DialogStudio: Towards Richest and Most Diverse Unified Dataset Collection for Conversational AI

Jianguo Zhang[*1], **Kun Qian**[*2], **Zhiwei Liu**[1], **Shelby Heinecke**[1], **Rui Meng**[1]
Ye Liu[1], **Zhou Yu**[2], **Huan Wang**[1], **Silvio Savarese**[1], **Caiming Xiong**[1]
[1] Salesforce AI [2] Columbia University
jianguozhang@salesforce.com, kq2157@columbia.edu

Abstract

Despite advancements in conversational AI, language models encounter challenges to handle diverse conversational tasks, and existing dialogue dataset collections often lack diversity and comprehensiveness. To tackle these issues, we introduce DialogStudio: the largest and most diverse collection of dialogue datasets, unified under a consistent format while preserving their original information. Our collection encompasses data from open-domain dialogues, task-oriented dialogues, natural language understanding, conversational recommendation, dialogue summarization, and knowledge-grounded dialogues, making it an incredibly rich and diverse resource for dialogue research and model training. To further enhance the utility of DialogStudio, we identify the licenses for each dataset, design external knowledge and domain-aware prompts for selected dialogues to facilitate instruction-aware fine-tuning. Furthermore, we develop conversational AI models using the dataset collection, and our experiments in both zero-shot and few-shot learning scenarios demonstrate the superiority of DialogStudio. To improve transparency and support dataset and task-based research, as well as language model pre-training, all datasets, licenses, codes, and models associated with DialogStudio are made publicly accessible[1].

1 Introduction

Recent years have seen remarkable progress in Conversational AI, primarily driven by the advent of approaches and language models (Shuster et al., 2022; Zhang et al., 2023; Longpre et al., 2023; Touvron et al., 2023). Despite the advancements, these models could fall short when handling various tasks in a conversation due to the lack of comprehensive and diverse training data. Current dialogue datasets (Lin et al., 2021; Asri et al., 2017) are typically limited in size and task-specific, which thus results in suboptimal ability in task-oriented model performance. Additionally, the lack of dataset standardization impedes model generalizability.

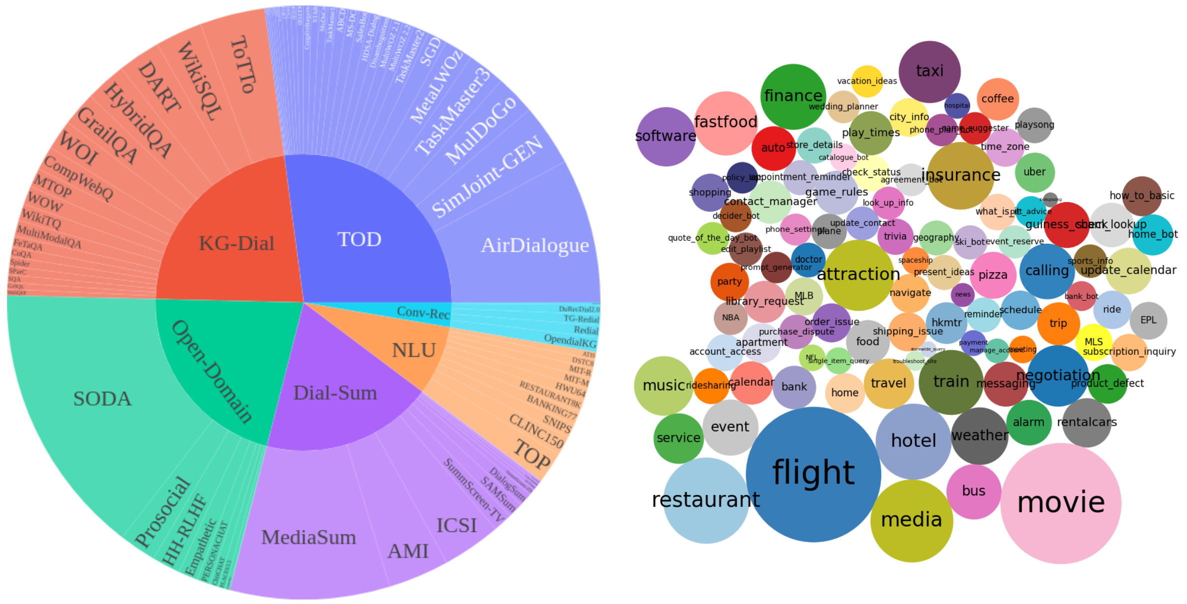
A few recent works (Gupta et al., 2022; Longpre et al., 2023; Ding et al., 2023) have introduced a large collection of datasets, which includes diverse tasks based on public datasets. For instance, FlanT5 (Longpre et al., 2023) presents the flan collections with a wide array of datasets and tasks. Despite this breadth, the coverage of dialogue datasets within the Flan collection remains notably sparse, featuring only about ten datasets. Although OPT (Iyer et al., 2022) have incorporated collections with several dialogue datasets, these collections remain inaccessible to the public. In contract, efforts like InstructDial (Gupta et al., 2022) and ParlAI (Miller et al., 2017) consist of more dialogue datasets, but they lack diversity and comprehensiveness. For instance, ParlAI mainly includes open-domain dialogue datasets, which are exclusively accessible through their platform. Other collections (Gupta et al., 2022; Kim et al., 2022a; Ding et al., 2023; Dubois et al., 2023) often distill single dataset from ChatGPT or process datasets into a sequence-to-sequence format to support language model training, featuring only input-output pairs such as dialogue context and system response. However, previous collections often overlook other crucial dialogue information, constraining their utility for research on individual datasets, tasks, and broader applications.

To overcome the aforementioned challenges, we introduce DialogStudio, the most comprehensive and diverse collection of publicly available dialogue datasets, unified under a consistent format. By aggregating dialogues from various sources, DialogStudio promotes holistic anal-

[*] Core contributors. Work completed during Kun's internship at Salesforce. Zhiwei is also a major contributor.

[1] https://github.com/salesforce/DialogStudio

Findings of the Association for Computational Linguistics: EACL 2024, pages 2299–2315
March 17-22, 2024



(a) Dataset Distribution

(b) Domain Coverage of TOD

Figure 1: (a) is the distribution of all datasets in DialogStudio. The outer and inner circle list names of datasets and the associated categories, respectively. (b) illustrates covered domains of Task-Oriented Dialogues in DialogStudio.

ysis and the development of models adaptable to a variety of conversational scenarios. The collection spans an extensive range of domains, aspects, and tasks, and it is inclusive of several categories: Open-Domain Dialogues, Task-Oriented Dialogues, Natural Language Understanding, Conversational Recommendation, Dialogue Summarization, and Knowledge-Grounded Dialogues. Thus, it can provide support for research in both individual dialogue tasks and large-scale language pre-training.

DialogStudio stands out not only for its comprehensive coverage but also for its accessibility. It offers easy access with a unified format and documents. A straightforward `load_dataset()` command through HuggingFace allows users to seamlessly interact with the collection, and we have included documentation for each dataset to enhance usability. We anticipate that this collection will enable comprehensive and standardized training and evaluations of dialogue models, fostering fair comparisons and propelling further advancements in Conversational AI.

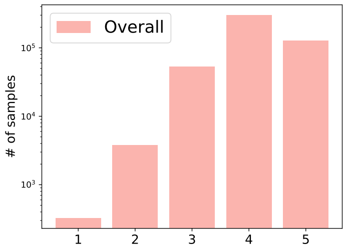
Furthermore, we identify dialogue domains, design external knowledge for available dialogues and create tailored prompts for selected datasets accordingly. Leveraging these datasets from DialogStudio, we have constructed instruction-aware models, with capacities ranging from 770M to 3B parameters. These models have the ability to handle various external knowledge and are adept at both response generation and general tasks, demonstrating the benefits of DialogStudio. The main contributions of this paper are as follows:

- We introduce DialogStudio, a meticulously curated collection of more than 80 dialogue datasets. These datasets are unified under a consistent format while retaining their original information. We integrate external knowledge, incorporate domain-aware prompts and identify dataset licenses, making DialogStudio an exceptionally rich and diverse resource for dialogue research and model training.
- We have made our datasets publicly available to enhance transparency and support research efforts. Additionally, we are committed to improving DialogStudio's usability and will persist in our efforts to refine it, ensuring an optimal user experience.
- We train conversational AI models based on DialogStudio, and these models have demonstrated superior performance over strong baselines in both zero-shot and few-shot learning scenarios.

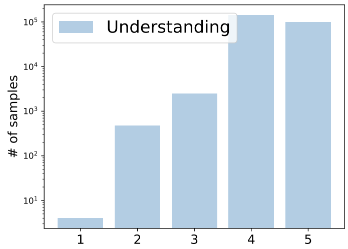
2 Data analysis

2.1 Data Visualization

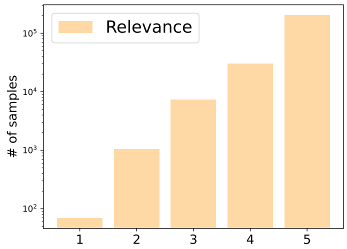
The dialogue datasets are compartmentalized into several categories: *Open-Domain Dialogues*,



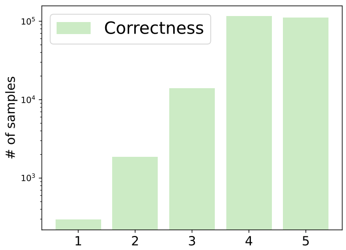
(a) Overall, $\mu = 4.14$



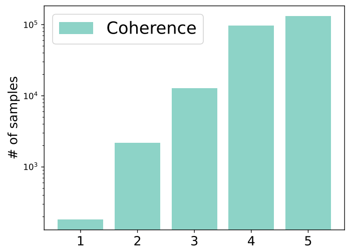
(b) Understanding, $\mu = 4.39$



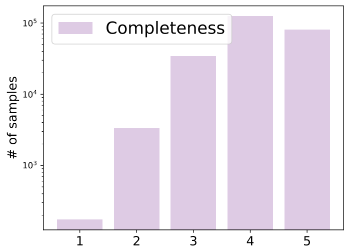
(c) Relevance, $\mu = 4.80$



(d) Correctness, $\mu = 4.38$



(e) Coherence, $\mu = 4.47$



(f) Completeness, $\mu = 4.16$

Figure 2: The score distribution for the dialogue quality.

Task-Oriented Dialogues (TOD), *Natural Language Understanding Dialogues (NLU)*, *Conversational Recommendation (Conv-Rec)*, *Dialogue Summarization (Dial-Sum)*, and *Knowledge-Grounded Dialogues (KG-Dial)*. Figure 1a presents an overview of DialogStudio's dataset categories. Note that the category boundaries are fuzzy as some datasets span multiple categories. For instance, SalesBot (Chiu et al., 2022) contains both casual and task-oriented conversations. Analogously, MultiWOZ (Budzianowski et al., 2018; Zang et al., 2020), a task-oriented dialogue corpus, incorporates knowledge bases and dialogue acts to enhance knowledge-grounded generation. Additionally, DialogStudio demonstrates its diversity by covering a wide range of domains, part of which is shown in Figure 1b.

2.2 Data Quality Investigation

Due to the existence of noise in dialogue, we develop a simple yet effective way to verify the quality of the datasets. Specifically, we employ ChatGPT (GPT-3.5-turbo) to evaluate the quality of system responses based on severall perspectives (Mehri et al., 2022; Kim et al., 2022a), *i.e.*, Understanding, Relevance, Correctness, Coherence, Completeness and Overall quality. Understanding assesses whether the model's responses accurately reflect the meaning and intent of the user's inputs. Relevance demonstrates whether the generated response should be directly related and appropriate to the preceding user input and the context of the conversation. Coherence measures the logical consistency of the model's responses within the context of the conversation. Completeness refers to whether the system's responses fully address the user's queries or tasks. Overall quality comprehensively rates the quality of dialogue. All scores are in the range of 1-5, and higher scores should only be given to truly exceptional examples. We delicately design the prompt and ask the ChatGPT model to *strictly* rate the score.

Since there are a lot of datasets in DialogStudio, we randomly select 33 multi-turn dialogue datasets and evaluate all the training dialogues of each dataset. To harmonize ChatGPT and human ratings, we take a random sample of 50 training dialogues from each dataset. These were then rated by three expert researchers using the five specified criteria. Post-alignment of ChatGPT and human evaluations, we view dialogues with a score above 3 as being of high quality. Figure 2 illustrates distributions of those scores. We also reveal the average score as the μ in each sub-caption. In general, the dialogues show high qualities regarding to the individual criteria and the overall quality.

3 Datasets Unification and Access

We collect and process a wide range of datasets, involving different domains, types, and tasks.

Since these datasets originally contain various information and format, we propose a unification strategy to process all the datasets such that they can be loaded in the same data loader.

3.1 Unification

Before unifying the format of those datasets, we fixed several issues as follows: 1) we remove those dialogues labeled as multi-turn dialogues, but actually with only one turn and miss either user utterance or system utterance. 2) We manually check the individual dialogues. If one dialogue contains one or more empty user or system utterances, we fill utterances based on corresponding dialogue contexts, dialogue acts, and dialogue information. In total, less than 0.5% of dialogues had these issues. To support research interest on individual datasets, we have flagged and rectified these problematic dialogues.

Additionally, we recognize the success of instruction tuning for dialogue models and thus we manually pre-define five different prompt templates for multi-turn dialogue datasets, such as *This is a bot helping users to {Task_Domain}. Given the dialogue context and external database, please generate a relevant system response for the user.* The {*Task_Domain*} is associated with the dialogue domain and we manually create a corresponding description. For example, if a dialogue is of domain *travel*, we set {*Task_Domain*} as *book a trip*. A concrete example of the prompt is demonstrated in Figure 3. Moreover, many datasets lack a direct mapping between dialogues and their domain information. To address this, we determine the domain of each dialogue using its intent, schema, APIs, and associated databases.

Next, we construct a uniform JSON dictionary format to store all relevant information of each dialogue as illustrated in Figure 3. Compared with existing works, DialogStudio covers more dialogue information and is easier to retrieve the information for arbitrary dialogue-related tasks. Concretely, we include all dialogue-related information, such as the dialogue ID, data split label, domain, task, and content. Additionally, we identify the external knowledge, dialogue state tracking (DST) knowledge, and intent knowledge in the dialogue, which are the most beneficial knowledge for a dialogue.

Regarding external knowledge, we construct it based on information such as databases and dialogue acts. Since each dialogue dataset focuses on specific tasks or domains and has a different database and annotation schema, we unify such information into *external knowledge*. For example, if the user is looking for a hotel and asking for its address, the system response should be based on both the search results from the database and the dialogue context. To simulate the realistic situation and avoid directly providing the model with the ground truth resulting hotel, we also randomly sample four other candidate results and mix them with the ground truth result. All information is flattened and converted into a string as external knowledge.

To complete tasks and generate coherent responses, a dialogue system needs to track users' requirements for the task. Those requirements are usually represented as dialogue states. For example, regarding the hotel booking task, a dialogue system needs to extract information such as price range and locations to enable searching hotels in the database. The type of dialogue states varies across different tasks and datasets. As such, it is hard for dialogue systems to predict the values of those dialogue states if unknowing the specific dialogue states the task covers. Therefore, we propose to insert the schema, consisting of predefined dialogue state types and values for each task, into the input sequence. For datasets like SGD (Rastogi et al., 2020), which already provides annotation schema, we directly convert the dictionary-structured schema into a string. For the rest datasets that have no such schema file, we iterate over all dialogues and collect potential state annotations to construct a schema. We provide domains, slot types, and slot values in the schema string. For those categorized dialogue slots like "hotel star-ratings", which have a fixed number of candidate values, we provide all possible values. For others that have unlimited possible values, e.g. "stay night", we randomly sample ten values, such that a model can learn what slot values are relevant to these slot types. We put the turn-level ground-truth DST information in "dst", and the general DST information under "dst knowledge", as presented in Figure 3.

Analogously, intent prediction also requires models to know all possible intent types for each task. Therefore, we extract the schema directly from the schema file if it exists. As to datasets without schema, we also iterate over all dialogue

```
"dialogue_id": "train_1",
"num_utterances": 14,
"utterances": [
  {
    "speaker": "USR",
    "text": "I'd like to book a trip to Atlantis from Caprica on
            Saturday, August 13, 2016 for 8 adults.",
    "ap_label": "",
    "da_label": "inform"
  },
  {
    "speaker": "USR",
    "text": "I have a tight budget of 1700.",
    "ap_label": "",
    "da_label": "inform"
  },
  {
    "speaker": "SYS",
    "text": "Hi...I checked a few options for you, and we do
            not currently have any trips that meet this criteria.",
    "ap_label": "",
    "da_label": "sorry",
    "slots": {
      "dst_city": "Atlantis",
      "or_city": "Caprica",
      "str_date": "Saturday, August 13, 2016",
      "n_adults": "8",
      "budget": "1700"
    }
  }
],
"scenario": {
  "db_id": "U22HTHYNP",
  "db_type": "booking",
  "task": "book"
}
```


(a) Original Data


```
"FRAMES--train--1": {
    "original dialog id": "train_1",
    "dialog index": 1,
    "original dialog info": {
      "scenario": {
        "db_id": "U22HTHYNP",
        "db_type": "booking",
        "task": "book"}}
    "log": [
      {
        "turn id": 1,
        "user utterance": "I'd like to book a trip to Atlantis from Caprica on Saturday,
                          August 13, 2016 for 8 adults. I have a tight budget of 1700.",
        "system response": "Hi...I checked a few options for you, and we do not currently
                           have any trips that meet this criteria.",
        "dialog history": "",
        "original user side information": {
          "da_label": "inform"
        },
        "original system side information": {
          "da_label": "sorry",
          "slots": {
            "dst_city": "Atlantis",
            "or_city": "Caprica",
            "str_date": "Saturday, August 13, 2016",
            "n_adults": "8",
            "budget": "1700"
          }
        },
        "intent": "inform",
        "dst": "book dst_city Atlantis, book or_city Caprica, book str_date Saturday, August
               13, 2016, book n_adults 8, book budget 1700"
      }
    ]
    "external knowledge": "( travel : (( trip : ( returning : ( duration : ( hours : 0 | min : 51...",
    "dst knowledge": " ( book : ( dst_city : ( Indianapolis | St. Loius | Le Paz | …) | or_city : (
                     PUebla | sf | toluca | San Francisco…",
    "intent knowledge": "( book : ( null | negate | request | goodbye | affirm))…",
    "prompt": [
      "This is a bot helping users to book a trip. Given the dialog context and external
       database, please generate a relevant system response for the user."
    ]
}
```


(b) DialogStudio Data

Figure 3: A dialogue format example. Left: original example, right: converted example. Here we only show the first turn and partial information.

in the dataset to collect all potential intents. Then, we put the turn-level ground-truth intent information into ”intent”, and the general intents under ”intent knowledge”, as presented in Figure 3. Note that not all datasets provide detailed annotation for dialogue states, intents, or even databases. For dialogue state tracking and intent classification tasks, we only process dialogues with corresponding annotations. Since all data is used for response generation, we leave the external knowledge value for the database blank if there is no related database in the original dataset.

3.2 Access and Maintenance

As aforementioned in the format, our DialogStudio data is easy to access via the JSON files. To make DialogStudio more maintainable and accessible, we will publish datasets on both GitHub and HuggingFace. GitHub mainly stores selected dialogue examples and relevant documents. We sample five original dialogues and five converted dialogues for each dataset to facilitate users in comprehending our format and examining the contents of each dataset. The complete DialogStudio dataset is maintained in our HugginFace repository, where all the datasets can be directly downloaded or loaded with the HuggingFace `load_dataset(dialogstudio, dataset_name)` API. Given the substantial volume of datasets, optimizing user experience poses a challenge and limitation. We will continuously maintain and update both GitHub and HuggingFace. DialogStudio is built upon public research datasets without individual or private information. We believe it is important to clearly present the license associated with each of these datasets. Consequently, we have included the original licenses for all datasets. All these datasets are supportive of academic research, and some of them also endorse commercial usage. The code that we employ falls under the widely accepted Apache 2.0 license. While we strictly require adherence to the respective dataset licenses for all intended usages on DialogStudio, there remains a possibility that some works might not fully comply with the licenses.

Regarding the other concerns such as ethical concern, we admit that DialogStudio is collected and maintained by the authors of this work and we did not hire external annotators. Since it contains unified datasets across several categories, it supports various research purposes from individual tasks and datasets to language model pre-training.

4 Experiments

In this section, we present the pre-training details, methodologies, and metrics used to assess the performance of our DialogStudio model. The evaluation process aims to measure the model's ability to both solve task-oriented dialogues and understand general prompt-based instruction.

4.1 Model Pre-training

In this section, we introduce more details about how we conduct our pre-training. In regards of training models, we mix several datasets from DialogStudio.

For task-oriented and conversational recommendation datasets, we selected dialogues from a range of sources including KVRET (Eric et al., 2017), AirDialogue (Wei et al., 2018), DSTC2-Clean (Mrkšić et al., 2017), CaSiNo (Chawla et al., 2021), FRAMES (El Asri et al.), WOZ2.0 (Mrkšić et al., 2017), CraigslistBargains (He et al., 2018), Taskmaster1-2 (Byrne et al., 2019), ABCD (Chen et al., 2021a), MulDoGO (Peskov et al., 2019), BiTOD (Lin et al., 2021), SimJoint (Shah et al., 2018), STAR (Mosig et al., 2020), SGD (Rastogi et al., 2020), OpenDialKG (Moon et al., 2019) and DuRecDial-2.0 (Liu et al., 2021).

Meanwhile, for knowledge-grounded dialogues, we drew upon dataset from SQA (Iyyer et al., 2017), SParC (Yu et al., 2019b), FeTaQA (Nan et al., 2022), MultiModalQA (Talmor et al., 2021), CompWebQ (Talmor and Berant, 2018), CoSQL (Yu et al., 2019a).

For open-domain dialogues, we sample dialogues from SODA (Kim et al., 2022a), ProsocialDialog (Kim et al., 2022b), Chitchat (Myers et al., 2020).

For each dialogue dataset, we sample at most 11k dialogues. Additionaly, we extracted around 11k dialogue turns from question-answering dialogues featured in RACE (Lai et al., 2017), NarrativeQA (Kočiskỳ et al., 2018), SQUAD (Rajpurkar et al., 2018), MCtest (Richardson et al., 2013), OpenBookQA (Mihaylov et al., 2018), MultiRC (Khashabi et al., 2018). Here, a dialogue turn refers to a pair consisting of a dialogue context and its corresponding system response. The rest datasets in DialogStudio are preserved for future evaluations and downstream fine-tuning.

For each dialogue during the training, we shape the available external knowledge into a string, which is included in dialogue context, and instruct the model to generate a dialogue response based on external knowledge. We use the format *Instruction \n <USER> user utterance <SYSTEM> system response <USER> ... <USER> user utterance \n <EXTERNAL KNOWLEDGE> supported knowledge* to train the model, where *<USER>*, *<SYSTEM>* and *<EXTERNAL KNOWLEDGE>* are special tokens.

We follow the public HuggingFace transformer code (Wolf et al., 2020; Wang et al., 2022) to train the model. For initializing our models, we adopt T5 (Raffel et al., 2020) and Flan-T5 (Longpre et al., 2023) as starting points to respectively establish DialogStudio-T5 and DialogStudio-Flan-T5. For the training of DialogStudio-Flan-T5, we exclude all translation-oriented tasks, limiting the sample size for each Flan task to a maximum of 150 examples. This leads to a cumulative total of 140,000 samples. We train the model up to 3 epochs with bfloat16 precision, a total batch size of 64. We set a constant learning rate 5e-5 and 3e-5 for the large model and the 3B model, respectively. Experiments are conducted using 16 A100 GPUs, each with 40GB of GPU memory.

4.2 Evaluation for Response Generation

Settings. We evaluate the performance on CoQA (Reddy et al., 2019) and MultiWOZ 2.2 (Zang et al., 2020). CoQA is a multi-turn conversational question answering dataset with 8k conversations about text passages from seven diverse domains. MultiWOZ 2.2 is one of the largest and most widely used multi-domain task-oriented dialogue corpora with more than 10000 dialogues. Each dialogue involves with one or more domains such as *Train, Restaurant, Hotel, Taxi,* and *Attraction*. The dataset is challenging and complex due to the multi-domain setting and diverse linguistic styles. Note that we exclude both datasets during the pre-training stage to prevent data leakage.

	CoQA		MultiWOZ	
	ROUGE-L	F1	ROUGE-L	F1
Flan-T5-3B (Longpre et al., 2023)	37.1	37.2	7.0	7.4
Flan-T5-Large (Longpre et al., 2023)	22.5	22.3	15.9	17.6
GODEL-Large (Peng et al., 2022)	43.2	43.3	18.5	19.3
DialogStudio-T5-Large	61.2	61.5	32.4	34.5
DialogStudio-Flan-T5-Large	63.3	63.5	33.7	35.9

Table 1: Zero-shot results on CoQA and MultiWOZ 2.2.

	CR (14 tasks)	DAR (7 tasks)	TE (27 tasks)	avg. (48 tasks)
OPT-30B (Zhang et al., 2022b)	21.3/1.1	35.2/4.1	40.3/0.9	32.3/2.0
OPT-IML-30B (Iyer et al., 2022)	37.4/41.6	51.4/51.8	54.7/47.8	47.9/47.1
OPT-175B (Zhang et al., 2022b)	21.0/4.2	37.1/16.8	41.6/2.2	33.3/7.7
OPT-IML-175B (Iyer et al., 2022)	39.0/49.8	**61.2**/60.2	54.3/51.0	**51.5**/53.6
Tk-INSTRUCT-11B (Wang et al., 2022)	32.3/**62.3**	51.1/**69.6**	**55.0/64.1**	46.1/**65.3**
Tk-INSTRUCT-3B (Wang et al., 2022)	38.4/51.3	45.7/58.5	48.4/52.8	44.2/54.2
DialogStudio-NIV2-T5-3B	**41.3**/59.8	57.5/63.7	52.3/55.1	50.4/59.5

Table 2: 0-shot/2-shot/5-shot ROUGE-L testing results on unseen datasets and unseen tasks. Results of baselines are reported by original papers. CR, DAR, and TE, avg. are abbreviations for Coreference Resolution, Dialogue Act Recognition, Textual Entailment, and average results, respectively.

For CoQA, we follow the original paper setting to answer question based on external passage. For MultiWOZ 2.2, we consider the lexicalized dialogue-act-to-response generation task where the model needs to consider both the dialogue context and the dialogue acts during generation. We follow the prompt from (Bang et al., 2023) to instruct models, i.e., *Continue the dialogue as a task-oriented dialogue system called SYSTEM. The answer of SYSTEM should follow the ACTION provided next while answering the USER's last utterance*.

We focus on zero-shot evaluation and report the ROUGE-L and F1 score (Miller et al., 2017), where ROUGE-L measures the longest common subsequence and F1 measures the Unigram F1 overlap between the prediction and ground-truth response.

Baselines. We consider GODEL (Peng et al., 2022) and Flan-T5 (Longpre et al., 2023) as our baselines. GODEL is a T5-based large pre-trained model for goal-oriented dialogues. It is pre-trained with 551M multi-turn Reddit dialogues and 5M knowledge-grounded and question-answering dialogues. Flan-T5 is an instruction-aware model. It is also initialized from T5 and pre-trained on the Flan collection, which consists of more than 1800 tasks and 400 datasets, including dialogue datasets.

Results. Table 1 depicts the results from both zero-shot and few-shot testing. Evidently, our models considerably surpass the baseline models in terms of zero-shot learning, exhibiting a robust generalized ability for response generation in a zero-shot scenario.

Flan-T5-3B, on the other hand, underperforms in the task of generating responses from dialog-acts. This model tends to produce incorrect dialog acts, unnatural utterances, or terminates with an empty end token. One explanation for this is that Flan-T5 models did not receive sufficient dialogue training during the instruction-training phase on the Flan collections. Comparisons between the performances of existing models before and after training on the unified dataset validate DialogStudio's usefulness.

4.3 Evaluation on Super-NaturalInstructions

Settings. NIV2 (Wang et al., 2022) introduces an instruction-tuning benchmark with more than 1600 tasks. We select 3 categories with 44 tasks from the held-out test set, which consists of 154

	MMLU		BBH
	0-SHOT	5-SHOT	3-SHOT
TK-INSTRUCT 11B (Wang et al., 2022)	-	41.1	32.9
LLAMA 13B (Touvron et al., 2023)	-	46.2	37.1
Vicuna 13B (Chiang et al., 2023)	-	49.7	37.1
Flan-T5-Large (Longpre et al., 2023)	41.5	41.9	37.1
Flan-T5-XL (Peng et al., 2022)	48.7	49.3	40.2
OPT-IML-Max 30B (Iyer et al., 2022)	46.3	43.2	31.3
DialogStudio-Flan-T5-Large	40.1	40.9	34.2
DialogStudio-Flan-T5-3B	48.3	47.8	40.3

Table 3: Test results on MMLU and BBH. Results come from original papers and InstructEval (Chia et al., 2023).

tasks, i.e., Coreference Resolution, Dialogue Act Recognition, and Textual Entailment. The selected tasks and datasets are unseen in the training stage. Specifically, we strictly follow all settings including metrics in (Wang et al., 2022), i.e., train models with instructions + 2 positive demonstrations and no negative demonstrations. We finetune DialogStudio-T5-3B on 756 training tasks.

Baselines. OPT (Zhang et al., 2022b) is a set of open decoder-only transformer models pre-trained on 180B tokens. OPT-IML (Iyer et al., 2022) is an instruction meta-learning model based on the OPT-IML bench with more than 1500 tasks. Tk-INSTRUCT (Wang et al., 2022) is initialized from T5 and further pre-trained based on NIV2 collections. Note that we neglect Flan-T5 because it trains with all the downstream datasets and tasks.

Results. Table 2 shows the 0-shot and 2-shot results on unseen datasets and unseen tasks. Based on the average performance on 48 tasks, DialogStudio-NIV2-T5-3B outperforms OPT-IML-175B by 5.9% on 2-shot learning with more than 50 times fewer model parameters, and it surpasses Tk-INSTRUCT-11B by 4.3% on 0-shot learning with more than 3 times fewer parameters. The performance demonstrates the strong generalization ability of DialogStudio model. Compared with Tk-INSTRUCT-3B, DialogStudio-NIV2-T5-3B achieves 6.2% and 5.3% improvements on 0-shot and 2-shot learning respectively. Given that both Tk-INSTRUCT and our DialogStudio-NIV2-T5-3B are fine-tuned from the T5 model, this improvement indicates the effectiveness of pre-training with our DialogStudio collection.

4.4 Evaluation on MMLU and BBH

Table 3 presents results on MMLU (Hendrycks et al., 2020) and Big Bench Hard (BBH) (Srivastava et al., 2022). When comparing the performance of DialogStudio-Flan-T5 with Flan-T5, we observe a minor decrease. However, this is accompanied by a significant improvement in dialogue relevant capabilities.

4.5 Evaluation on Alternative Benchmarks

DialogStudio encompasses not just public realistic dialogue datasets, but also those derived from or shared with ChatGPT, such as SODA (Kim et al., 2022a) and ShareGPT. Due to GPU constraints, we employ techniques like LoRA (Hu et al., 2021) to fine-tune llama (Touvron et al., 2023). When using equivalent datasets from DialogStudio, we observed performance comparable to other models, e.g., Vicuna (Chiang et al., 2023), on benchmarks like AlpacaEval (Dubois et al., 2023) and MT-Bench (Zheng et al., 2023). This demonstrates that DialogStudio caters to research interests in both specific datasets and generalized instruction tuning.

5 CONCLUSION

In this study, we have introduced DialogStudio, a comprehensive collection that aggregates more than 80 diverse dialogue datasets while preserving their original information. This aggregation not only represents a significant leap towards consolidating dialogues from varied sources but also offers a rich tapestry of conversational patterns, intents, and structures, capturing the nuances and richness of human interaction. Utilizing DialogStudio, we developed corresponding models, demonstrating superior performance in both zero-shot and few-shot learning scenarios. In the spirit of open research and advancing the field, we are committed to releasing DialogStudio to the broader research community.

References

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv:1704.00057*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 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 (EMNLP)*.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

Iñigo Casanueva, Ivan Vulić, Georgios Spithourakis, and Paweł Budzianowski. 2022. Nlu++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1998–2013.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. Casino: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021a. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3002–3017.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. Places: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 814–838.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022a. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021b. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul A Crook, and William Yang Wang. 2022b. Ketod: Knowledge-enriched task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Ssu Chiu, Maolin Li, Yen-Ting Lin, and Yun-Nung Chen. 2022. Salesbot: Transitioning from chit-chat to task-oriented dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6143–6158.

Samuel Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. 2020. Span-convert: Few-shot span extraction for dialog with pretrained conversational representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 107–121.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur D. Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail S. Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (convai2). *ArXiv*, abs/1902.00098.

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*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. Frames: A corpus for adding memory to goal-oriented dialogue systems.

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 Twelfth Language Resources and Evaluation Conference*, pages 422–428.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880, Online. Association for Computational Linguistics.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. TWEETSUMM - a dialog summarization dataset for customer service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 245–260, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pages 3477–3488.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey P Bigham. 2022. Instructdial: Improving zero and few-shot generalization in dialogue through instruction tuning. *arXiv preprint arXiv:2205.12673*.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2787–2792.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. Inspired: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022a. Soda: Million-scale dialogue distillation with social commonsense contextualization. *ArXiv*, abs/2212.10465.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022b. Prosocialdialog: A prosocial backbone for conversational agents. *arXiv preprint arXiv:2205.12688*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316.

S Lee, H Schulz, A Atkinson, J Gao, K Suleman, L El Asri, M Adada, M Huang, S Sharma, W Tay, et al. 2019. Multi-domain task-completion dialog challenge. *Dialog system technology challenges*, 8(9).

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018a. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*.

Xiujun Li, Yu Wang, Siqi Sun, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018b. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv preprint arXiv:1807.11125*.

Yu Li, Kun Qian, Weiyan Shi, and Zhou Yu. 2020. End-to-end trainable non-collaborative dialog system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8293–8302.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. *NeurIPS 2021 Track on Datasets and Benchmarks*.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8386–8390. IEEE.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. *arXiv preprint arXiv:1903.05566*.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. Durecdial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Scott Martin, Shivani Poddar, and Kartikeya Upasani. 2020. Mudoco: corpus for multidomain coreference resolution and referring expression generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 104–111.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, et al. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges. *arXiv preprint arXiv:2203.10012*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Johannes EM Mosig, Shikib Mehri, and Thomas Kober. 2020. Star: A schema-guided dialog dataset for transfer learning. *arXiv preprint arXiv:2010.11853*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Will Myers, Tyler Etchart, and Nancy Fulda. 2020. Conversational scaffolding: An analogy-based approach to response prioritization in open-domain dialogs.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, et al. 2022. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, et al. 2021. Dart: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. *arXiv preprint arXiv:2206.11309*.

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (multidogo): Strategies toward curating and annotating large scale dialogue data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4526–4536.

Kun Qian, Ahmad Beirami, Zhouhan Lin, Ankita De, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2021. Annotation inconsistency and entity bias in multiwoz. *ArXiv*, abs/2105.14150.

Kun Qian, Satwik Kottur, Ahmad Beirami, Shahin Shayandeh, Paul A Crook, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2022. Database search results disambiguation for task-oriented dialog systems. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1158–1173.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. Gecor: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4547–4557.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134, Online. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *ACL*.

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.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Virgile Rennard, Guokan Shang, Julie Hunter, and Michalis Vazirgiannis. 2023. Abstractive meeting summarization: A survey. *Transactions of the Association for Computational Linguistics*, 11:861–884.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203.

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*.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *ICLR*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Supernaturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. Airdialogue: An environment for goal-oriented dialogue research. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3844–3854.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206.

Tao Yu, Rui Zhang, Heyang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, et al. 2019a. Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1962–1979.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, et al. 2019b. Sparc: Cross-domain semantic parsing in context. *arXiv preprint arXiv:1906.02285*.

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.

Jianguo Zhang, Kazuma Hashimoto, Yao Wan, Zhiwei Liu, Ye Liu, Caiming Xiong, and S Yu Philip. 2022a. Are pre-trained transformers robust in intent classification? a missing ingredient in evaluation of out-of-scope intent detection. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 12–20.

Jianguo Zhang, Stephen Roller, Kun Qian, Zhiwei Liu, Rui Meng, Shelby Heinecke, Huan Wang, Silvio Savarese, and Caiming Xiong. 2023. Enhancing performance on seen and unseen dialogue scenarios using retrieval-augmented end-to-end task-oriented system. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–518.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022b. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In *North American Association for Computational Linguistics (NAACL)*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*.

Appendix

Table 4 and Table 5 lists datasets included in DialogStudio. Initially, we present a partial list of these datasets. More and latest information are available in GitHub[2].

[2] https://github.com/salesforce/DialogStudio

NLU	NLU++ (Casanueva et al., 2022)
	BANKING77-OOS (Zhang et al., 2022a)
	BANKING77 (Casanueva et al., 2020)
	RESTAURANTS8K (Coope et al., 2020)
	CLINC150 (Larson et al., 2019)
	CLINC-Single-Domain-OOS-banking (Zhang et al., 2022a)
	CLINC-Single-Domain-OOS-credit_cards (Zhang et al., 2022a)
	HWU64 (Liu et al., 2019)
	SNIPS (Coucke et al., 2018)
	SNIPS-NER (Coucke et al., 2018)
	DSTC8-SGD (Coope et al., 2020)
	TOP (Gupta et al., 2018)
	TOP-NER (Gupta et al., 2018)
	ATIS-NER (Hemphill et al., 1990)
	ATIS (Hemphill et al., 1990)
	MIT-MOVIE (Liu et al., 2013)
	MIT-RESTAURANT (Liu et al., 2013)
TOD	KVRET (Eric et al., 2017)
	AirDialogue (Wei et al., 2018)
	DSTC2-Clean (Mrkšić et al., 2017)
	CaSiNo (Chawla et al., 2021)
	FRAMES (El Asri et al.)
	WOZ2.0 (Mrkšić et al., 2017)
	CraigslistBargains (He et al., 2018)
	Taskmaster1 (Byrne et al., 2019)
	Taskmaster2 (Byrne et al., 2019)
	Taskmaster3 (Byrne et al., 2019)
	ABCD (Chen et al., 2021a)
	MulDoGO (Peskov et al., 2019)
	BiTOD (Lin et al., 2021)
	SimJointGEN (Shah et al., 2018)
	SimJointMovie (Shah et al., 2018)
	SimJointRestaurant (Shah et al., 2018)
	STAR (Mosig et al., 2020)
	SGD (Rastogi et al., 2020)
	MultiWOZ2_1 (Eric et al., 2020)
	MultiWOZ2_2 (Zang et al., 2020)
	MultiWOZ2_2+ (Qian et al., 2021)
	HDSA-Dialog (Chen et al., 2021a)
	MS-DC (Li et al., 2018b)
	GECOR (Quan et al., 2019)
	Disambiguation (Qian et al., 2022)
	MetaLWOZ (Lee et al., 2019)
	KETOD (Chen et al., 2022b)
	MuDoCo (Martin et al., 2020)

Table 4: List of datasets included in DialogStudio (a).

KG-Dial	SQA (Iyyer et al., 2017) SParC (Yu et al., 2019b) FeTaQA (Nan et al., 2022) MultiModalQA (Talmor et al., 2021) CompWebQ (Talmor and Berant, 2018) CoSQL (Yu et al., 2019a) CoQA (Reddy et al., 2019) Spider (Yu et al., 2018) ToTTo (Parikh et al., 2020) WebQSP (Yih et al., 2016) WikiSQL (Zhong et al., 2017) WikiTQ (Pasupat and Liang, 2015) DART (Nan et al., 2021) GrailQA (Gu et al., 2021) HybridQA (Chen et al., 2020) MTOP (Chen et al., 2020) UltralChat-Assistance (Ding et al., 2023) Wizard_of_Wikipedia (Dinan et al., 2018) Wizard_of_Internet (Komeili et al., 2022)
Dial-Sum	TweetSumm (Feigenblat et al., 2021) SAMSum (Gliwa et al., 2019) DialogSum (Chen et al., 2021b) AMI (Kraaij et al., 2005; Rennard et al., 2023) ICSI (Janin et al., 2003) QMSum (Zhong et al., 2021) MediaSum (Zhu et al., 2021) ECTSum (Mukherjee et al., 2022) SummScreen_ForeverDreaming (Chen et al., 2022a) SummScreen_TVMegaSite (Chen et al., 2022a) CRD3 (Rameshkumar and Bailey, 2020) ConvoSumm (Fabbri et al., 2021)
Open-Domain	ChitCHAT (Myers et al., 2020) SODA (Kim et al., 2022a) Prosocial (Kim et al., 2022b) HH-RLHF (Bai et al., 2022) Empathetic (Rashkin et al., 2019) ConvAI2 (Dinan et al., 2019) AntiScam (Li et al., 2020) ShareGPT (Zheng et al., 2023) PLACES3.5 (Chen et al., 2023)
Conv-Rec	SalesBot (Chiu et al., 2022) Redial (Li et al., 2018a) Inspired (Hayati et al., 2020) DuRecDial 2.0 (Liu et al., 2021) OpendialKG (Moon et al., 2019)

Table 5: List of datasets included in DialogStudio (b).