`Wizard of Tasks`: A Novel Conversational Dataset for Solving Real-World Tasks in Conversational Settings

Jason Ingyu Choi, Saar Kuzi, Nikhita Vedula, Jie Zhao, Giuseppe Castellucci, Marcus Collins, Shervin Malmasi, Oleg Rokhlenko and Eugene Agichtein
Amazon
{chojson,skuzi,veduln,zhaozjie,giusecas}@amazon.com
{collmr,malmasi,olegro,eugeneag}@amazon.com

Abstract

Conversational Task Assistants (CTAs) are conversational agents whose goal is to help humans perform real-world tasks. CTAs can help in exploring available tasks, answering task-specific questions and guiding users through step-by-step instructions. In this work, we present `Wizard of Tasks`, the first corpus of such conversations in two domains: Cooking and Home Improvement. We crowd-sourced a total of 549 conversations (18,077 utterances) with an asynchronous Wizard-of-Oz setup, relying on recipes from WholeFoods Market for the cooking domain, and WikiHow articles for the home improvement domain. We present a detailed data analysis and show that the collected data can be a valuable and challenging resource for CTAs in two tasks: Intent Classification (IC) and Abstractive Question Answering (AQA). While on IC we acquired a high performing model ($\geq$85% F1), on AQA the performance is far from being satisfactory ($\sim$27% BertScore-F1), suggesting that more work is needed to solve the task of low-resource AQA.

1 Introduction

In recent years, the way to access web information has changed from using keyword- and semantics-based search (Manning et al., 2008), to Question Answering (QA) (Chen et al., 2017) and Conversational Agents (CAs) (Radlinski and Craswell, 2017). CAs have evolved to support different types of interaction and information: there are CAs for chatting (Zhou et al., 2020) and CAs that let users interact with existing information systems to accomplish specific tasks, e.g., booking a restaurant, as in task-oriented dialogues (Bobrow et al., 1977; Wen et al., 2017).

A specific type of information humans are looking for is how to perform tasks, e.g., cooking a dish or fixing a household problem. The Web contains articles accompanied by images or videos with step-by-step instructions to perform variegate tasks. Existing CAs can be used mainly to browse tasks or to answer specific questions. However, they fail in providing a comprehensive natural conversation that includes search, context-aware QA, step-by-step instructions, and multi-modal information sharing.

The Alexa Prize TaskBot[1] Challenge (Gottardi et al., 2022) is a research challenge sponsored by Amazon to foster research on CTAs to assist humans in executing real-world tasks. The targeted tasks require multiple steps and decisions, including multi-modal (voice and visual) user experiences. The challenge includes two domains: Cooking, i.e., guiding people in preparing recipes; and Home Improvement, i.e., guiding people through common household do-it-yourself tasks such as painting a wall or pruning trees. As shown in Figure 1, CTAs should support QA capabilities on Web sources as well as selected recipes or articles, dialog management to support step-by-step instruction navigation, and multi-modal interaction.

In this paper, we present `Wizard of Tasks`[2] (WoT), the first dataset for CTAs. We collected a total of 549 conversations with $\sim$ 18K utterances in two domains with a Wizard Of Oz (WOZ) crowd-sourcing setting, where one worker is willing to execute a task, while another worker has the relevant knowledge to perform it and guides the first towards its execution. We adopted an asynchronous strategy to collect conversations, so that neither worker needs to wait for the other to respond and can multi-task better, enabling faster data collection. We present a detailed analysis of the dataset as well as experiments in two tasks: Intent Classification (IC) and Abstractive Question Answering (AQA). We show that a transformer-

[1] https://www.amazon.science/alexa-prize/taskbot-challenge
[2] https://registry.opendata.aws/wizard-of-tasks/

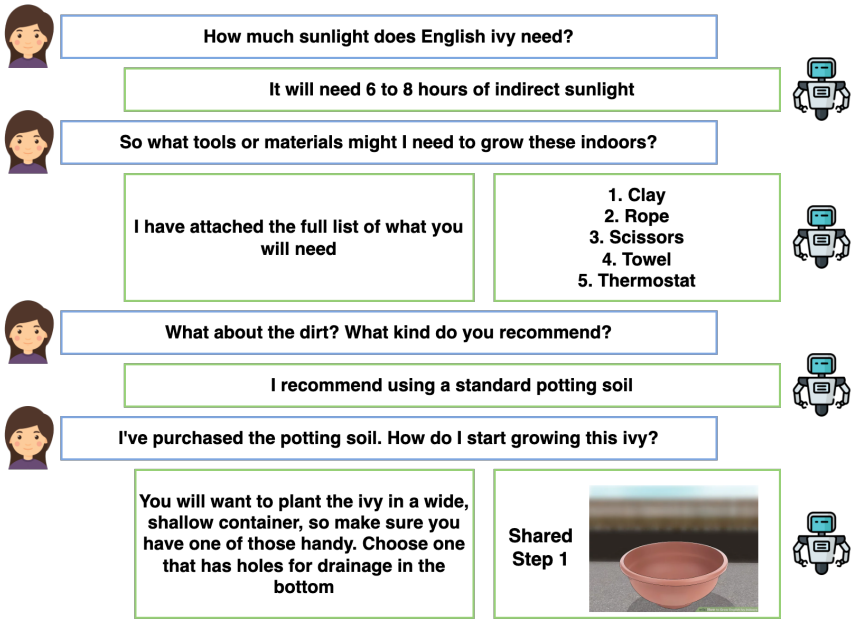

Figure 1: First few turns of one `Wizard of Tasks` conversation from the Home Improvement (DIY) domain.

based IC model can achieve ≥85% F1. In contrast, the performance on AQA is still far from satisfactory (∼27% BertScore-F1), suggesting the need for better QA models for low-resource settings.

In the rest of the paper, Section 2 discusses related work. Section 3 presents our data collection strategy. In sections 4 and 5 we discuss the data analysis and experimental evaluation. Finally, in Section 6, we state our conclusions.

2 Related Work

There is a large body of work focusing on the development of training corpora for conversational agents. In the following section, we summarize the work done to collect conversations for task-oriented and open-domain CAs.

2.1 Task-oriented Agents

The goal of task-oriented agents is to assist users in completing a task that is grounded in a knowledge base. For example, an agent can assist users in making a restaurant reservation by eliciting their preferences, building a database query and sharing the available options. While some of the previous work has focused on single-domain data sets (Moon et al., 2020; Wen et al., 2017), most of it focused on multi-domain data (Budzianowski et al., 2018; El Asri et al., 2017; Shah et al., 2018).

One of the approaches for collecting task-oriented conversations is to use computer simulations and templates to generate synthetic data sets (Shah et al., 2018; Rastogi et al., 2020; Zhao and Eskenazi, 2018), possibly also rephrasing using crowd-sourcing (Rastogi et al., 2020; Shah et al., 2018). A second approach is to let users interact with existing dialog systems (Williams et al., 2016) to improve their performance.

Perhaps the most well-known approach is to crowdsource conversations using a pool of public workers. Specifically, the WOZ paradigm is often used where one human is playing the role of a conversational agent and the other one is playing that of a user (Wen et al., 2017; Budzianowski et al., 2018; Zang et al., 2020; Eric et al., 2020; Moon et al., 2020; El Asri et al., 2017). The advantage is that the resulting data consists of natural conversations compared to using computer simulations. In some studies (Wen et al., 2017; Ikeda and Hoashi, 2018), the data was collected in an asynchronous manner by assigning each conversation turn to an available worker who writes an utterance based on all previous turns.

2.2 Open-domain Agents

Open-domain dialog agents can converse with users about topics without a clear goal. The conversations are usually grounded in some knowledge source, e.g., a Wikipedia page. The main approach for collecting data for open-domain dialog systems is to collect conversations between humans using a crowd-sourcing platform and the WOZ setting. In some studies, only one worker has access to the

knowledge source (Dinan et al., 2018). In other works, both sides have access to some knowledge source (Gopalakrishnan et al., 2019; Zhou et al., 2018; Moghe et al., 2018; Zhang et al., 2018) to simulate a scenario where two humans share some amount of background knowledge.

3 Crowd-sourcing a CTA Dataset

We designed a crowd-sourcing task to create a dataset suitable for a CTA to assist users in completing complex tasks requiring multiple steps and reasoning. We collected conversations for two target domains[3]: i) Cooking, i.e., assisting users in performing recipes and ii) Home Improvement (DIY), i.e., assisting users in performing tasks to improve their home.

3.1 Worker Roles and Expectations

Our data collection adopts a WOZ paradigm, where one worker (i.e., the student) communicates with another worker (i.e., the teacher) about tasks and how to perform them. Each worker is assigned only one of the two roles for the duration of the study to avoid potential quality issues.

3.1.1 Teacher Role

The teacher is defined as a knowledgeable expert who instructs the student to complete an assigned task, while keeping the conversation engaging and natural. The teacher is given a set of informative documents about the task to help the student.

A conversation starts with a student asking about the task. The teacher is expected to understand the document, find relevant instructions and respond to the student. The teacher can also access external resources to search for the needed information with their preferred search engine. In this case, we ask the teacher to provide the URL of any reference used to produce their response. One interesting feature in our dataset is the adoption of multi-modality to share information between the agent and the user. Creating such a multi-modal experience is not trivial, and requires more understanding on how people behave in such a setting. Thus, the teacher can share various types of content with the student regarding the task to enrich their response (e.g., step images, step text, ingredients, and tools).

Teacher Task. The teachers are required to answer four domain-independent questions for each turn before submitting their responses:

- Is the last student message relevant and coherent to the conversation history?
- Is the last student message useful for proceeding to the next steps?
- What is the action of your message?
- Write your response to the last student.

The first two questions provide relevance and usefulness labels for the previous student turn. The answer to these questions is a binary label (yes/no). The last two questions ask for the action of the teachers and their response.[4] The teachers can choose among the following actions: i) return a list of ingredients/tools; ii) return the next step; iii) answer a question only using the current task document; iv) answer a question using external knowledge (e.g., via common sense or search); v) ask a question to the student.[5]

3.1.2 Student Role

The student is defined as a curious learner who is willing to complete a task, while keeping the conversation engaging and natural. Initially, the student is given only a task title, and is expected to start the conversation by asking about the task, the required ingredients or the next steps. As the task progresses, the student is responsible for following the teacher instructions and moving toward task completion. The student is encouraged to ask at least one open-ended question about ingredients or general questions about the current step before moving to the next step. This helps to minimize the chance that the student simply requests the next step in the task without meaningful interactions.

Student Task. The students must answer six questions to submit their responses as follows:

- Is the last teacher's message relevant and coherent to the chat history?
- Is the last message of the teacher useful for proceeding to the next steps?
- Which shared content, if any, can cause potential harm to people or property damage?

[3] Whole Foods Market (https://www.wholefoodsmarket.com/recipes) for cooking tasks and Wikihow (https://www.wikihow.com/Main-Page) for DIY tasks.

[4] We disabled the copy-paste function to force the teachers to write their final response.

[5] This is to encourage teachers to ask questions to produce more natural conversations.

- What would you do in real-life if you were doing the task?
- What is the intent of your response?
- What is your response to the previous teacher?

While the first two questions are similar to the ones asked to the teacher, the third is added to label potentially dangerous content or activities provided by the teacher. The fourth question is meant to create a more realistic annotation experience. Since we do not expect workers to really do the tasks, by asking what the worker would do in real-life (e.g., "walk to the refrigerator, grab a pear and start cutting it"), we hope they can better focus on the tasks and generate more realistic sentences.

The last two questions are used to collect the intent and the message of the student. The options are: i) request ingredients/tools; ii) request next step; iii) ask a question about ingredients/tools; iv) ask a question about a step; v) stop. More details about crowdsourcing tasks are available in Appendix A.

3.2 Crowdsourced Data Collection

We used Amazon Mechanical Turk[6] to collect `Wizard of Tasks`. We paid \$0.20 for each completed task.[7] We also used an on-boarding task to filter out low-quality workers. The conversational data was collected in two batches to enable a data quality check during the process. The first batch of Cooking data (66%) was collected between Oct. 5 and Oct. 8 (2021) and the second batch between Jan. 1 and Jan. 4 (2022). The first batch of the DIY data (75%) was between Oct. 14 and Oct. 30 (2021) and the second batch between Jan. 4 and Jan. 12 (2022).

3.2.1 Asynchronous Strategy for Conversational Data Collection

Unlike previous data collection approaches for conversations (Budzianowski et al., 2018; Zang et al., 2020; Eric et al., 2020; Moon et al., 2020), we adopted an asynchronous strategy (Ikeda and Hoashi, 2018) to collect `Wizard of Tasks`. This approach has several advantages over synchronous conversation. The main advantage is that two workers are not required to be online at the same time. This allows them to work on multiple assignments simultaneously, a common practice among crowd-sourcing workers. Moreover, decoupling the workers will free them from waiting for the other party to reply, making the collection more time efficient, and thus, reducing costs for each task. As a side effect, more than two workers participate in a single conversation. This may bring more diversity into the language and writing styles as against using the same two workers.

In practice, for each turn i of a conversation, the worker (either a student or a teacher) can see some information, including the task title, the history of the conversation up to turn $i-1$, the content shared by the teachers, and, only for the teacher, the document content. The worker is expected to use the included information to understand the context and decide how to reply (i.e., to provide the i^{th} turn). After answering the required questions, the worker can submit a task. This triggers the automatic creation of a new task for the next turn $i+1$, which can be accessed by the next worker, who may be different from the previous ones.

3.2.2 Automated Quality Assurance

One disadvantage of the asynchronous strategy is that workers can abuse it.[8] We developed a heuristic to block malicious workers. When workers submitted a task, we retrieved their k=5 most recent submissions and evaluated the average relevancy and usefulness. If either value fell below a threshold (p=0.5), those workers were temporarily blocked. Each day, we manually analyzed sample submissions from temporarily blocked workers to decide whether each worker should be permanently blacklisted or unblocked. Finally, if both the average relevancy and usefulness were greater than p=0.9 in the latest k=5 submissions, we paid a bonus of \$0.08.

4 Wizard of Tasks Analysis

In this section, we present the analysis of the `Wizard of Tasks` data. Dataset statistics are presented in section 4.1, followed by a linguistic analysis in section 4.2.[9]

4.1 Dataset Statistics

The dataset consists of 272 (277) conversations in the Cooking (DIY) domain comprising 7,908

[6] https://www.mturk.com/

[7] To estimate the task price, we recruited five domain experts to work on demo tasks for 60 minutes and we computed the single task price according to a pay of \$12.5 per hour.

[8] In an early phase, we observed some workers inputting random and repetitive responses.

[9] We report additional statistics in Appendix D and conversations examples in Appendix C.

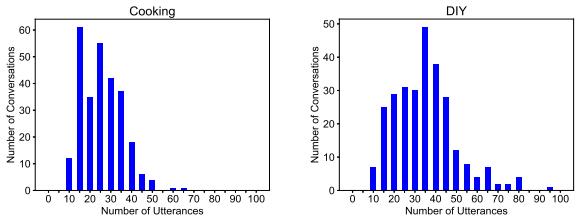

Figure 2: Conversation Length Histogram. Each point x in the x-axis corresponds to the interval $[x, x_r)$ where x_r is the closest point to the right of x.

$(10, 169)$ utterances. 238 (159) workers participated in the Cooking (DIY) experiments with an average number of utterances per worker of 33.2 (63.9). The average number of unique tasks per worker is 20.1 (35.1) for Cooking (DIY). The average number of utterances per conversation is 29.1 for Cooking and 36.7 for DIY. Figure 2 shows histograms of the conversation length. We speculate that the difference between the two domains is that DIY tasks are generally more complex and require more information for their completion.

	Cooking		DIY	
Role	Relevance	Usefulness	Relevance	Usefulness
Student	97.2%	90.2%	97.1%	90.9%
Teacher	96.5%	94.4%	97.7%	95.7%

Table 1: The percentage of student and teacher utterances that were marked as relevant or useful by crowdsourcing workers.

Utterance Quality. In Table 1, we report the percentage of utterances that were assessed positively by other workers with respect to relevance and usefulness. The results demonstrate the high quality of the collected utterances with averaged relevancy higher than 95% in both domains. Please note that teacher utterances achieved higher usefulness than student utterances. This can be attributed to the asymmetry of roles: while the student mostly asks questions, the teacher has to answer them which is more informative and thus useful.

External Resources and Shared Information. In 6.4% (3.5%) of teacher utterances for the Cooking (DIY) domain, the workers reported that they used external URLs. In 56.4% (63.4%) of teacher utterances in Cooking (DIY), the workers shared information from the recipe/article itself. A possible explanation of the difference is that recipes are generally shorter, thus teachers were willing to share more in the DIY domain.

Student Intents	Cooking	DIY
Request Step (Previous or Next)	45.8%	51.1%
Steps Question	32.7%	30.8%
Request Ingredients/Tools	13.8%	11.0%
Stop	6.8%	5.9%
Chit-chat	0.4%	1.1%
Other	0.5%	0.2%

Teacher Actions	Cooking	DIY
Return Step (Previous or Next)	49.1%	54.1%
Internal Fact Answer	17.0%	19.1%
External Fact Answer	23.7%	18.7%
Return Ingredients/Tools	6.7%	5.2%
Chit-chat	1.5%	0.7%
Other	0.2%	0.3%
Ask Question	1.8%	1.9%

Table 2: The percentage of student and teacher utterances in each intent/action type.

Intent and Action Types. In Table 2, we report the percentage of utterances for each intent/action. The two most common student intents (>78%) are requesting a step or asking a question about ongoing steps. The two common teacher actions (>66%) are either returning a step or answering a question about the given steps/tasks. The large portion of question/answer interactions attests to the complexity of the underlying tasks.

4.2 Linguistic Analysis

Utterance Length. The average number of tokens in utterances is 18.5 (21.7) and 14.2 (15.6) for teachers and students, respectively, in the Cooking (DIY) domain.[10] Figure 3 shows the full distribution of utterance length across conversations. The histogram shows different distributions for students and teachers, which we attribute again to the asymmetry of the roles.

The average number of sentences per utterance is slightly higher for teachers, i.e., 1.4 (1.4) and 1.5 (1.6) average sentences per utterance for students and teachers, respectively, in Cooking (DIY). In general, the low values demonstrate the conversational nature of the dataset. Even though teacher utterances have more words, the workers used a small number of sentences to construct them.

Linguistic Patterns Analysis. To investigate the linguistic patterns in the collected data, we ran a dependency parser on the sentences. To construct a pattern from a sentence, we first identify the child tokens of the sentence root. Then, we concatenate their dependency to generate a template.[11] Examples of the three most frequent patterns for students

[10] To perform the linguistic analysis, we used the spaCy library (https://spacy.io/).

[11] The dependencies are concatenated based on the order of the corresponding tokens in the sentence.

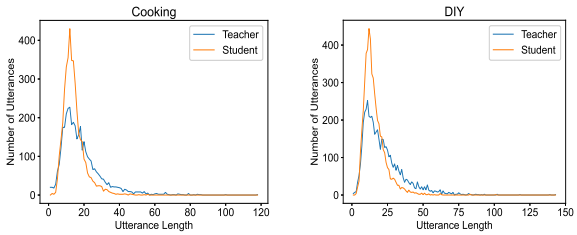

Figure 3: Utterance Length Distribution. The number of tokens in utterances of students and teachers.

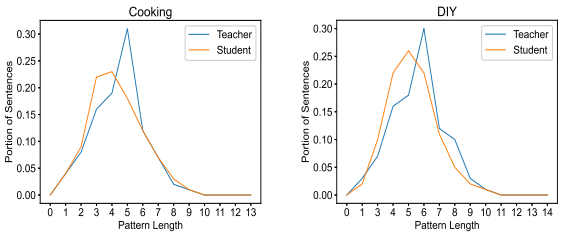

Figure 4: Linguistic Pattern Length Distribution. The portion of sentences with a specific pattern length.

and teachers are presented in Table 3 (for the DIY domain, refer to Table 11 in the Appendix).

In Table 4, we report statistics of the patterns. The results demonstrate diverse linguistic patterns in the data. In both domains, the average number of sentences per pattern (i.e., the average number of sentences expressing a pattern) is around 3. Furthermore, we can observe a slightly lower average number for teachers compared to students. A possible explanation is that teachers have access to the article/recipe from different sources which may have increased the linguistic variance of their responses. The table also shows that the average length of patterns (i.e., the average number of children of a root node) is around 4. Finally, we measured the similarity between teacher and student patterns. Specifically, the Jaccard index between the set of unique patterns of students and teachers is 0.101 (0.119) for Cooking (DIY) which shows the different language used in the different roles.

In Figure 4, we further look at the patterns' length. It shows higher variance in length in teacher utterances compared to student ones. We argue that student utterances are focused on asking for guidance and are usually not grounded in any document. On the other hand, there can be a large variety of possible responses for the teachers.

Finally, we further examined the frequency of the different patterns by computing the percentage of linguistic patterns appearing in different sentences. About 70% of the patterns appear only in a single sentence. This further demonstrates the linguistic diversity in the data. We also observe that the percentage of patterns with more than 10 appearances is substantial. These are common utterances across all conversations, such as asking for the next step or asking for tools/ingredients.

5 Experiments and Results

In this section, we experiment with two tasks: intent classification and abstractive question answering, to demonstrate the value of the `Wizard of Tasks` dataset for improving existing CTAs.

5.1 User Intent Classification

For this task, we aim to predict the intent of student utterances. Overall, there are six possible intents that students annotated according to Table 2. We will only focus on the four most common intents since we observed that the Chit-Chat and Other labels cover only about 1.0% of the data.

For modeling, we fine-tuned a pre-trained BERT (Devlin et al., 2018) model with the collected conversations. We experimented with two variants: i) encoding only the last user utterance (Utt_i); ii) encoding the last utterance along with the k previous turns ($Utt_i + Utt_{i-1}$ ... Utt_{i-k}).

With these experiments, we aim to verify the potential improvements provided by contextual information. The assumption is that the context could provide useful information in disambiguating intents, especially when ambiguous (e.g., Step Questions vs. Ingredients/Tools Questions).

In the first setting, each turn i is encoded as $[CLS]\ Utt_i\ [SEP]$. The second setting uses $[CLS]\ Utt_i\ [SEP]\ Utt_{i-1}$ ... $Utt_{i-k}\ [SEP]$.[12]

The final classification is done by applying a linear transformation layer to the encoding of the [CLS] token where intent probabilities are computed using the softmax function.

Experimental Setup and Results. To evaluate the performance of the models, we used 5-fold cross validation. We trained all models for 10 epochs with early stopping. In Table 5, we report the performance of the model as a function of k, the number of previous turns. The table includes the

[12]Notice that more conversational oriented encodings could be adopted, but this is out of the scope of this analysis.

Linguistic Pattern	Examples
Students	
dobj aux nsubj ROOT advcl punct	What do I do when the skillet is hot? | What should I do once I'm done slicing the omelet?
dobj aux nsubj ROOT advmod punct	What can I do now? | What am I doing next?
dobj aux nsubj ROOT prep punct	What do I do after heating the oil? | What should I do after my oven preheats?
Teachers	
nsubj aux ROOT dobj punct	I've included the step for reference. | You will need tofu and various seasonings.
nsubj ROOT xcomp punct	The first step is to combine ingredients in a blender. | The first step is to heat the oil.
nsubj aux ROOT xcomp punct	You will need to swirl the pan to coat grains with oil. | You are going to cover with a plate or a heavy can to drain.

Table 3: The three most frequent linguistic patterns in student and teacher utterances (Cooking); only patterns with at least two dependencies are included in the table.

	Cooking		DIY	
	Student	Teacher	Student	Teacher
# Sentences	5764	5621	7609	8244
# Unique Patterns	1606	1934	2128	2349
Avg. # Sentences per Pattern	3.6	2.9	3.6	3.5
Avg. Pattern Length	4.4	4.3	4.6	4.2

Table 4: Linguistic Pattern Statistics. "Avg. # Sentences per Pattern" is the avg. number of sentences expressing a pattern (# Sentences / # Unique Patterns).

k	Accuracy	F1				
		Macro-Avg	Ingr./Tools Question	Steps Question	Request Step	Stop
Cooking						
0	0.876	**0.867**	**0.736**	0.822	0.939	0.970
1	0.877	0.865	0.720	0.829	0.939	0.970
3	**0.879**	**0.867**	0.720	0.834	0.939	**0.974**
5	**0.879**	0.865	0.715	**0.835**	**0.940**	0.969
DIY						
0	0.863	**0.821**	**0.599**	0.805	**0.936**	0.944
1	0.857	0.812	0.579	0.797	0.931	0.942
3	0.862	0.816	0.581	0.810	0.933	0.941
5	**0.864**	**0.821**	0.592	**0.812**	0.933	**0.947**

Table 5: IC results w.r.t. the number of previous conversation turns (k). In bold, the best result in a column.

Test Domain	Cooking		DIY	
Train Domain	Cooking	DIY	DIY	Cooking
Ingr./Tools Question	0.736	0.538	0.599	0.454
Steps Question	0.822	0.804	0.805	0.771
Request Step	0.939	0.929	0.936	0.926
Stop	0.970	0.959	0.944	0.931
Macro-F1	0.867	0.807	0.821	0.771

Table 6: Comparing the performance of an intent classification model in a cross-domain setting ($k = 0$).

overall performance of the model (Accuracy and Macro-F1) as well as the F1 score of each label.

As shown in Table 5, the overall accuracy of the model is higher than 0.85 for both domains and for all values of k. The results show that the classification task is slightly more challenging for the DIY domain than Cooking. A possible explanation is the linguistic differences (e.g., the length of utterances and the number of linguistic patterns) between the two domain, as analyzed in Section 4. Surprisingly, we did not observe noticeable difference in F1 after using the conversation history, possibly due to a limited number of training conversations.

Next, focusing our attention on the per-label performance in Table 5, we can see that some labels are substantially harder to predict than others. Specifically, the F1 of Request Step is at least 12% (27%) larger than the F1 of Steps Questions (Ingredients/Tools Questions) for both domains and all values of k. A possible reason for this can be that student questions can be very specific to the task performed while other type of utterances (e.g., Request Step utterances) use language that is shared across tasks to a greater extent. Breaking out the performance by labels, we can also see that using the conversational history improves the performance of some labels and lowers the performance of others. For example, the F1 of Steps Question improves from 0.822 (0.805) to 0.835 (0.812) in Cooking (DIY) when changing the value of k from 0 to 5; an opposite trend is observed in Ingr./Tools Questions. This result suggests that the amount of conversational history used by the model should potentially vary across different types of utterances, which is an interesting direction for future work.

Finally, in Table 6, we analyze the performance of using a model that was trained on one domain to predict the intent of utterances in the other domain. The results show the overall importance of domain-specific information for the task. Still, the necessary information for the prediction of some labels is shared between domains to a greater extent than in the case of other labels. For example, while in the case of Request Step both models perform similarly in both domains, using the cross-domain model substantially deteriorates the performance of the Ingr./Tools Question prediction.

Error Analysis. We observed that our model produced the least number of errors on the Stop intent, probably because Stop utterances often contain informative keywords (e.g., "stop" and "done"). We also observed that some Request Step utterances can contain similar keywords (e.g., "I am done with this step"). This explains why sometimes the model confused a Stop intent with a Request Step intent in both the Cooking and the DIY domains.

We also observed a challenge in distinguishing between Ingredient/Tool Question intents and Steps Question intents, primarily due to some ambiguity between these question types (Table 7). For instance, the model predicted Steps Question when the user asked about *plastic bag* substitution, but predicted Ingredient/Tool Question for *zipper* substitution. Conversely, the model predicted Ingredient/Tool Question when the user asked about *edge sander*. This is because while the first part of the sentence is a substitution question, the second sentence actually links back *edge sander* to the next step. For further analysis, refer to Appendix E.

Cooking:
True: Request Step, Predicted: Steps Question
Do I need to let the cake cool first?
True: Ingr./Tools Question, Predicted: Steps Question
Does the vinegar's flavor profile change when it becomes a syrup?
True: Steps Question, Predicted: Ingr./Tools Question
Should these other ingredients be sprinkled in a specific order?
True: Steps Question, Predicted: Request Step
What do I need to do in order to get my grill prepped?
DIY:
True: Request Step, Predicted: Steps Question
Now it is planted. Should I water the plant?
True: Ingr./Tools Question, Predicted: Steps Question
Could I substitute a plastic bag if I don't have a paper bag?
True: Steps Question, Predicted: Ingr./Tools Question
Can you teach me how to replace a zipper?
I have an edge sander. Do I need it for the next step?
True: Steps Question, Predicted: Request Step
What should I do if I see a few drips of water when I run the water?

Table 7: Intent classification errors.

5.2 Abstractive Question Answering

In this task, we seek to generate natural language answers for the questions asked by student workers. The ground truth answers are provided by the teacher workers. We focus on utterances with the following student intents to prepare our dataset for this task: Step Questions and Ingredients/Tools Questions (from Table 8). We randomly sample 80% of the available question-answer (QA) pairs as training data, with the remaining 20% equally split into validation and test sets. We only use information from the available document (recipe or article) and the conversational history to generate answers to questions, without considering any external knowledge. Thus, our test set only consists of QA pairs where answers from the teacher do not contain any links to external resources.

Domain	Total # QA pairs	# Internal QA pairs
Cooking	1378	538
DIY	1435	684

Table 8: Statistics of student-teacher QA pairs.

Experimental Setup and Results. We fine-tune two state-of-the-art pre-trained language models, BART (Lewis et al., 2019) and T5 (Raffel et al., 2019) with our identified QA pairs for both domains, for the task of natural language answer generation. As input, we provide the models the user question and a context, which consist of document (recipe or article) text, a list of ingredients or tools associated with the document, and the prior conversational history. We evaluate different variations of the context using the natural language generation metrics of BLEU, ROUGE and BERT-score (Zhang et al., 2019). In the interest of space, Table 10 shows the results for the two fine-tuned models, using two turns of conversational history and the entire content of the document as input context for both domains. We also report as a baseline answering model, the document step most similar to the input question (MSS). We compute the most similar step according to the cosine similarity between the steps' and questions' representations, obtained by using Sentence-BERT (Reimers and Gurevych, 2019). BART outperforms MSS in both domains, while T5 only outperforms MSS on BERT-score in the Cooking domain. MSS performs better on DIY, than the Cooking domain. A possible explanation could be that for DIY tasks teachers seem to rely more on the document content to answer questions, whereas for cooking tasks, they are more likely to summarize or paraphrase the document content to generate answers. We also observe that the fine-tuned BART model outperformed the T5 model by about 1-13% across different context settings and domains in terms of generated answer quality. This may be due to a smaller number of allowed input tokens (512 for T5 vs 1024 for BART), since the average (standard deviation) of input context length is 286 (80) for the recipe domain and 1532 (451) for the DIY domain.

We present the BERT-score for the fine-tuned BART model in Figure 5, across different contextual settings for both domains. We observe a maximum model performance of 0.27 BERT-score F1 points for both the Cooking and DIY domains. Including conversational history as part of the input context improves the answer generation performance by 1-3% across both domains. This gain is further enhanced by 1-2% with the addition of the list of tools/ingredients to the context. In general, the inclusion of a greater number of document steps in the context contributes the most to the answer

generation quality, followed by the conversational history and the list of ingredients/tools.

The low performance (<0.3) using state-of-the-art NLG models for both domains emphasizes the challenges involved in solving tasks associated with low-resource settings and indicates that there is still a lot of room for improvement. We observed that these models commit various types of factual errors and stylistic or grammatical errors during answer generation (examples can be found in Table 9 and in Appendix F), which points out the scope of tackling these open research problems in the future. In addition, in this work we simply concatenated the different sources of information (conversational history, list of tools/ingredients, document steps) to create the input context. Evaluating different ways to generate a concise, useful context, with the possible use of some external knowledge, can also help improve the answer quality.

Error Analysis. Overall, we observed a better sentence structure and a much fewer number of grammatical or stylistic errors in the generated answers for the DIY domain, as compared to the Cooking domain (Table 9). Our model produced the largest number of factual errors which are numerical in nature, as compared to the other types of errors. This includes both hallucinating numerical terms that did not exist in the input context, as well as generating numbers or units that are different from their ground truth values. This is possibly because the pre-trained NLG language models we used are not well trained or equipped to encode and predict the numerical information. Incorporating the list of ingredients/tools in the input text when generating an answer helps reduce the erroneous generation of specific ingredients, tools or materials used in the tasks (e.g., "knife" vs. "thick spoon" in the top part of Table 9 or the definition of "cheesecloth" in the bottom part).

Cooking:
Factual, Numerical Errors → Correct Answer
Bake the cake for **15 minutes** → You should bake for 20 minutes.
Factual, Non-numerical Errors → Correct Answer
You can **turn on a stovetop** and cook until smooth. → No stovetop is required.
Grammatical/Stylistic Errors → Correct Answer
Yes can use the same knife. → You can use the same knife
DIY:
Factual, Numerical Errors → Correct Answer
Yes, it takes about **8 to 10 days** for the seeds to grow. → The sprouts take four days to grow.
Factual, Non-numerical Errors. → Correct Answer
The cheesecloth is a **thin layer of plastic**. → A cheesecloth is a loosely woven cotton cloth.
Grammatical/Stylistic Errors → Correct Answer
Yes, there are any safety concerns with opening the valve. → Open with caution in well ventilated room and no flame exposed.

Table 9: Abstractive QA errors. Erroneous terms in utterances are boldfaced.

Model	Ctxt.	Hist.	BLEU	ROUGE	BERTScore
Cooking					
MSS	None	0	0.077	0.148	0.071
T5-base	All	2	0.066	0.136	0.119
BART-base	All	2	0.116	0.211	0.270
DIY					
MSS	None	0	0.125	0.208	0.224
T5-base	All	2	0.053	0.134	0.183
BART-base	All	2	0.119	0.235	0.276

Table 10: Abstractive QA results. 'Ctxt.' and 'Hist.' refer to context and history, respectively.

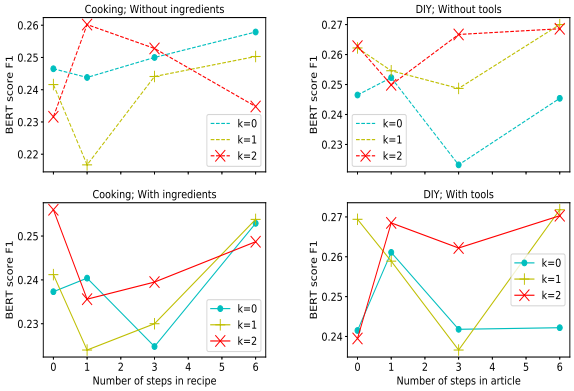

Figure 5: BERT-score of fine-tuned BART using different context settings on Cooking (left) and DIY (right). k denotes history size, while the x-axis shows the number of recipe/article steps used as part of the input context. The upper two figures (dashed line) refer to using context without ingredients/tools, whereas the lower two figures refer to using them.

6 Conclusions

In this paper, we presented `Wizard of Tasks`, a novel dataset for conversations in the area of Conversational Task Assistants. This dataset is inspired by the Alexa Prize TaskBot Challenge, where the participants have to create a dialog agent on top of the Alexa platform to help users in performing real world tasks in two domains: Cooking and Home Improvement. To the best of our knowledge, our task-oriented conversational dataset is the first of its kind in these two domains. We discussed our dataset collection asynchronous strategy in a crowdsourcing setting. We reported multiple analyses of the collected data as well as initial experimental evaluations on two tasks: Intent Classification and Abstractive Question Answering. Despite the small size of the collected data, we hope it can foster research in the novel area of CTAs. In the future, we plan to expand the dataset to additional domains.

References

Daniel G. Bobrow, Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry S. Thompson, and Terry Winograd. 1977. Gus, A frame-driven dialog system. *Artif. Intell.*, 8(2):155–173.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Layla El Asri, Hannes Schulz, Shikhar Kr Sarma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Anna Gottardi, Osman Ipek, Giuseppe Castellucci, Shui Hu, Lavina Vaz, Yao Lu, Anju Khatri, Anjali Chadha, Desheng Zhang, Sattvik Sahai, Prerna Dwivedi, Hangjie Shi, Lucy Hu, Andy Huang, Luke Dai, Bofei Yang, Varun Somani, Pankaj Rajan, Ron Rezac, Michael Johnston, Savanna Stiff, Leslie Ball, David Carmel, Yang Liu, Dilek Hakkani-Tur, Oleg Rokhlenko, Kate Bland, Eugene Agichtein, Reza Ghanadan, and Yoelle Maarek. 2022. Alexa, let's work together: Introducing the first alexa prize taskbot challenge on conversational task assistance. In *Alexa Prize TaskBot Challenge Proceedings*.

Kazushi Ikeda and Keiichiro Hoashi. 2018. Utilizing crowdsourced asynchronous chat for efficient collection of dialogue dataset. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, USA.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332.

Seungwhan Moon, Satwik Kottur, Paul A Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, et al. 2020. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121.

Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, page 117–126, New York, NY, USA. Association for Computing Machinery.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tiancheng Zhao and Maxine Eskenazi. 2018. Zero-shot dialog generation with cross-domain latent actions. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 1–10.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics*, 46(1):53–93.

A Data Collection User Interface Design

As shown in Figure 6 and 7, we developed four widgets to optimize the presentation: i) annotation guidelines; ii) chat history; iii) task content; and iv) annotation questions and answers. All of these widgets are slightly tuned to support both teacher and student roles. For example, we do not show any article/recipe content to students except for their title. On the other hand, every bit of information including a list of ingredient/tools, steps text, images, and summary are available to teachers. There are also variations in annotation guidelines and the type of questions we ask to students vs. teachers.

For chat history, we always show the entire conversation history since our data collection is done asynchronously. This is important because for anyone to contribute to an existing conversation, understanding the context is a prerequisite for generating any next response. Workers were also able to click the "Share" button to see the information that was shared on a particular turn. We developed these mechanisms to help workers to quickly figure out what is happening inside different conversations and decide the next possible response.

B Additional Details of the Data Collection

In total, 238 (159) crowd workers participated in the Cooking (DIY) experiments and the average number of utterances per worker is 33.2 (63). The average number of unique tasks per worker is 20.1 (35.1) for Cooking (DIY). Finally, we observed that 93 workers participated in both tasks. The difference between domains might be due to the higher complexity and diversity in DIY tasks compared to Cooking, which might attract fewer workers that are interested in participating in the task.

C Examples of Collected Conversations

We report in Table 18 and 19 two `Wizard of Tasks` conversations from the Cooking and the DIY domain.

D Additional Dataset Statistics

In Table 12, we report the distribution of utterance length in terms of the number of sentences. The average number of sentences per utterance is slightly higher for teachers (i.e., 1.4/1.5 and 1.5/1.7 average sentences per utterance for students and teachers, respectively, in Cooking/DIY). In general, the low values demonstrate the conversational nature of the data set: even though teacher utterances have more words, the workers used a small number of sentences to construct them.

In Table 13, we further examine the frequency of the different patterns. In the table, we can see the percentage of linguistic patterns with a specific number of sentences that they appear in. The results show that about 70% of patterns appear only in a single sentence which demonstrates the linguistic diversity in the data. We also observe that the percentage of patterns with more than 10 appearances is substantial. These patterns can be common utterances across all conversations, such as asking for the next step or asking for a list of tools or ingredients.

E Error Analysis: Intent Classification

We present the confusion matrix for both domains in Table 14 and 15 to see the performance trade-offs between different intent labels.

The tables show that our model produced the least number of errors on the Stop intent. This is expected because Stop utterances often contain informative keywords (e.g., "stop" and "done"). Interestingly, we also observed that some Request Step utterances can contain similar keywords (e.g., "I am done with this step"). This explains why sometimes the model confused a Stop intent with a Request Step intent in both the Cooking and the DIY domains.

The confusion matrices also demonstrate the great challenge in distinguishing between Ingredient/Tool Question intents and Steps Question intents. Although our intent labels were designed to capture the differences between these two types of questions, we noticed that this boundary can become ambiguous from time-to-time.

F Error Analysis: Abstractive Question Answering

In this section, we report the errors committed by our fine-tuned BART answer generation model in the abstractive question answering task, where it fails to generate the correct answer to a user's question. A sample of the different types of errors are shown in Tables 16 and 17, and the erroneously generated terms are shown in bold.

Linguistic Pattern	Examples
Students	
dobj aux nsubj ROOT prep punct	What should I do after that? | What should I do after getting rid of the clippings?
ccomp punct dobj aux nsubj ROOT advmod punct	Okay I have watered them what do I do now? | I've gathered the materials requested what do i do next?
dobj aux nsubj ROOT advmod punct	What should I do next? | What do I do now?
Teachers	
nsubj aux ROOT dobj punct	I've shared details about the mulch here. | Shade will reduce the potency of the lavender plant.
nsubj aux ROOT dobj advmod punct	I've shared the details here. | You're doing great so far!
nsubj ROOT xcomp punct	We want to water our soil until saturated. | A sauna helps to cleanse the skin and make you feel healthy.

Table 11: The three most frequent linguistic patterns in student and teacher utterances (DIY); only patterns with at least two dependencies are included in the table.

	Cooking		DIY	
# Sentences	Teacher	Student	Teacher	Student
1	63%	68%	53%	65%
2	29%	24%	32%	27%
3	5%	7%	11%	7%
>3	2%	1%	4%	2%

Table 12: The distribution of number of sentences for teacher and student utterances.

	Cooking		DIY	
# Sentences	Student	Teacher	Student	Teacher
$(0, 1]$	69%	73%	70%	71%
$(1, 10]$	26%	23%	26%	24%
$(10, \infty)$	5%	4%	4%	4%

Table 13: Pattern Frequency. The percentage of patterns that appear in a given number of sentences.

	Request Step	Ingr./Tools Question	Steps Question	Stop
Request Step	1799	14	49	11
Ingr./Tools Question	24	427	112	1
Steps Question	133	156	1046	2
Stop	2	1	0	276

Table 14: Intent classification confusion matrix (Cooking).

	Request Step	Ingr./Tools Question	Steps Question	Stop
Request Step	2538	18	101	10
Ingr./Tools Question	41	302	231	0
Steps Question	166	115	1296	11
Stop	10	0	3	292

Table 15: Intent classification confusion matrix (DIY).

Factual, Numerical Errors	**Correct Answer**
Bake the cake for **15 minutes**	You should bake for 20 minutes.
You should preheating the oven to **475F**.	It takes about 15-20 minutes to preheat the oven.
Yes, the recipe **does not include popcorn**.	Yes! You'll need 2 cups of popcorn.
Factual, Non-numerical Errors	**Correct Answer**
You can **turn on a stovetop** and cook until smooth.	No stovetop is required.
You can use a **knife**.	You will need a thick spoon.
Grammatical/Stylistic Errors	**Correct Answer**
Yes can use the same knife.	You can use the same knife
You should preheating the oven to 475F.	You should preheat the oven to 475F.

Table 16: Abstractive QA errors (Cooking). Erroneous terms in boldface.

Factual, Numerical Errors	**Correct Answer**
Yes, you will need to cut them every **two weeks**	Yes because they prevent the rhubarb from growing big and strong
Yes, it takes about **8 to 10 days** for the seeds to grow.	The sprouts take four days to grow.
Factual, Non-numerical Errors.	**Correct Answer**
The cheesecloth is a **thin layer of plastic**.	A cheesecloth is a loosely woven cotton cloth.
The soil will have to be **moist**.	The soil just needs to be very loose with chunks of bark and other organic matter
Grammatical/Stylistic Errors	**Correct Answer**
Yes, there are any safety concerns with opening the valve.	Open with caution in well ventilated room and no flame exposed.
Yes, there is no set amount of space in a small closet.	A great option is to use space bags. I've shared more information with you.

Table 17: Abstractive QA errors (DIY). Erroneous terms in boldface.

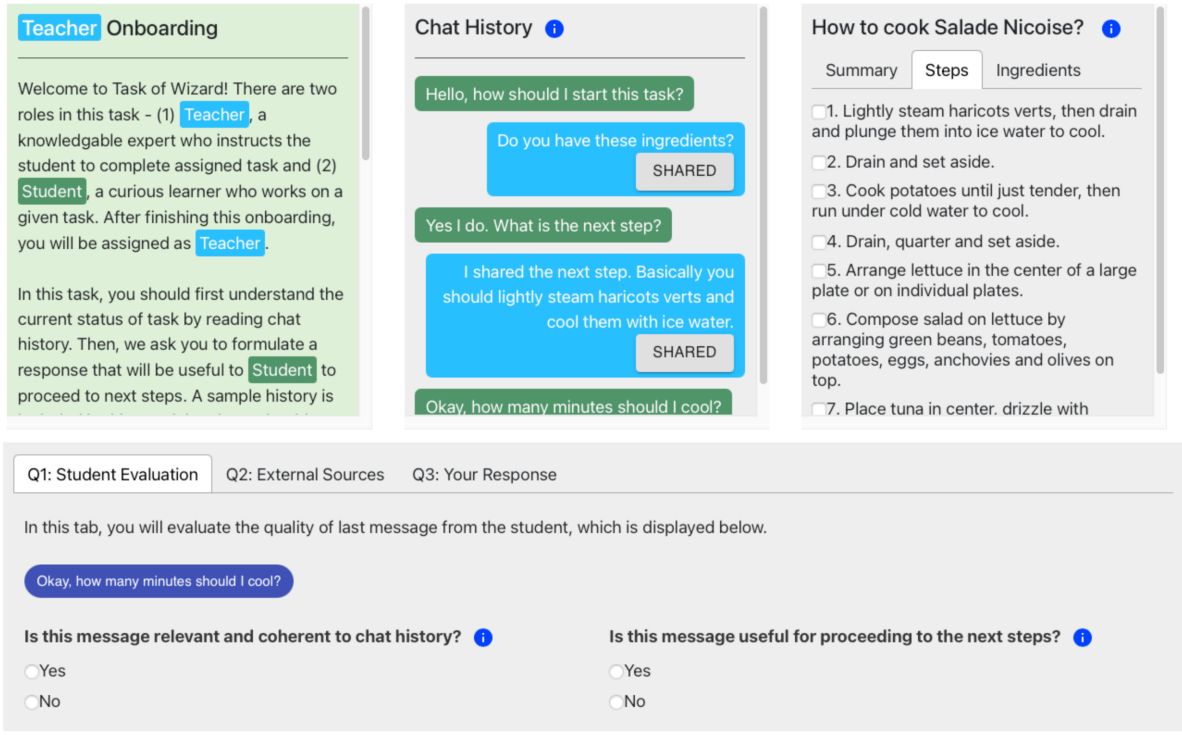

Figure 6: Teacher task user interface. Teachers are expected to find and provide relevant answers to students' questions. Teachers can also use the checkmarks in the task information section to share selected content if the textual response is not sufficient to deliver the full information.

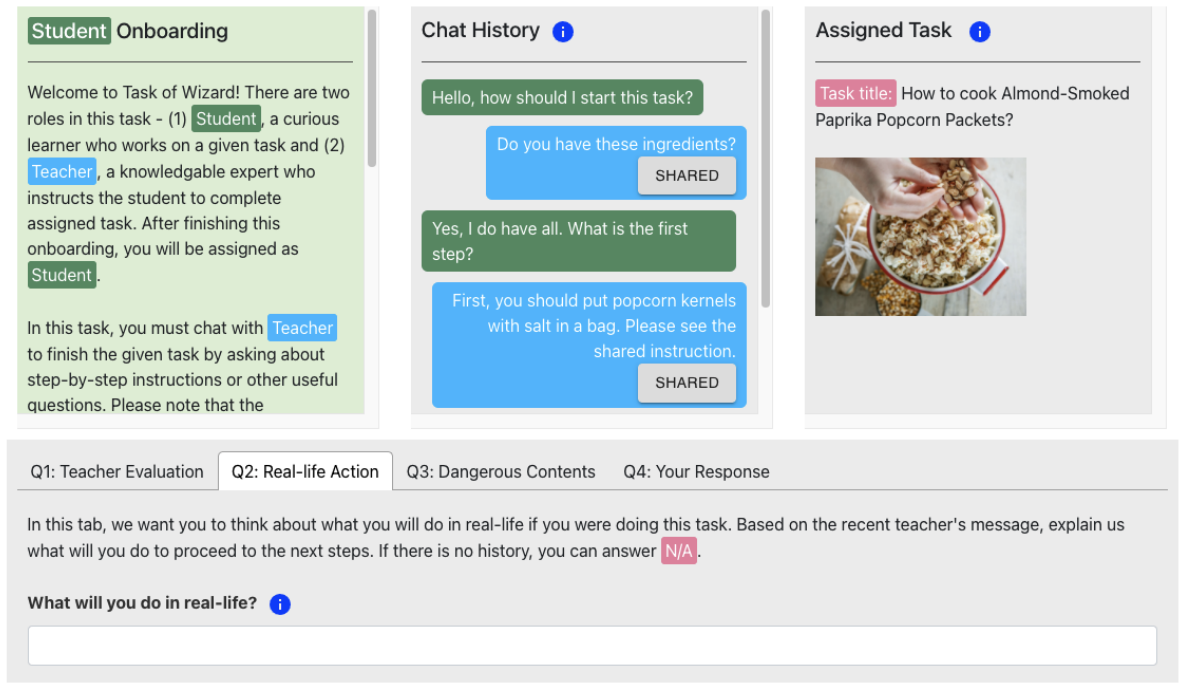

Figure 7: Student task user interface. Students are expected to understand the task status at each turn and engage with the teacher to progress toward task completion. The students can click on "Shared" button to view shared content of the current and previous turns. By default, only the task title and image are shown to the student.

	Utterance	Notes
User	*What do I need besides Salmon for this recipe?*	
Agent	*Here is what you need*	[Teacher shares some ingredients]
User	*What is the first step that should be performed after preparing these ingredients?*	
Agent	*The first step is to cut the salmon. I've shared below*	[Teacher shares a step]
User	*Now that the salmon is in the bowl, what is next?*	
Agent	*The next step is to mix it with some seasoning. I've shared the step below*	[Teacher shares a step]
User	*What specific seasonings will I need?*	
Agent	*You will need soy sauce, sesame oil, and sesame seeds.*	[Teacher shares some ingredients]
User	*The salmon and seasonings are now mixed in the bowl. What comes next?*	
Agent	*Refrigerate the mixture for a minimum of 15 minutes.* *You can leave it in for up to 4 hours for better flavor.*	
User	*Does the refrigeration time affect the texture of the mixture?*	
Agent	*It will not affect the texture, it just makes the flavors meld together better.*	
User	*Am I done? let me know please?*	
Agent	*Yes, now you're just plating and enjoying!*	[Teacher shares the last step]
User	*I have an odd question, Why is this called Poke (Salmon poke)?*	
Agent	*Poke means "cut into pieces", and is how Hawaiians first described the dish.*	[Answer found in external URL]
User	*Good to know! Thanks for the help, it's time to go eat now!*	

Table 18: Example of one full `Wizard of Tasks` conversation from the Cooking domain.

	Utterance	Notes
User	*Can you help me with cleaning of wool rugs?*	
Agent	*Sure! First you should gather all the supplies needed that I have listed.*	[Teacher shares some tools]
User	*What should I do now I have the supplies?*	
Agent	*Start with taking the rug outside and shaking it off,* *using a broom to help get off the dirt and dust.*	[Teacher shares a step]
User	*After I've shaken as much dust and dirt out of the rug,* *what do I need to do a deeper job, that doesn't seem like enough.*	
Agent	*No, that in itself won't be enough. Now run a vacuum cleaner over your rug.*	[Teacher shares a step]
User	*Ok, the rug has been shaken of dirt and dust and then vacuumed...* *is there anything else to do?*	
Agent	*Yes, you will need to shampoo your rug.* *Please follow the attached instruction for shampooing your rug.*	[Teacher shares a step]
User	*Do I need to use a sponge or can I use a cleaning scrub brush of* *sorts to get out extras stains? Is this a situation that I can* *deviate a little or do I need to be precise?*	
Agent	*A sponge is less harsh but either will work.*	
User	*I've rubbed a lot of soap through it and rinsed until the soap is gone.* *What should I do with it next?*	
Agent	*You want to dry your rug immediately*	[Teacher shares a step]
User	*Why does the rug need to be dried immediately?*	
Agent	*Because it needs extensive drying to get rid of the moisture*	[Teacher shares the last step]
User	*Squeezing it with my hands left a lot of water still in.* *Could I make a pad of dry towels on the floor, then lay the rug down,* *then put more towels on top and apply pressure to it to blot up more water?*	
Agent	*That would work you could also try hang drying the wool rug for the final step*	
User	*Okay, good to know! What should I do after drying the rug?*	
Agent	*Cleaning and drying wool rug is complete*	
User	*Could the steps that I just took for my wool rug work on my cotton rug as well?*	
Agent	*I would check the care tag on the rug before doing anything.*	
User	*What about these instructions makes this care particularly* *well-suited for wool rugs?*	
Agent	*Wool is very delicate and needs a lot of care*	
User	*Thanks for the help cleaning my rug!*	

Table 19: Example of one full `Wizard of Tasks` conversation from the Home Improvement (DIY) domain.