QookA: A Cooking Question Answering Dataset

Alexander Frummet
University of Regensburg
Regensburg, Bavaria, Germany
alexander.frummet@ur.de

David Elsweiler
University of Regensburg
Regensburg, Bavaria, Germany
david.elsweiler@ur.de

ABSTRACT

Conversational agents have become increasingly integrated into our daily lives, including assisting with cooking-related tasks. To address these issues and supplement other datasets, we introduce QookA—a unique dataset featuring spoken queries, associated information needs, and answers rooted in cooking recipes. QookA overcomes shortcomings in existing datasets, laying the foundation for more effective conversational agents tailored to cooking tasks. This paper outlines the dataset construction process, analyzes the data, and explores research applications, providing a valuable resource to enhance conversational agents in the cooking domain.

CCS CONCEPTS

• **Information systems** → **Question answering**.

KEYWORDS

cooking, conversational search, question answering

1 INTRODUCTION

In recent years, conversational agents have become integrated into our daily routines, simplifying tasks from setting timers to promptly answering user queries. Users now seek information not only on the weather or local events but also for precise guidance during procedural tasks, such as constructing furniture or cooking a meal. Amazon's launch of the Alexa TaskBot Challenge, with a particular focus on DIY and cooking tasks due to their growing popularity [9], underscores the importance of creating adept conversational agents that can effectively handle a wide range of user inquiries within these two domains. To address such needs, models are trained using conversational question-answering datasets [e.g. 2, 7], which serve various purposes, spanning from general information seeking [2, 10] to domain-specific inquiries like travel, movies, and cooking [1]. Some datasets also delve into specific dialogue aspects, including follow-up questions [11], or focus on question answering in procedural tasks [4]. Cooking is an example of such a procedural task, involves multiple steps, and cooking-related queries are predominantly grounded in a single document: the recipe [6].

To ensure the reliability of question-answering models within an authentic cooking environment, it is imperative to have user queries that authentically mirror those generated by individuals during the actual cooking process. These queries should not only reflect the information needs that naturally arise but also document the contextual backdrop that initiates these needs. Despite the presence of some question-answering datasets in the cooking domain, these struggle to fulfil these essential criteria.

For example, the *Cookversational Search* dataset [6] captures naturalistic human-human dialogues occurring during real-life cooking interactions, providing insight into the kinds of information needs people have in these contexts and how they converse to address these. These needs encompass questions about ingredient quantities, equipment usage during cooking, cooking times, temperatures, and cooking techniques. However, as they are dialogues between humans they do not accurately represent how individuals query agents conversationally. The authors documented language aspects unlikely in human-agent dialogues.

In contrast, both the Wizard of Tasks [3] and CookDial [7] datasets were crowdsourced using a Wizard-of-Oz (WoZ) approach to attain utterances realistic to a human-agent conversation. Their main drawback, however, is that these datasets do not come close to realistically representing cooking information needs as described in Frummet et al. [6]. Moreover, no efforts have been made to simulate a cooking environment, evident in the high proportion (45.8%) of *Request Step* intents such as "What's next?". The proportion of such questions is much lower (28.71%) in real cooking scenarios [6]. Similarly, the *CookDial* dataset [7] by Jiang et al. primarily involves questions related to reading recipe instructions. This means that these collections fail to accurately replicate the distribution and, more importantly, the diversity of information needs that arise during the cooking process. The Wizard of Tasks dataset comprises questions and conversational responses generated by multiple contributors across various sessions, but it lacks annotations for grounding answers within the recipe making effective QA model development difficult since only one conversational answer is provided as ground truth. As a result, the dataset is primarily suitable for generative question answering tasks, where models must generate answers, but not for extractive question answering tasks that require finding the answer span within the conversation. Furthermore, the dataset's evaluation is constrained to embedding and n-gram based metrics, such as BERTScore, BLEU, and ROUGE. This limitation makes it challenging to determine if the correct piece of information is included in the conversational response.

To address these limitations and help foster the development of effective conversational agents in the cooking domain, we introduce **QookA**[1], a novel dataset featuring **spoken natural language queries** that align with **realistic information needs** in a **simulated cooking context**. This dataset includes queries with information need (and other) annotations, and answers that are (mostly) grounded in the recipe document.

The paper is structured as follows: Section 2 provides a detailed explanation of the dataset construction process, Section 3 presents

[1] https://github.com/AlexFrummet/QookA

basic data analysis, and in Section 4, we elucidate how our dataset can be utilized to address a variety of research questions.

2 DATASET CONSTRUCTION

To fulfil our data needs, we devise a crowd-sourced method simulating the cooking process to replicate real-world information needs. Crowdworkers were directed to envision engaging a conversational assistant similar to Alexa while following each recipe step. At each step, they created questions about the highlighted information in the recipe text, supported by visual cues from recipe images. Questions were recorded via a web interface and auto-transcribed. We will detail our approach below.

2.1 Strengths of our approach

Our dataset construction approach offers three key advantages over the WoZ approaches used in the datasets discussed in Section 1. First, our dataset reflects **realistic information needs**. To achieve this, we were guided by Frummet et al.'s taxonomy for cooking information needs [6] and collected queries reflecting the breadth of need types, which is not the case with existing datasets.

Secondly, we tasked participants with creating **spoken questions** to highlighted answers that they might ask in the context of the current cooking step when following the recipe. Existing datasets predominantly consist of written questions that lack the contextual connection to a cooking step. This makes them less representative of spoken conversational assistance scenarios, such as those that occur in a kitchen with a voice assistant, where queries are always issued within the context of the current step a user is in.

Thirdly, in contrast to existing datasets, we situated our queries in a simulated **cooking context** that was both conversational and visual. To ensure participants clearly understood the context, they were guided step by step throught the recipe. By highlighting answers within the step descriptions, our participants were able to generate contextualized questions that were directly related to the cooking process. Additionally, we incorporated images illustrating the expected outcome of each step, providing some visual context of the cooking process. These images depict what the participants would see upon completing each step.

2.2 Data Collection Tool

A React-based data collection tool, as shown in Figure 1, facilitated the data collection process.

2.2.1 Information Need-based word highlighting. To achieve our aims, we implemented a rule-based named entity recognition approach that leveraged cooking wiki lists and spaCy's Entity Matcher[2] [8] to extract and highlight words related to cooking, e.g., ingredients, verbs, equipment etc. Different words were randomly highlighted to provoke information needs of different types. For instance, when we highlighted the word "butter" in the screenshot, we anticipated queries like "Apart from water, what else do I need to add to the saucepan?" – representing an information need related to **ingredients** according to [6]. For the highlighted "wooden spoon" in the screenshot we expected queries like "What should I stir it with?" – an equipment information need according to [6]. If a quantity, e.g., "150 grams" was highlighted, we expected a question representing an Amount information need and so on. In the provided screenshot, you can see that at each step, three words are highlighted. These highlighted entities were strategically chosen to elicit questions aligning with the classes in the taxonomy [6].

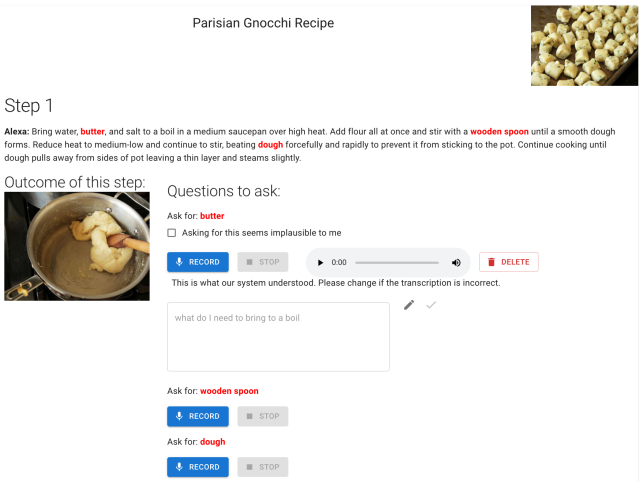

Figure 1: Screenshot of the Data Collection Tool

2.2.2 Spoken User Queries with the Recorder. Spoken user queries were recorded using a widget in the web interface. The tool, illustrated in Figure 1, allowed participants to start, stop, delete, and review their recordings. Google's Speech-to-Text API transcribed these automatically and participants could verify and correct any transcription errors as appropriate.

2.3 Recipe Selection and Parsing

Fifty recipes from *SeriousEats*[3] were selected based on specific criteria, ensuring step-by-step images for essential cooking context. A minimum of three steps per recipe was required to ensure detailed and context-rich conversations. To avoid meal or diet biases, we prioritized diversity, encompassing various courses and diets such as vegan, vegetarian, and meat recipes. The chosen recipes were obtained using a custom-built Python crawler.

2.4 Study Protocol

Before beginning data collection, the initial page of our data collection tool provided participants with task instructions and examples. Participants then gave informed consent and completed a survey to provide demographic information and details about their everyday use of conversational agents. The informed consent detailed the study's objectives, processes, and participant rights. After the survey, participants accessed a page similar to the one shown in Figure 1, where they recorded questions matching the answers highlighted within the recipe steps. Upon completing all the recipe steps, we expressed our gratitude and provided the payment code.

[2] https://spacy.io/usage/rule-based-matching

[3] https://www.seriouseats.com/

2.5 Participants

94 participants were recruited via Prolific, each receiving ≈ GBP 4 per experiment. Participants had an average age of 35 years ($min = 19.00, x_{.25} = 24.50, \tilde{x} = 33.00, x_{.75} = 40.50, max = 72.00, sd = 12.28$). 73% of our participants identified as female, 26% as male, and 1% as non-binary. In terms of educational background, the majority held academic degrees: 55.79% had bachelor's degrees, 21.05% had master's degrees, and 1.05% had PhDs. Additionally, 20% had high school diplomas, and 2.11% had vocational education. In terms of smart assistant usage, most participants indicated frequent use: 13.68% used them several times a day, 28% multiple times a week, and 26.23% less than once a week. Additionally, 17.89% used smart assistants less than once a month, and 13.68% had no prior experience with smart assistants. The primary purposes for using smart assistants were searching the web (50.52%), setting timers or alarms (48.42%), and checking the weather (37.89%).

2.6 Annotation

To ensure thorough information need annotations, we applied labels according to [6] and categorised all user queries. We also annotated whether the question required reasoning to answer which was the case, e.g., in yes/no questions (see third example in Table 1) or questions that require multiple knowledge sources to be answered, and whether the answer has to be derived from external knowledge sources. Each question was linked to the surrounding sentence where the answer could be located, the corresponding step text, the step number, and the related recipe. To accomodate multiple answers for some questions, we included a column for alternative responses. Representative examples can be found in Table 1.

2.7 Data Cleaning

In our pursuit of dataset quality, we took several post-processing steps. First, we manually reviewed all transcriptions, correcting any transcription errors in questions by re-listening to the recordings. We systematically assessed the entire dataset to ensure that each question aligned with the anticipated information need and corresponded to the answer, exemplified by the highlighted word in Figure 1. When discrepancies were identified, we either assigned the correct information need or adjusted the expected answer to match the posed question. To ensure question realism, we filtered out "implausible questions". These encompassed questions that were excessively specific (e.g., Q: "What weight of **dry ingredients** do I need?" A: "650g" – as knowing there are "dry ingredients" means the user already reasoned about a step), commands rather than questions (e.g., "Add a large chicken to my shopping basket!"), queries demanding a fundamental knowledge of kitchen equipment (e.g., Q: "How do I achieve a medium-low heat?" – assuming the user should understand stove settings), or implausible questions (e.g., Q: "What should I marinate for this recipe?" A: "chicken" – when the recipe name is "Yogurt and Mint Marinated Chicken")

3 QOOKA ANALYSIS

Our data collection process yielded 1268 utterances (see Table 2).

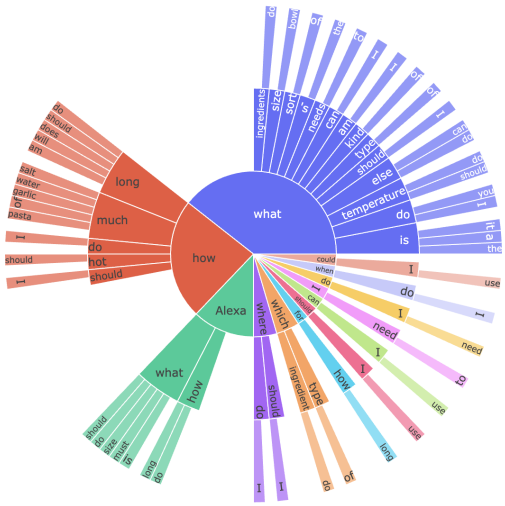

Figure 2: Distribution of the 50 most common trigram prefixes of questions in QookA.

3.1 Information Need Distribution

The information need annotation process, described in Section 2.6, identified nine distinct cooking information needs based on Frummet's taxonomy [6]. In their in-situ study, they identified a total of 11 different need types. The two absent categories include queries related to recipe search and miscellaneous queries that are somewhat unrelated to the actual cooking process. Hence, our dataset encompasses queries that represent a comprehensive range of information needs one might encounter when cooking from a given recipe. This diversity in information needs sets *QookA* apart, offering a more extensive variety compared to Wizard of Task and CookDial. Additionally, the queries in our dataset build upon each other, forming a sequence related to different recipe steps.

3.2 Linguistic Analysis

Table 2 demonstrates how QookA compares to prominent existing datasets. We present fundamental metrics, such as the number of utterances per conversation, and the number of tokens per question and answer (or utterance in the case of CookDial). These analyses reveal that our sessions were similar in length to previously released collections. Figure 2 offers a more in-depth exploration of the linguistic diversity in our dataset. This figure provides an overview of the 50 most common question trigrams in our dataset, showcasing a wide range of question formulations and underlining the linguistic richness of our dataset.

4 POTENTIAL APPLICATIONS

Our dataset supports diverse use cases, including developing effective QA models and exploring linguistic aspects of user utterances.

4.1 Developing effective QA Models

As mentioned in the introduction, datasets like QuAC [2], CoQA [10], and others can be used for fine-tuning question answering models. In the cooking domain, the CookDial and Wizard of Tasks

Question	Answer	Alt. Answer(s)	Info. Need	Reasoning	Internal Knowledge
How much salt should I use?	1 teaspoon	–	Amount	false	true
What temperature do I need to heat it?	medium-high heat	medium-high	Temperature	false	true
Do I add tomatoes to the pita bread?	Yes.	–	Ingredient	true	true
How do I transfer dough onto the metal sheet?	unanswerable	–	Cooking Technique	false	false

Table 1: Examples of annotated turns.

	QookA	CookDial	Wizard of Tasks
questions	1268	9068†	7908†
conversations	94	260	272
tokens/question	8.63	11.1†	14.2
tokens/answer	3.13	NA	18.5
utterances/conv.	26.98	34.9	29.1
% yes/no	2.68	NA	NA
% unanswerable	7.57	NA	NA
% answers grounded in document	85.88	NA	NA

Table 2: Statistics summarising the QookA dataset with values for CookDial and Wizard of Tasks for comparison where available. †indicates that the values were calculated for both question and answers as other statistics are unavailable.

Information Need	N	%
Ingredient	503	39.67
Equipment	287	22.63
Time	165	13.01
Preparation	98	7.73
Amount	93	7.33
Temperature	93	7.33
Cooking Technique	22	1.74
Knowledge	6	0.47
Meal	1	0.08

Table 3: Information Need distribution according to Frummet et al.'s taxonomy [6] in the QookA dataset.

datasets can be employed for the development of question answering models. For instance, Choi et al. [3] use their Wizard of Task dataset for generative question answering and explore methods to enhance question answering performance with varying amounts of context. Similarly, Jiang et al. [7] investigate the performance implications of different conversational contexts when using CookDial. Our QookA dataset can also serve these purposes, offering the additional benefit of incorporating spoken natural language queries that closely simulate interactions with a cooking assistant. Another advantage over related datasets is that the QookA queries are contextually embedded in the cooking process. Therefore, scholars can not only employ QookA to explore the influence of different context representations on QA model perfomance based on the conversational context, but also investigate if and how cooking context influences QA model performance and explicitly test models designed to exploit this.

4.2 Information Need Classification

All user queries in this dataset are annotated with the whole range of real-world cooking information needs. For instance, "How much salt should I use?" represents an Amount information need, while "How do I whisk?" concerns technical aspects of the cooking process, i.e. representing the Cooking Technique information need. Recognising and distinguishing these information needs is important for several reasons. Firstly, knowing the user information need is crucial for precise information retrieval from the appropriate knowledge source [6]. Questions related to ingredient quantities are typically found in the recipe, whereas those about cooking techniques are often not addressed within the same recipe [6]. Secondly, previous studies investigating human-human conversations in the cooking domain indicated that users might prefer different answer formulations based on the information need they have [5, 6]. This shows that effective information need detection is an important step towards engaging cooking assistant. Our QookA dataset offers all the relevant annotations to perform cooking information need detection and, subsequently, to conduct such studies. CookDial and Wizard of Task, however, do not possess these relevant annotations.

4.3 Investigating question/answer formulations

Future user studies can investigate user preferences for addressing information needs, specifically focusing on response formulation. For instance, analysing responses to the question "How much salt should I use?"—answered as "1 teaspoon"—can reveal variations in contextual expression, such as "You need to use 1 teaspoon of salt." Our dataset uniquely embeds answers in the cooking process, linked to corresponding steps, unlike related datasets. Further research can explore user preferences for answer formulations contextualised within the cooking process, a feature lacking in other datasets like CookDial/Wizard of Tasks, where queries lack a sequence related to recipe steps. Figure 2 highlights the linguistic diversity of user queries in our dataset, providing a foundation for future studies on how question framing varies based on underlying information needs and understanding the reasons for these variations.

REFERENCES

[1] Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - Accessing Domain-Specific FAQs via Conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7302–7314. https://doi.org/10.18653/v1/2020.acl-main.652

[2] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2174–2184. https://doi.org/10.18653/v1/D18-1241

[3] 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, Gyeongju, Republic of Korea, 3514–3529. https://aclanthology.org/2022.coling-1.310

[4] Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 8118–8128. https://doi.org/10.18653/v1/2020.emnlp-main.652

[5] 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 co-located with the 18th International Conference of the Italian Association for Artificial Intelligence (AIIA 2019), Rende, Italy, November 19th-22nd, 2019 (CEUR Workshop Proceedings, Vol. 2521)*, Mehwish Alam, Valerio Basile, Felice Dell'Orletta, Malvina Nissim, and Nicole Novielli (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-2521/paper-02.pdf

[6] 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 (2022), 81:1–81:32. https://doi.org/10.1145/3498330

[7] Yiwei Jiang, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2022. CookDial: A Dataset for Task-Oriented Dialogs Grounded in Procedural Documents. *Applied Intelligence* 53, 4 (jun 2022), 4748–4766. https://doi.org/10.1007/s10489-022-03692-0

[8] Moran Mizrahi and Dafna Shahaf. 2021. *50 Ways to Bake a Cookie: Mapping the Landscape of Procedural Texts.* Association for Computing Machinery, New York, NY, USA, 1304–1314. https://doi.org/10.1145/3459637.3482405

[9] NHPA. 2022. *Home Improvement Market Size in the United States from 2008 to 2025 (in Billion U.S. Dollars) [Graph].* Statista. https://www.statista.com/statistics/239753/total-sales-of-home-improvement-retailers-in-the-us/

[10] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266. https://doi.org/10.1162/tacl_a_00266

[11] Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of Natural Language Rules in Conversational Machine Reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2087–2097. https://doi.org/10.18653/v1/D18-1233