Simulated Task Oriented Dialogues for Developing Versatile Conversational Agents

Xi Wang[1](✉), Procheta Sen[2], Ruizhe Li[3], and Emine Yilmaz[1]

[1] Unversity College London, London, UK
{xi-wang,emine.yilmaz}@ucl.ac.uk
[2] University of Liverpool, Liverpool, UK
procheta.sen@liverpool.ac.uk
[3] University of Aberdeen, Aberdeen, UK
ruizhe.li@abdn.ac.uk

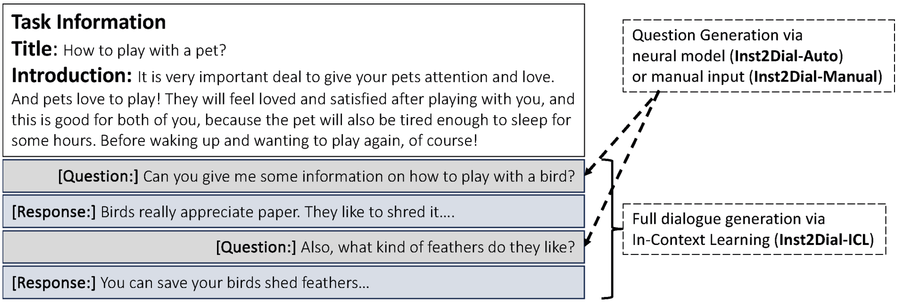
Abstract. Task-Oriented Dialogue (TOD) Systems are increasingly important for managing a variety of daily tasks, yet often underperform in unfamiliar scenarios due to limitations in existing training datasets. This study addresses the challenge of generating robust and versatile TOD systems by transforming instructional task descriptions into natural user-system dialogues to serve as enhanced pre-training data. We explore three strategies for synthetic dialogue generation: crowdsourcing, encoder-decoder models, and in-context learning with large language models. The evaluation of these approaches, based on a comprehensive user study employing 10 different metrics, reveals the top quality of the dialogues generated by learning an encoder-decoder model as per human evaluation. Notably, employing this synthetic dialogue further improves the performance of advanced TOD models, especially in unfamiliar domains, with improvements spanning 5.5% to as much as 20.9% in combined evaluation scores. Our findings advocate for the use of specialised, task-oriented knowledge bases and step-wise dialogue generation techniques to advance the capabilities and generalizability of TOD systems.

1 Introduction

Task-Oriented Dialogue (TOD) Systems have recently proven valuable in helping users with various daily tasks like restaurant bookings [10,16]. For these systems to be effective, they must understand tasks and offer relevant suggestions. A key challenge is arming TOD models with extensive knowledge for effective responses across multiple tasks. Researchers tackle this by fine-tuning models with rich, large-scale datasets [36]. However, a notable gap exists in available datasets that cover a broad range of tasks and details. Existing datasets, such as MultiWoZ [40], Frames [11], and SGD [30], mainly focus on common scenarios like travel. While SGD covers 16 domains, it lacks comprehensive instructional content crucial for task-specific adaptation. One solution is creating enriched, larger-scale datasets, but this faces two main hurdles. First, the scarcity of clean, structured, domain-specific knowledge. A notable and recent advancement in

N. Goharian et al. (Eds.): ECIR 2024, LNCS 14608, pp. 157–172, 2024.
https://doi.org/10.1007/978-3-031-56027-9_10

this regard is Task2KB [25], which compiles task instructions and a wealth of associated information from WikiHow[1] – an online platform offering detailed guides for diverse tasks. However, a second challenge remains: the significant resource investment required to develop high-quality, large-scale TOD datasets. Traditional human labelling methods are both time-consuming and often result in noisy or unreliable data [6]. Therefore, it is essential to leverage existing publicly available task-oriented knowledge bases, like Task2KB, to develop rich, reliable and diverse task-oriented dialogue datasets, so as to further benefit the advancement of TOD models.



Task Information
Title: How to play with a pet?
Introduction: It is very important deal to give your pets attention and love. And pets love to play! They will feel loved and satisfied after playing with you, and this is good for both of you, because the pet will also be tired enough to sleep for some hours. Before waking up and wanting to play again, of course!

[Question:] Can you give me some information on how to play with a bird?

[Response:] Birds really appreciate paper. They like to shred it....

[Question:] Also, what kind of feathers do they like?

[Response:] You can save your birds shed feathers...

Question Generation via neural model (**Inst2Dial-Auto**) or manual input (**Inst2Dial-Manual**)

Full dialogue generation via In-Context Learning (**Inst2Dial-ICL**)

Fig. 1. An illustrative example of synthetic dialogues that can be generated using different methodologies.

Concurrent with recent advancements in generative models, such as GPT-3.5/4 and Flan-T5, research has demonstrated the feasibility of generating high-quality dialogue data using descriptive text as input, such as Wikipedia passages [9]. In light of these findings, our study aims to transform instructional task descriptions into natural user-system dialogues. These dialogues intend to serve as training data to enhance both the quality of responses generated and the generalizability of state-of-the-art TOD models. Specifically, we systematically explored three strategies for generating synthetic data using task-related instructional information. These strategies encompass a) *step-wise dialogue generation employing encoder-decoder models*, b) *crowdsourcing* and c) *in-context learning* with large language models. In Fig. 1, we exemplify the use of the three strategies with the corresponding resulting datasets, Inst2Dial-Auto/Manual/ICL, which generate questions and use instructions as responses (Inst2Dial-Auto&Manual) or generate full dialogue at once (Inst2Dial-ICL).

To provide a comprehensive evaluation of these methodologies, we conducted both offline assessments and user studies. The evaluation results, based on a user study employing 10 different metrics-tailored to whether additional task information was provided-demonstrate that step-wise dialogue generation with encoder-decoder models consistently outperforms the other two strategies. Specifically, synthetic dialogues produced through such a strategy have been shown to enhance the performance of state-of-the-art models, especially in domains with

[1] https://wikihow.com.

limited knowledge. Improvements ranged from a minimum of 5.5% to as much as 20.9% in the combined evaluation score.

This study presents three pivotal contributions to the field of Task-Oriented Dialogue (TOD) systems. **Firstly**, we introduce novel approaches for simulating task-oriented dialogues by leveraging instructional documents from WikiHow, thereby creating rich training datasets. **Secondly**, we implement rigorous quality control mechanisms to ensure the generated dialogues are both contextually relevant and of high quality. **Thirdly**, we demonstrate the practical application of large-scale simulated dialogues (full Inst2Dial-Auto) by utilizing them as pre-training data, which results in significant performance improvements in state-of-the-art TOD models.

2 Related Work

Task-Oriented Dialogue (TOD) systems, a subdomain of conversational systems, act in the role of task completion assistant with a requirement of comprehensive task knowledge [19]. Traditional techniques have aimed to improve various components, such as dialogue state tracking [39], action prediction [3] and response generation [7,13]. Recent end-to-end approaches started leveraging advanced language models as backbones for enhanced performance in natural language understanding and generation [5,14,38]. In line with the advancements in conversational systems, such as UBAR [38] and JSA-TOD [5], many conversational datasets have emerged. These include the Schema-Guided Dialogue (SGD) [30], MultiWoZ [40] and RiSAWOZ [27] datasets. However, these datasets are primarily limited by restricted domain coverage (as exemplified by MultiWoZ's coverage of only eight domains and SGD's extension to a current maximum of 16) and insufficient instructional information, impairing the effectiveness of TOD models trained on them.

To address these limitations, various research efforts have either augmented existing datasets or employed human engagement to create more inclusive dialogue datasets. However, these approaches are often contained by their considerable financial implications and intricate design requirements [9]. Meanwhile, there exists a thread of work leveraging simulation techniques for dialogue generation [9,20,34]. Classical approaches predominantly utilise rule-based methods [23,34]. A contemporary instance is [9], which employed a BERT-like architecture to generate dialogues based on given texts, culminating in the release of the WikiDialog corpus. Each dialogue in this corpus is synthesised from a corresponding Wikipedia passage. For task-oriented dialogues, Mohapatra et al., [20] fine-tuned a GPT-2 model [28] and applied it to a certain dialogue context to generate simulated dialogues, aiming to improve the performance of TOD models, particularly in a low-resource setting. While their work aligns closely with our own contributions, it fully relies on existing task completion scenarios within the dataset, thereby neglecting the rich instructional data contained in external knowledge bases. This oversight results in the persistence of limited domain coverage, an issue we previously identified. In summary, extant research has not leveraged step-wise instructional content to develop synthetic dialogues, a gap that our work aims to fill, especially considering recent advancements in generative models.

3 INST2DIAL Synthetic Dialogue Development

In this section, we discuss three strategies for generating synthetic dialogues, leveraging rich task-specific instructional content from an external knowledge base, Task2KB [25].

Problem Description. Formally, each task t is characterised by its title τ^t, introduction i^t and a series of k instructional steps $S^t = \{s_1^t, s_2^t, ..., s_k^t\}$. The process of generating synthetic dialogues is modelled as $\hat{d}^t = f(\tau^t, i^t, S^t)$. For readability, subsequent descriptions will omit the task index t. We conceptualise $f(\cdot)$ as either a question generator paired with selected instructions as responses or as a full-dialogue generator using the instructions as input. Next, we proceed to detail the three strategies we propose to implement $f(\tau^t, i^t, S^t)$.

3.1 INST2DIAL-Auto

In this approach, we employ an advanced encoder-decoder model to generate a full conversation. This model uses instructional steps for task completion as responses while automatically generating pertinent questions. Unlike in open-domain scenarios, the question generator is specifically designed to ask task-relevant questions that help progress task completion. The generation process is divided into three stages: (1) neural question generator learning, (2) formulating the input for dialogue generation, and (3) the actual dialogue generation.

Neural Question Generator Learning. The aim of this module is to fine-tune an encoder-decoder model, such as Flan-T5 [8], to produce high-quality questions for synthetic task-oriented dialogues. Inspired by the methodology in [9], we employ sequential masking on questions within existing dialogues to create input-output pairs (i.e., dialogue inpainting). In this format, the input contains masked text designated to be filled with a generated question. For a dialogue d consisting of two sequential Question-Answer (QA) pairs, $d = \{q_0, a_0, q_1, a_1\}$, we derive two input-output pairs as follows:

$$\begin{aligned}&\textbf{input}: [MASK][SEP]a_0 \rightarrow \textbf{output}: \hat{q}_0\\&\textbf{input}: q_0[SEP]a_0[SEP][MASK][SEP]a_1 \rightarrow \textbf{output}: \hat{q}_1\end{aligned}$$

Here, $[MASK]$ marks where the generated question should be inserted, and $[SEP]$ separates questions and answers. The model is then trained to predict suitable questions for such structured inputs. To tailor the question generator for task completion, we propose training on specialised TOD datasets rather than commonly-used open-domain datasets [2,9].

Input Formation. In the next stage, we focus on input formation using our trained task-oriented question generator. The goal is to capture the logical progression of a task in dialogue form. To this end, we use step-wise task instructions

S as answers to prospective user questions, forming a chain of linked responses with missing questions for a task t. We also introduce strategies to balance dialogue length and information content. Specifically, we investigate two methods: (1) employing topic sentences from each step description, and (2) choosing the most specific sentence as determined by a text specificity predictor, speciteller [18]. After a thorough evaluation, we find that using topic sentences leads to higher-quality synthetic dialogues – they are both fluent and informative with less noise. We publicly make available both implementations in our GitHub repository for the details.

Dialogue Generation. Finally, we turn to generating the missing questions in our prepared dialogues from the previous stage, represented as $d = \{\square,\ a_0, \square,\ a_1,\ ...,\ \square,\ a_{|d|}\}$. We introduce three strategies, Single-QA, Last-QA and Full-QA, that take different input contexts into account. All strategies initiate the dialogue using the task's title τ and introduction i from the Task2KB knowledge base [33] as the opening QA pair to set the context. Subsequent dialogue is generated incrementally, with each new input influenced by prior ones. The input formatting is as follows:

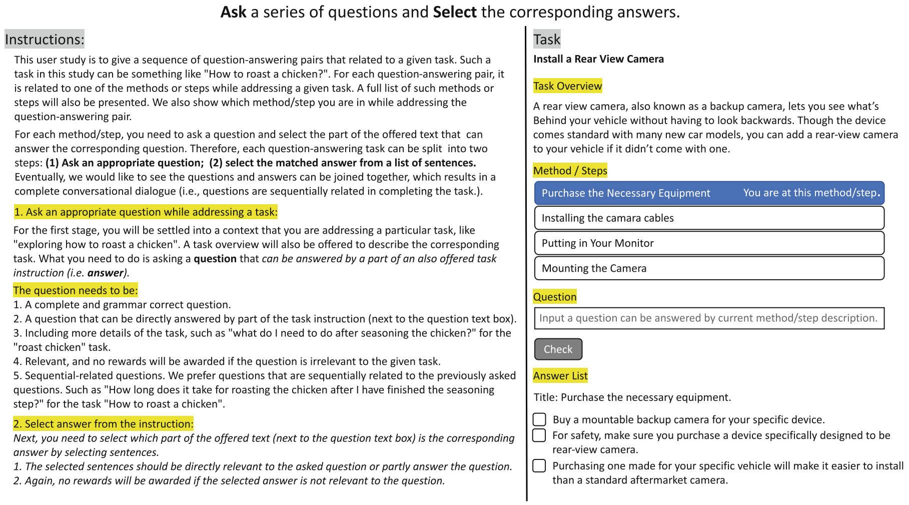
$$\begin{aligned}
&\textbf{input 1}: q^{\tau}[SEP]a^{i}[SEP][MASK][SEP]a_0\\
&\textbf{input 2}: q^{\tau}[SEP]a^{i}[SEP](\hat{q}_0[SEP]a_0)[SEP][MASK][SEP]a_1\\
&\textbf{input n}: ...
\end{aligned}$$

The second input incorporates the question generated from the first, serving as an extended context. We differentiate between the three strategies by varying the scope of the input: Single-QA includes only the immediate preceding answer, Last-QA adds the last QA pair, and Full-QA incorporates all previous QA pairs. Upon optimising these generation methods, we produce a comprehensive set of automated task-oriented dialogues (INST2DIAL-Auto) covering a broad array of tasks within a large-scale knowledge base, Task2KB.

3.2 INST2DIAL-Manual

Next, we introduce our second approach for generating synthetic task-oriented dialogues, Crowdsourced Question Generation (i.e., INST2DIAL-Manual). Unlike the first strategy, which relies on a learned generative model, this method leverages human efforts for dialogue generation, particularly focusing on generating high-quality questions that integrate seamlessly into full dialogues with step-by-step instructional responses.

To achieve this, we designed a user study, creating a custom interface that enables crowd workers to generate the INST2DIAL-Manual dataset. In Fig. 2, we present an example task, "Install a Rear View Camera", to demonstrate the interface used by the crowd workers. The interface is split into two sections. The left side provides detailed guidelines for the task, while the right side is designed for worker input. As can be seen in Fig. 2, crowd workers are instructed to perform two actions for each step of a task: (1) formulate a relevant question concerning the task, and (2) select an appropriate response from the instructional material. Due to space limitations, Fig. 2 only provides the first step (in



Ask a series of questions and **Select** the corresponding answers.

Instructions:

This user study is to give a sequence of question-answering pairs that related to a given task. Such a task in this study can be something like "How to roast a chicken?". For each question-answering pair, it is related to one of the methods or steps while addressing a given task. A full list of such methods or steps will also be presented. We also show which method/step you are in while addressing the question-answering pair.

For each method/step, you need to ask a question and select the part of the offered text that can answer the corresponding question. Therefore, each question-answering task can be split into two steps: **(1) Ask an appropriate question; (2) select the matched answer from a list of sentences.** Eventually, we would like to see the questions and answers can be joined together, which results in a complete conversational dialogue (i.e., questions are sequentially related in completing the task.).

1. Ask an appropriate question while addressing a task:

For the first stage, you will be settled into a context that you are addressing a particular task, like "exploring how to roast a chicken". A task overview will also be offered to describe the corresponding task. What you need to do is asking a **question** that *can be answered by a part of an also offered task instruction (i.e.* ***answer****).*

The question needs to be:

1. A complete and grammar correct question.
2. A question that can be directly answered by part of the task instruction (next to the question text box).
3. Including more details of the task, such as "what do I need to do after seasoning the chicken?" for the "roast chicken" task.
4. Relevant, and no rewards will be awarded if the question is irrelevant to the given task.
5. Sequential-related questions. We prefer questions that are sequentially related to the previously asked questions. Such as "How long does it take for roasting the chicken after I have finished the seasoning step?" for the task "How to roast a chicken".

2. Select answer from the instruction:

Next, you need to select which part of the offered text (next to the question text box) is the corresponding answer by selecting sentences.

1. The selected sentences should be directly relevant to the asked question or partly answer the question.

2. Again, no rewards will be awarded if the selected answer is not relevant to the question.

Task

Install a Rear View Camera

Task Overview

A rear view camera, also known as a backup camera, lets you see what's Behind your vehicle without having to look backwards. Though the device comes standard with many new car models, you can add a rear-view camera to your vehicle if it didn't come with one.

Method / Steps

Purchase the Necessary Equipment You are at this method/step.

Installing the camara cables

Putting in Your Monitor

Mounting the Camera

Question

Input a question can be answered by current method/step description.

Check

Answer List

Title: Purchase the necessary equipment.

□ Buy a mountable backup camera for your specific device.

□ For safety, make sure you purchase a device specifically designed to be rear-view camera.

□ Purchasing one made for your specific vehicle will make it easier to install than a standard aftermarket camera.

Fig. 2. The interface for crowd workers. A complete instruction is in the left, which stands next to an example interface that allows workers to write a question for the first step of an example task (i.e., install a rear-view camera) and select the corresponding sentence-wise answer.

the bottom right) of the full instructions available to workers. Post data collection, we implement quality control measures, involving manual labelling by domain experts. They assess each dialogue on three criteria: the relevance and meaningfulness of the questions, the alignment of the questions with the context, and the compatibility between questions and answers. Dialogues that fail any of these tests are excluded to ensure a high-quality dataset.

3.3 INST2DIAL-ICL

In addition to the previous methods, we also explore the potential of advanced language models, like GPT [21], for generating synthetic dialogues through in-context learning [37]. This approach has demonstrated strong performance in various generative tasks. We craft a specialised prompt that incorporates task-related information, including the task title, introduction and step-by-step instructions. The model employed for this study is GPT-3.5, and the corresponding prompt is as follows:

> You help in generating a mix-initiative synthetic multi-turn task-oriented dialogue, each utterance starts with [user] or [system], while [user] initiates the conversations, and they take turns (with a similar number of turns to the number of steps) in giving utterances, by leveraging the task instructions as input but without explicitly mentioning the steps. Example input: [task title: □, instruction: □], Output: {[user] □, [system] □, ..., [system] □}.

□ denotes placeholder information, omitted here due to space constraints. Using Task2KB, this approach allows us to automatically produce complete dialogues for each task without the need for heavy model training or human intervention.

To summarise, this section outlines three strategies employed in this study to produce synthetic task-oriented dialogues. These strategies aim to enhance TOD models and ultimately improve user experience through more accurate and contextually relevant system responses. Details for the implementation of each method will be covered in the subsequent section.

4 Experimental Setup

In this section, we outline the experimental framework designed to implement and assess the three strategies for generating synthetic datasets and their effectiveness in enhancing task-oriented dialogue systems. We commence with an investigation into the optimization of the INST2DIAL-Auto strategy, guided by the following research question:

RQ1: Which strategy learns the best question generator (INST2DIAL-Auto) for the generation of task-oriented dialogues?

In line with the methodology articulated in [9], our first step involves conducting a user study to evaluate the quality of the synthetic dialogues we generate, especially given the absence of ground-truth benchmarks. Subsequently, we use these high-quality synthetic dialogues as model pre-training data, thereby boosting more robust model performance. With this approach, we intend to address the subsequent research questions:

RQ2: Which among the three strategies for synthetic dialogue generation yields the highest quality of dialogues as per human evaluation?

RQ3: Does the top-performing strategy with the resulting synthetic dialogues also contribute to improving state-of-the-art TOD models?

To answer these questions, we have crafted a comprehensive set of experiments for each of the three strategies with code and resources publicly[2]:

Table 1. Statistic of datasets for learning question generators.

Type	Dataset	Train	Valid	Test
# Dialogues	QReCC	11,020	1,409	1,409
	ORConvQA	8,766	980	1,542
	MultiWoZ	8,437	1,000	1,000
# QA pairs	QReCC	52,481	6,816	6,817
	ORConvQA	22,760	2,450	4,029
	MultiWoZ	16,352	2,085	2,133

Table 2. The Perplexity score of running various language models on multiple conversational datasets.

Models	ORConvQA	QReCC	MultiWoZ
T5-base	4.6423	3.7980	3.8165
Flan-T5-base	**4.5446**	**3.7214**	3.6611
Flan-T5-large	5.3126	3.8204	**3.6058**

[2] https://github.com/wangxieric/task2kb-resource.

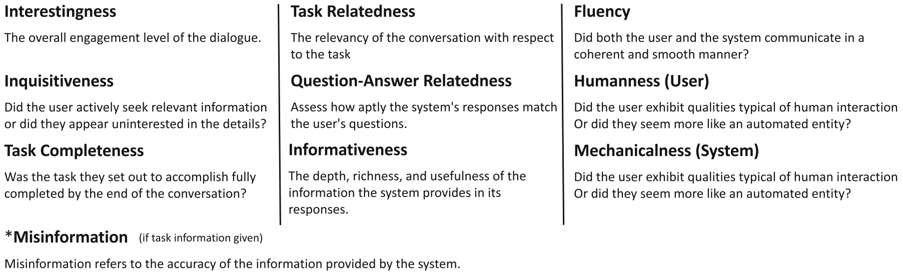
INST2DIAL-Auto: At first, we explore the performance differences between open-domain and task-oriented dialogues for fine-tuning Language Models (LMs) as question generators. We use T5-base, Flan-T5-base, and Flan-T5-large models, fine-tuning them on two open-domain datasets – ORConvQA [26] and QReCC [1] – as well as one task-oriented dataset, MultiWoZ. Table 1 provides a summary of these datasets, highlighting similarities in dialogue count but variations in conversation length (measured by the number of QA turns). QReCC encompasses longer conversations compared to the other two datasets, which have similar turns of QAs. Then, we fine-tune these models as question generators using a default learning rate of 1e-4, the widely-used Adam optimizer [17] and cross-entropy loss. Our evaluation proceeds in several steps: We first assess the overall performance of each fine-tuned model across the three datasets. We then evaluate the effectiveness of these models specifically for question generation in task-oriented dialogues, aiming to highlight any performance disparities. Finally, we examine the impact of context input types-Single-QA, Last-QA, and Full-QA-each of which incorporates historical utterances differently. This aspect of the evaluation is particularly focused on the MultiWoZ dataset and the inputs for final generation (TOC-Auto) to explore the influence of varying input lengths.

INST2DIAl-Manual: To collect INST2DIAL-Manual, we conduct a user study using the Amazon Mechanical Turk platform[3]. To ensure a high quality of written English, we restricted participation to individuals residing in English-speaking countries, as identified by the relevant Wikipedia page[4]. Further refining our participant tool, we required each worker to have successfully completed at least 2,000 previous tasks (known as HITs) with an approval rate exceeding 95%. To guide question construction, we implemented specific guidelines into the "Check" button functionality. These guidelines mandate each question should begin with one of the updated 5Ws ('how', 'what', 'when', 'where', or 'why'), contain at least five words, and conclude with a question mark. After crafting a question, workers are obliged to select the corresponding answer from the task's step description before proceeding to the next step or submitting their final responses. To minimize the risk of low-quality or copied inputs, we deactivated the copy-paste feature within the text box designated for question formulation and collected a minimum of two dialogues for each task. For their contributions, workers were compensated at a rate of US$0.10 per completed dialogue.

INST2DIAL-ICL: For the generation of the INST2DIAL-ICL dataset, we followed the prompt configuration detailed in Sect. 3.3. We utilized the pretrained GPT-3.5 model, specifically the gpt-3.5-turbo version, in conjunction with OpenAI's Chat Completion API. We adhered to the API's default settings for the dialogue generation process.

To address RQ2, we evaluate the three types of synthetic dialogues generated from a shared random sample of 100 tasks in 19 categories from Task2KB. Conducted on Amazon Mechanical Turk, our user study employs 10 criteria

[3] https://www.mturk.com/.
[4] https://en.wikipedia.org/wiki/English-speaking_world.



Interestingness

The overall engagement level of the dialogue.

Inquisitiveness

Did the user actively seek relevant information or did they appear uninterested in the details?

Task Completeness

Was the task they set out to accomplish fully completed by the end of the conversation?

Task Relatedness

The relevancy of the conversation with respect to the task

Question-Answer Relatedness

Assess how aptly the system's responses match the user's questions.

Informativeness

The depth, richness, and usefulness of the information the system provides in its responses.

Fluency

Did both the user and the system communicate in a coherent and smooth manner?

Humanness (User)

Did the user exhibit qualities typical of human interaction Or did they seem more like an automated entity?

Mechanicalness (System)

Did the user exhibit qualities typical of human interaction Or did they seem more like an automated entity?

***Misinformation** (if task information given)

Misinformation refers to the accuracy of the information provided by the system.

Fig. 3. Dialogue Evaluation Aspects

based on a recent dialogue quality framework [32]. These 10 evaluative aspects are detailed in Fig. 3. Participants evaluate dialogues in two scenarios: with and without task instructions. The ‘misinformation’ metric is applied only when instructions are available. Each dialogue undergoes six independent evaluations-three per scenario-and workers receive US$0.10 per evaluation.

Table 3. Generated questions linked to an example response selected from a task step description. Key information are highlighted in bold. Sentences that are not proper questions are marked with a [×] symbol.

Task Title		How to learn music theory online?
LLM	Dataset	Generated Dialogue Examples
Flan-T5-Base	**QReCC**	You can **learn music theory online** for **free**? [×]
	ORConvQA	How to learn **music theory online**?
	MultiWoZ	Can you recommend a good place to find a **music theory lesson**?
Flan-T5-Large	**QReCC**	if you’re looking for **a low cost way** to **learn music theory**? [×]
	ORConvQA	You can find a tutor or teacher to teach you? [×]
	MultiWoZ	What is **the best way** to find a **free course**?
Example Response		***Online learning*** is **a great way** to find a **lesson** taught by a professional **without having to pay the cost**.”

5 Results

In this section, we present the experimental results and conduct a thorough analysis that answers the first two research questions.

Optimise Question Generator for INST2DIAL-Auto (RQ1): To optimise the configuration for our learned question generator, we initially evaluate various encoder-decoder models across three datasets: ORConvQA, QReCC and MultiWoZ. The results of this evaluation can be found in Table 2. Notably, the

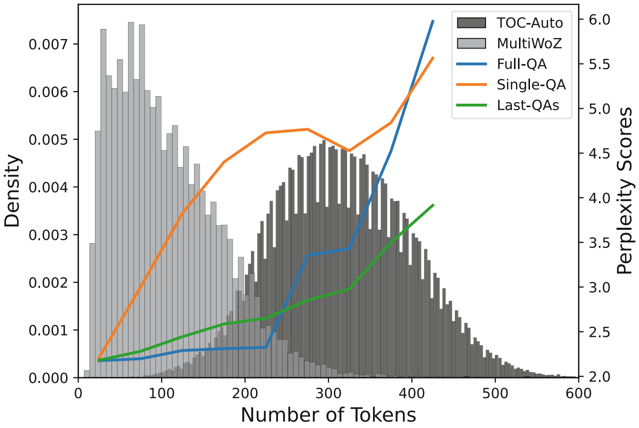


Fig. 4. Evaluation of single-QA, Pair-QAs and Full-QAs on MultiWoZ with distinct input lengths.

Flan-T5 model [8] outperforms T5 [29] with lower perplexity scores, underscoring the effectiveness of language models fine-tuned on instructional resources or chain-of-thought documents. To elucidate the disparities among questions generated by different language models, we provide an illustrative example, using a response as input and comparing various generated questions. These are detailed in Table 3. In our observations, models trained on open-domain dialogues struggle to generate contextually accurate and proper questions. In addition, the Flan-T5-Large model, which scored best in our initial evaluation, after trained on MultiWoZ, adeptly captures critical response elements like 'a great way', 'lesson' and 'without having to pay the cost', resulting in highly relevant questions. Therefore, we employ the Flan-T5-Large model, fine-tuned on the MultiWoZ dataset, for our subsequent dialogue generation experiments. We also assess model performance under varying conditions by examining average perplexity scores as the length and complexity of conversational history change. Figure 4 visualizes this, juxtaposing three question-generating setups: Single-QA, Last-QA, and Full-QA. Our findings indicate that all three approaches experience a performance decline in generating questions for longer conversations. Specifically, Single-QA consistently lags behind, primarily due to its disregard for conversational history. Full-QA excels in shorter dialogues, but its performance decreases as the input length increases. Last-QA proves to be the most resilient, maintaining stable performance irrespective of input length. To sum up, in response to Research Question 1 (RQ1), the Last-QA approach, when used in conjunction with the Flan-T5-Large model fine-tuned on the MultiWoZ dataset, achieves an effective balance between historical context retention and adaptability to longer dialogues.

Quality of Synthetic Dialogues as per Human Evaluation (RQ2): To evaluate the quality of dialogues generated through three distinct strategies –

Inst2Dial-Auto/Manual/ICL – we conducted a comprehensive user study, collecting feedback on 100 sampled dialogues from each group. Specifically, each dialogue was evaluated by at least 3 crowd workers as per 10 evaluation criteria (see Fig. 3). These criteria were applied both with and without the provision of task instructions. Figure 5 displays the average scores for these metrics, which range from 1 to 4, for each dataset. Upon examining the evaluations under both scenarios, with and without task instructions, it became evident that participants gave more varied scores when additional contextual information was available. Across all three strategies, the generated dialogues were found to be interesting, fluent, and task-relevant. However, when metrics such as task completeness, question-answer relevance, and informativeness were considered, Inst2Dial-Auto consistently outperformed the other two datasets. This was particularly true for the informativeness metric when complete information was provided for comparison. Interestingly, Inst2Dial-ICL scored the lowest in terms of accurate information dissemination, as reflected by its misinformation scores – a finding that aligns with the observed hallucination issue of LLMs [35]. Overall, Inst2Dial-Auto emerged as the most satisfactory dataset under both evaluation conditions. Therefore, in response to RQ2, Inst2Dial-Auto yields the most satisfactory dialogues according to user feedback, compared to the other two synthetic dialogue types.

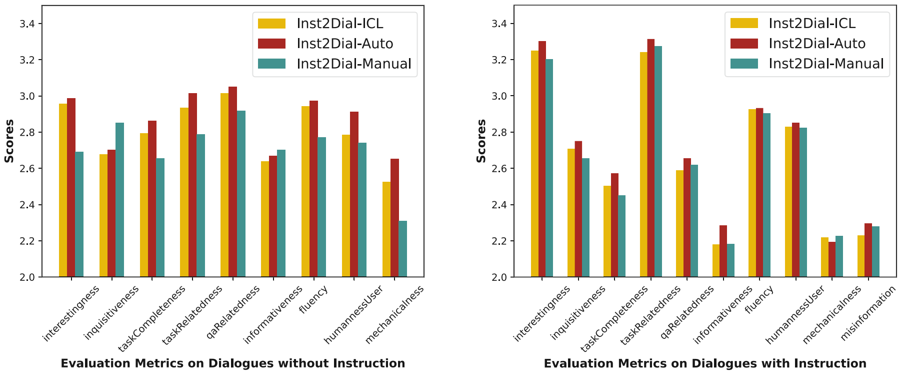


Fig. 5. Dialogue Quality Comparison on Essential Aspects as per User Feedback.

6 Application

Based on the evaluation results from the previous section, the INST2DIAL-Auto dataset – generated using a fine-tuned encoder-decoder model – emerged as the best-performing strategy. This leads us to RQ3, which investigates the impact of using INST2DIAL-Auto on SOTA TOD models. While incorporating synthetic data into the training set could improve model performance, it also risks complicating dialogue structuring and demands additional engineering efforts. To mitigate these challenges, we opt to fine-tune a pre-existing language model, DistilGPT2 [31], on our synthetic dataset to encode task-specific knowledge. We then integrate this fine-tuned model into two recent advanced neural TOD models to assess the impact on performance. Additionally, we introduce a knowledge-augmented loss function specifically designed for fine-tuning Language Models (LMs) to effectively generate task-oriented responses on the full INST2DIAL-Auto dataset with all tasks, 251,433 in total, from Task2KB. The loss function for response generation ($\mathcal{L}_{RG}$) is defined as $\mathcal{L}_{RG} = -\sum_{t=1}^{T} \log p(r_t|q_t, r < t, q^{title}, a^i)$, where q_t denotes the question asked during the t-th conversational turn and $r < t$ represents the modelling of prior conversations. We employ the Adam optimiser [17] with a learning rate of $2e^{-}5$. For empirical validation, the fine-tuned LMs serve as substitute backbones for two state-of-the-art TOD models: **(1) UBAR** [38], an advanced model that extends SimpleTOD [14] and Soloist [24], giving it complete access to full dialogues, beliefs, database states and system acts for each conversational turn. **(2) JSA-TOD** [5], which employs a recent joint stochastic approximation algorithm [22] for semi-supervised learning, concentrating on leveraging both labelled and unlabelled dialogue data.

We assess the performance of the trained DistilGPT2 across multiple training settings: full supervision, few-shot learning and limited domain knowledge setting. For full supervision, we adhere to the publicly accessible implementations of baseline models, using fully labelled training data. In the few-shot learning context, we train UBAR and JSA-TOD models on randomly sample 10% of the training data. Notably, for JSA-TOD, we use a semi-supervised setup, comprising 3% labelled and 7% unlabeled dialogue data sampled from the training set. Upon limited domain knowledge setup, we substitute 10% of the training data with domain-specific dialogues, such as those related to hotels or restaurants, and then evaluate performance on a test set containing a diverse array of tasks. Conversely, we further evaluate our fine-tuned backbone model by comparing it to an LM trained on a contemporary synthetic dialogue dataset known as WikiDialog (wdl) [9]. Similar to our dataset, WikiDialog converts documents into conversational dialogues; however, instead of employing task instructions for dialogue generation, it uses passages extracted from Wikipedia. To ensure a fair comparison, we sample an equal number of dialogues from WikiDialog as the size of ours, which is smaller. Subsequently, we adhere to an identical procedure in training a DistilGPT2 model for comparison. In particular, we evaluate various UBAR and JSA-TOD implementations on the MultiWoZ datasets, version 2.0 [4] and 2.1 [12], respectively, as they reported in their paper. For a comprehensive evaluation, we rely on key metrics associated with the MultiWoZ dataset: informativeness, success rate, BLEU scored and combined performance metrics.

In Table 4, we present the experimental results. First, under a mixed-domain setup, we note tangible enhancements in both UBAR and JSA-TOD models when employing DistilGPT2 trained on INST2DIAL-Auto in a few-shot learning context. These gains do diminish when juxtaposed with the full supervision baseline, a phenomenon attributable to the well-understood issue of catastrophic forgetting [15]. Nevertheless, the use of INST2DIAL-Auto continues to offer observable benefits. Subsequently, we scrutinize the performance of these models in sce-

Table 4. The experimental results of UBAR and JSA-TOD on MultiWoZ. ‘re’, ‘wdl’ and ‘our’ refer to the use of initial DistilGPT2, the one trained on Wikidialog and our synthetic data, respectively. The improvement ratio is compared to the reproduced model on combined scores.

UBAR					
Setups	**Inform**	**Success**	**BLEU**	**Combined**	**Impr. %**
Few Shot Setting with Mixture Domains					
Few Shot + re	50.55	37.94	10.88	55.12	-
Few Shot + wdl	53.95	41.94	11.49	59.44	7.8%
Few Shot + our	**54.85**	**42.34**	**11.59**	**60.19**	**9.1%**
Full Supv. + re	87.69	75.88	**14.87**	96.65	-
Full Supv. + wdl	89.99	78.58	14.83	99.11	2.5%
Full Supv. + our	**90.19**	**79.08**	14.83	**99.46**	**2.9%**
Domain Generalisability					
Few Shot (hotel) + re	41.74	25.33	10.81	44.34	-
Few Shot (hotel) + wdl	45.75	29.43	11.78	49.36	11.3%
Few Shot (hotel) + our	**48.25**	**32.13**	**11.86**	**52.05**	**17.4%**
Few Shot (train) + re	51.65	32.23	10.79	52.73	-
Few Shot (train) + wdl	48.15	28.63	10.99	49.38	-6.3%
Few Shot (train) + our	**56.45**	**37.73**	**11.56**	**58.65**	**11.2%**
Few Shot (attraction) + re	47.45	31.53	11.13	50.62	-
Few Shot (attraction) + wdl	51.65	32.93	9.92	52.21	3.1%
Few Shot (attraction) + re	**55.16**	**39.04**	**11.52**	**58.62**	**15.8%**
Few Shot (restaurant) + re	47.75	29.13	10.69	49.13	-
Few Shot (restaurant) + wdl	51.75	34.53	12.17	55.31	12.6%
Few Shot (restaurant) + our	**54.95**	**37.74**	**12.33**	**58.68**	**19.4%**
Few Shot (taxi) + re	37.14	18.62	8.58	36.46	-
Few Shot (taxi) + wdl	45.65	24.12	8.98	43.87	20.3%
Few Shot (taxi) + our	**45.75**	**24.22**	**9.10**	**44.09**	**20.9%**

JSA-TOD					
Setups	**Inform**	**Success**	**BLEU**	**Combined**	**Impr.%**
Few Shot Setting with Mixture Domains					
Few Shot + re	53.10	36.30	**14.86**	59.56	
Few Shot + wdl	55.80	37.40	13.38	59.98	0.7%
Few Shot + our	**57.80**	**40.40**	12.96	**62.06**	**4.2%**
Full Supv. + re	86.30	75.90	19.08	100.18	
Full Supv. + wdl	83.40	73.40	18.65	97.05	-3.1%
Full Supv. + our	**87.51**	**76.20**	**19.87**	**101.73**	**1.5%**
Domain Generalisability					
Few Shot (hotel) + re	28.90	11.40	11.09	31.24	
Few Shot (hotel) + wdl	**29.50**	13.40	11.12	32.57	4.2%
Few Shot (hotel) + our	28.10	**15.60**	**12.06**	**33.91**	**8.5%**
Few Shot (train) + re	28.40	11.80	11.00	31.10	
Few Shot (train) + wdl	27.40	10.80	12.61	31.71	2.0%
Few Shot (hotel) + our	**28.80**	**14.60**	**12.64**	**34.34**	**10.4%**
Few Shot (attraction) + re	27.60	**13.90**	8.17	28.92	
Few Shot (attraction) + wdl	25.90	11.30	11.00	29.60	2.4%
Few Shot (attraction) + our	**28.50**	12.30	**10.11**	**30.51**	**5.5%**
Few Shot (restaurant) + re	28.90	17.20	12.65	35.70	
Few Shot (restaurant) + wkl	27.40	14.80	**13.25**	34.35	-3.8%
Few Shot (restaurant) + our	**33.60**	**24.00**	13.02	**41.82**	**17.1%**
Few Shot (taxi) + re	23.90	9.80	**8.35**	25.20	
Few Shot (taxi) + wkl	26.10	9.50	7.96	25.76	2.2%
Few Shot (taxi) + our	**27.30**	**11.80**	8.26	**27.81**	**10.4%**

narios with restricted domain knowledge. Here, we identify marked and consistent performance uplifts when integrating our INST2DIAL-Auto dataset, as substantiated by the data in Table 4. Thus, we affirm that employing INST2DIAL-Auto can augment the capabilities of pre-trained language models when deployed in task-oriented dialogue systems. This elevation in performance is particularly pronounced in both few-shot learning and out-of-domain application scenarios.

Next, we assess the utility of encoding task-specific instructions as opposed to utilizing open-domain resources like Wikipedia. Firstly, under both few-shot and full supervision paradigms, a DistilGPT2 model trained on WikiDialog fails to consistently outperform the baseline, particularly for the JSA-TOD model. In fact, it results in a 3.1% decrease in the combined score under full supervision. Secondly, when juxtaposing DistilGPT2 models fine-tuned on both WikiDialog and INST2DIAL-Auto to test domain generalizability, INST2DIAL-Auto emerges as demonstrably more effective, driving both models to achieve superior results. To address **RQ3**, our conclusions indicate that the INST2DIAL-Auto dataset is highly effective in enhancing the ability of task-oriented dialogue models to accurately interpret and respond to task-specific information across diverse categories. This results in noticeably improved conversational responses.

7 Conclusions

In this study, we explored three innovative strategies for generating synthetic dialogues with the aim of enhancing Task-Oriented Dialogue (TOD) models. The first approach leverages a sophisticated neural question generator within an optimized pipeline to produce the Inst2Dial-Auto dataset. For the other two datasets, Inst2Dial-Manual/ICL, we deployed a carefully designed user study and in-context learning prompts, respectively. Our empirical evaluation, rooted in human evaluation metrics, revealed that dialogues produced via a finely-tuned question generator-Inst2Dial-Auto-consistently yielded the highest quality. When applied to state-of-the-art TOD models, this dataset contributed to substantial improvements, most notably in scenarios with limited domain knowledge, registering a minimum uplift of 5.5% in combined evaluation scores.

Acknowledgement. This research is supported by the Alan Turing Institute under the EPSRC grant [EP/N510129/1] and the EPSRC Fellowship titled "Task Based Information Retrieval" [EP/P024289/1].

References

1. Anantha, R., Vakulenko, S., Tu, Z., Longpre, S., Pulman, S., Chappidi, S.: Open-domain question answering goes conversational via question rewriting. In: Proceedings of NAACL (2021)
2. Bao, J., et al.: A synthetic data generation framework for grounded dialogues. In: Proceedings of ACL (2023)
3. Boyer, K., Ha, E.Y., Phillips, R., Wallis, M., Vouk, M., Lester, J.: Dialogue act modeling in a complex task-oriented domain. In: Proceedings of SIGDIAL (2010)
4. Budzianowski, P., et al.: MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In: Proceedings of EMNLP (2018)
5. Cai, Y., Liu, H., Ou, Z., Huang, Y., Feng, J.: Advancing semi-supervised task oriented dialog systems by JSA learning of discrete latent variable models. In: Proceedings of SIGDIAL (2022)
6. Chen, D., Yu, Z.: Sources of noise in dialogue and how to deal with them. In: Proceedings of SIGDIAL (2023)
7. Chen, X., Xu, J., Xu, B.: A working memory model for task-oriented dialog response generation. In: Proceedings of ACL (2019)
8. Chung, H.W., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
9. Dai, Z., et al.: Dialog inpainting: turning documents into dialogs. In: Proceedings of ICML (2022)
10. De Cicco, R., Silva, S.C.L.d.C.e., Alparone, F.R.: "It's on its way": chatbots applied for online food delivery services, social or task-oriented interaction style? J. Foodserv. Bus. Res. **24**(2), 140–164 (2021)
11. El Asri, L., et al.: Frames: a corpus for adding memory to goal-oriented dialogue systems. In: Proceedings of SIGdial (2017)
12. Eric, M., et al.: Multiwoz 2.1: a consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In: Proceedings of LREC (2020)
13. Hosseini-Asl, E., McCann, B., Wu, C.S., Yavuz, S., Socher, R.: A simple language model for task-oriented dialogue. In: Proceedings of NeurIPS (2020)
14. Hosseini-Asl, E., McCann, B., Wu, C., Yavuz, S., Socher, R.: A simple language model for task-oriented dialogue. In: Proceedings of NeurIPS (2020)
15. Hu, W., et al.: Overcoming catastrophic forgetting for continual learning via model adaptation. In: Proceeding of ICLR (2019)

16. Jin, D., Kim, S., Hakkani-Tur, D.: Can i be of further assistance? using unstructured knowledge access to improve task-oriented conversational modeling. In: Proceedings of DialDoc (2021)
17. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. In: Proceedings of ICLR (2015)
18. Li, J.J., Nenkova, A.: Fast and accurate prediction of sentence specificity. In: Proceedings of AAAI (2015)
19. Madotto, A., Wu, C.S., Fung, P.: Mem2seq: effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In: Proceedings of ACL (2018)
20. Mohapatra, B., Pandey, G., Contractor, D., Joshi, S.: Simulated chats for building dialog systems: learning to generate conversations from instructions. In: Proceedings of EMNLP (2021)
21. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
22. Ou, Z., Song, Y.: Joint stochastic approximation and its application to learning discrete latent variable models. In: Proceedings of UAI (2020)
23. Papangelis, A., Wang, Y.C., Molino, P., Tur, G.: Collaborative multi-agent dialogue model training via reinforcement learning. In: Proceedings of SIGDIAL (2019)
24. Peng, B., Li, C., Li, J., Shayandeh, S., Liden, L., Gao, J.: SOLOIST: building task bots at scale with transfer learning and machine teaching. Trans. Assoc. Comput. Linguist. **9**, 824–907 (2021)
25. Procheta, S., Xi, W., Ruiqing, X., Emine, Y.: Task2kb: a public task-oriented knowledge base. In: Proceedings of AAAI (2023)
26. Qu, C., Yang, L., Chen, C., Qiu, M., Croft, W.B., Iyyer, M.: Open-retrieval conversational question answering. In: Proceedings of SIGIR (2020)
27. Quan, J., Zhang, S., Cao, Q., Li, Z., Xiong, D.: Risawoz: a large-scale multi-domain wizard-of-oz dataset with rich semantic annotations for task-oriented dialogue modeling. In: Proceedings of EMNLP (2020)
28. Radford, A., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
29. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 5485–5551 (2020)
30. Rastogi, A., Zang, X., Sunkara, S., Gupta, R., Khaitan, P.: Towards scalable multi-domain conversational agents: the schema-guided dialogue dataset. In: Proceedings of AAAI (2020)
31. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
32. See, A., Roller, S., Kiela, D., Weston, J.: What makes a good conversation? how controllable attributes affect human judgments. In: Proceedings of NAACL-HLT (2019)
33. Sen, P., Wang, X., Xu, R., Yilmaz, E.: Task2kb: a public task-oriented knowledge base. In: Proceedings of AAAI (2023)
34. Shah, P., et al.: Building a conversational agent overnight with dialogue self-play. arXiv preprint arXiv:1801.04871 (2018)
35. Shuster, K., Poff, S., Chen, M., Kiela, D., Weston, J.: Retrieval augmentation reduces hallucination in conversation. In: Proceedings of EMNLP (Findings) (2021)
36. Srivastava, M., Lu, Y., Peschon, R., Li, C.: Pretrain-finetune based training of task-oriented dialogue systems in a real-world setting. In: Proceedings of NAACL (2021)
37. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of NeurIPS (2022)

38. Yang, Y., Li, Y., Quan, X.: Ubar: towards fully end-to-end task-oriented dialog system with gpt-2. In: Proceedings of AAAI (2021)
39. Ye, F., Wang, X., Huang, J., Li, S., Stern, S., Yilmaz, E.: Metaassist: robust dialogue state tracking with meta learning. In: Proceedings of EMNLP (2022)
40. Zang, X., Rastogi, A., Sunkara, S., Gupta, R., Zhang, J., Chen, J.: Multiwoz 2.2 : a dialogue dataset with additional annotation corrections and state tracking baselines. arXiv preprint arXiv:2007.12720 (2020)