposium on Computational Models for Mixed Initiative Interaction*.

[32] Katherine E. Olson, Marita A. O'Brien, Wendy A. Rogers, and Neil Charness. 2011. Diffusion of technology: Frequency of use for younger and older adults. *Ageing Int.* 36, 1 (01 03 2011), 123–145. DOI: https://doi.org/10.1007/s12126-010-9077-9

[33] Andrea Papenmeier, Alexander Frummet, and Dagmar Kern. 2022. "Mhm..."—Conversational strategies for product search assistants. In *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR'22)*. Association for Computing Machinery, New York, NY, 36–46. DOI: https://doi.org/10.1145/3498366.3505809

[34] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vassilis Plachouras, Tim Rocktaschel, and Sebastian Riedel. 2021. KILT: A benchmark for knowledge intensive language tasks. In *NAACL*.

[35] 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'18)*. Association for Computing Machinery, New York, NY, 1–12. https://doi.org/10.1145/3173574.3174214

[36] Chen Qu, Liu Yang, W. Bruce Croft, Yongfeng Zhang, Johanne R. Trippas, and Minghui Qiu. 2019. User intent prediction in information-seeking conversations. In *Proceedings of the Conference on Human Information Interaction and Retrieval (CHIIR'19)*. Association for Computing Machinery, New York, NY, 25–33. DOI: https://doi.org/10.1145/3295750.3298924

[37] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the Conference on Conference Human Information Interaction and Retrieval (CHIIR'17)*. Association for Computing Machinery, New York, NY, 117–126. DOI: https://doi.org/10.1145/3020165.3020183

[38] Peter Salovey and John D. Mayer. 1990. Emotional intelligence. *Imagin., Cognit. Personal.* 9, 3 (1990), 185–211.

[39] Ameneh Shamekhi, Mary Czerwinski, Gloria Mark, Margeigh Novotny, and Gregory A. Bennett. 2016. An exploratory study toward the preferred conversational style for compatible virtual agents. In *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, 40–50.

[40] Sosuke Shiga, Hideo Joho, Roi Blanco, Johanne R. Trippas, and Mark Sanderson. 2017. Modelling information needs in collaborative search conversations. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)*. Association for Computing Machinery, New York, NY, 715–724. DOI: https://doi.org/10.1145/3077136.3080787

[41] Alessandro Speggiorin, Jeffrey Dalton, and Anton Leuski. 2022. TaskMAD: A platform for multimodal task-centric knowledge-grounded conversational experimentation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'22)*. Association for Computing Machinery, New York, NY, 3240–3244. DOI: https://doi.org/10.1145/3477495.3531679

[42] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and alignment in information-seeking conversation. In *Proceedings of the Conference on Human Information Interaction & Retrieval*. ACM, 42–51.

[43] Christoph Trattner and David Elsweiler. 2017. Investigating the healthiness of internet-sourced recipes: implications for meal planning and recommender systems. In *Proceedings of the 26th International Conference on World Wide Web*. 489–498.

[44] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the design of spoken conversational search: Perspective paper. In *Proceedings of the Conference on Human Information Interaction & Retrieval (CHIIR'18)*. Association for Computing Machinery, New York, NY, 32–41. DOI: https://doi.org/10.1145/3176349.3176387

[45] Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. 2019. QRFA: A data-driven model of information-seeking dialogues. In *Advances in Information Retrieval*, Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra (Eds.). Springer International Publishing, Cham, 541–557.

[46] Ellen M. Voorhees. 2003. Overview of the TREC 2003 question answering track. In *Proceedings of the 12th Text REtrieval Conference (TREC'03) (NIST Special Publication)*, Ellen M. Voorhees and Lori P. Buckland (Eds.), Vol. 500-255. National Institute of Standards and Technology (NIST), 54–68. Retrieved from http://trec.nist.gov/pubs/trec12/papers/QA.OVERVIEW.pdf

[47] Alexandra Vtyurina and Adam Fourney. 2018. 5 seconds after: Exploring user actions with voice assistants in the moments after a system response. Retrieved from https://cs.uwaterloo.ca/~avtyurin/papers/5_Seconds_After.pdf

[48] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the role of conversational cues in guided task support with virtual assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'18)*. Association for Computing Machinery, New York, NY, 1–7. DOI: https://doi.org/10.1145/3173574.3173782

[49] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2023. Conversational information seeking. *Found. Trends Inf. Retr.* 17, 3-4 (2023), 244–456. DOI: https://doi.org/10.1561/1500000081

Received 26 May 2023; revised 18 January 2024; accepted 18 February 2024