InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning

Prakhar Gupta[♣] **Cathy Jiao**[♣] **Yi-Ting Yeh**[♣] **Shikib Mehri**[♣]
Maxine Eskenazi[♣] **Jeffrey P. Bigham**[♣,♡]
[♣]Language Technologies Institute, Carnegie Mellon University
[♡]Human-Computer Interaction Institute, Carnegie Mellon University
{prakharg,cljiao,yitingye,amehri,max,jbigham}@cs.cmu.edu

Abstract

Instruction tuning is an emergent paradigm in NLP wherein natural language instructions are leveraged with language models to induce zero-shot performance on unseen tasks. Instructions have been shown to enable good performance on unseen tasks and datasets in both large and small language models. Dialogue is an especially interesting area to explore instruction tuning because dialogue systems perform multiple kinds of tasks related to language (e.g., natural language understanding and generation, domain-specific interaction), yet instruction tuning has not been systematically explored for dialogue-related tasks. We introduce INSTRUCTDIAL, an instruction tuning framework for dialogue, which consists of a repository of 48 diverse dialogue tasks in a unified text-to-text format created from 59 openly available dialogue datasets. Next, we explore cross-task generalization ability on models tuned on INSTRUCTDIAL across diverse dialogue tasks. Our analysis reveals that INSTRUCTDIAL enables good zero-shot performance on unseen datasets and tasks such as dialogue evaluation and intent detection, and even better performance in a few-shot setting. To ensure that the models adhere to instructions, we introduce novel meta-tasks. We establish benchmark zero-shot and few-shot performance of models trained using the proposed framework on multiple dialogue tasks.[1]

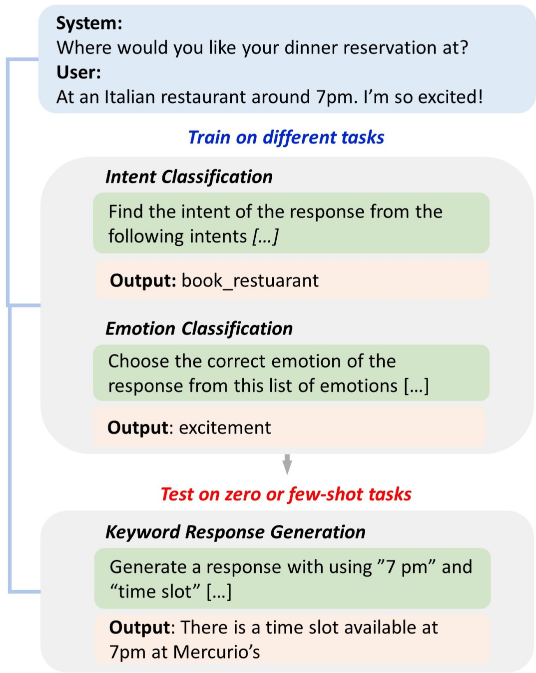

Figure 1: In this work, we investigate instruction tuning on dialogue. Instruction tuning involves training a model on a mixture of tasks defined through natural language instructions. The model then exhibits zero-shot or few-shot generalization to new tasks.

1 Introduction

Pretrained large language models (LLMs) (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020) are not only few-shot learners, but can also perform numerous language tasks without the need for fine-tuning. However, LLMs are expensive to train and test. Instruction tuning has emerged as a tool for directly inducing zero-shot generalization on unseen tasks in language models by using natural language instructions (Mishra et al., 2021; Sanh et al., 2022; Wei et al., 2022; Ouyang et al., 2022). Natural language instructions can contain components such as task definitions, examples, and prompts which allow different instructions to be created for multitask learning. Instruction tuning enables developers, practitioners, and even non-expert users to leverage language models for novel tasks by specifying them through natural language, without the need for large training datasets. Furthermore, instruction tuning can work for models that are significantly smaller than LLMs (Mishra et al., 2021; Sanh et al., 2022), making them more practical and affordable.

Most recent work (Mishra et al., 2021; Sanh

[1]Code available at https://github.com/prakharguptaz/Instructdial

et al., 2022; Wei et al., 2022) on instruction tuning has focused on general NLP tasks such as paraphrase detection and reading comprehension, and not specifically on dialogue, with the majority being classification tasks. While some work such as Wang et al. (2022a) include a few dialogue tasks, those tasks are collected through crowdsourcing and do not provide good coverage over the variety of dialogue tasks and domain. No prior work has examined how training a model on a wide variety of dialogue tasks with a variety of instructions may affect a system's ability to perform both core dialogue tasks, such as intent detection, and domain-specific tasks, such as emotion classification. In this work, we introduce INSTRUCTDIAL, a framework for instruction tuning on dialogue tasks. We provide a large curated collection of 59 dialogue datasets and 48 tasks, benchmark models, and a suite of metrics for testing the zero-shot and few-shot capabilities of the models.

The INSTRUCTDIAL repository consists of multiple dialogue tasks converted into a text-to-text format. In particular, we include dialogue generation, classification, and evaluation tasks. In addition, we include task-oriented and open-ended dialogue tasks drawn from a variety of domains (Figure 1). We release INSTRUCTDIAL to facilitate the development of dialogue models that can generalize to unseen tasks through natural language instructions.

Furthermore, we address general issues in instruction tuning with respect to our proposed work. Webson and Pavlick (2021) found that instruction tuned models may ignore instructions and attain the same performance with irrelevant prompts. We address this issue in two ways: (1) we train the models with a variety of outputs given the same input context by creating multiple task formulations, and (2) we propose two instruction-specific meta-tasks (e.g., select an instruction that matches with an input-output pair) to encourage models to adhere to the instructions.

The main contributions of this paper are:

- We introduce INSTRUCTDIAL, a framework to systematically investigate instruction tuning in the dialogue domain on a large collection of dialogue datasets (59 datasets) and tasks (48 tasks) for both task-oriented and open-domain dialogue. Our framework will be open-sourced to allow easy incorporation and configuration of new datasets and tasks;

- Through extensive experiments we show that instruction-tuning models on a variety of dialogue tasks enhances zero-shot and few-shot performance on different tasks; and,

- We provide various analyses and establish baseline and upper bound performance for multiple tasks. We also provide implementations or integration of various task-specific dialogue metrics.

Our experiments reveal further room for improvement. This includes work to make performance invariant to instruction wording and task interference. We hope that INSTRUCTDIAL will facilitate further progress on instruction-tuning systems for dialogue tasks.

2 Related Work

Pre-training and Multi-Task learning in Dialogue Large-scale transformer models (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020) pre-trained on massive text corpora have brought substantial performance improvements in natural language processing. Similar trends have occurred in the dialogue domain, where models such as DialoGPT (Zhang et al., 2020), Meena (Adiwardana et al., 2020), Blenderbot (Roller et al., 2021) and PLATO (Bao et al., 2020, 2021) trained on sources such as Reddit, Twitter, Weibo, or on human-annotated datasets show great capabilities in carrying open-domain conversations. Large-scale pretraining has also shown success in task-oriented dialogue (TOD). (Budzianowski and Vulić, 2019; Hosseini-Asl et al., 2020; Ham et al., 2020; Lin et al., 2020; Yang et al., 2021) utilized pretrained language models such as GPT-2 to solve TOD tasks such as language generation or act prediction. Similarly, BERT-type pretrained models have been used for language understanding in TOD tasks (Wu et al., 2020a; Mi et al., 2021b). Several of these works (Hosseini-Asl et al., 2020; Yang et al., 2021; Liu et al., 2022; Su et al., 2022a) have shown improved performance by performing multi-task learning over multiple tasks. Multi-task pretraining also helps models learn good few-shot capabilities (Wu et al., 2020a; Peng et al., 2021). Our work covers both open-domain and TOD tasks and goes beyond multi-tasking as it incorporates additional structure regarding task definitions and constraints through instructions.

Instruction tuning Constructing natural language prompts to perform NLP tasks is an active area of research (Schick and Schütze, 2021; Liu et al.,

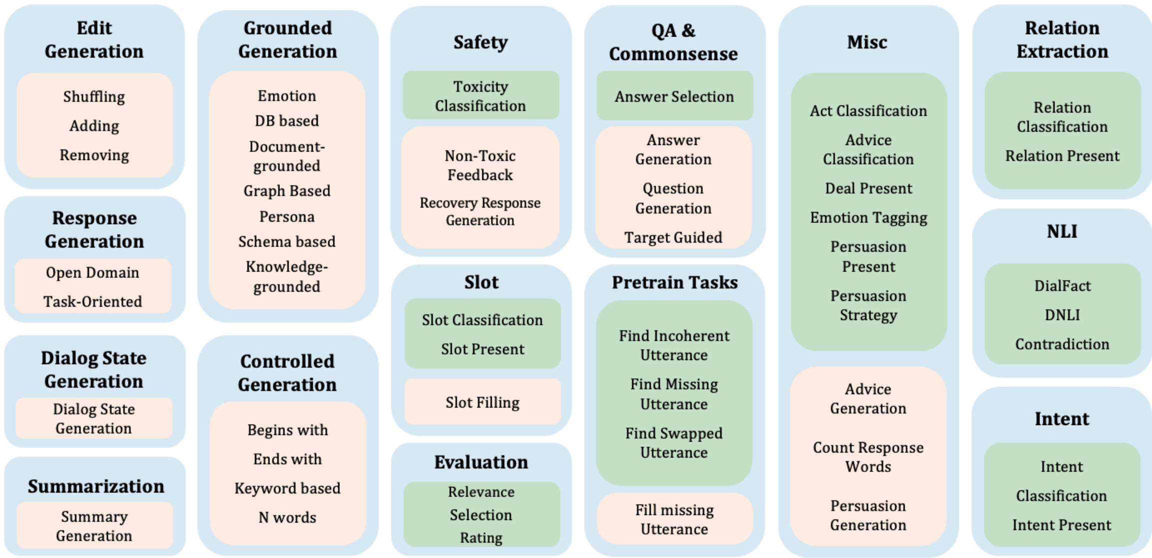

Figure 2: INSTRUCTDIAL task taxonomy. Green color represents classification tasks and orange color represents generation tasks.

2021a). However, prompts are generally short and do not generalize well to reformulations and new tasks. Instruction tuning is a recent paradigm where models are trained on a variety of tasks with natural language instructions. Going beyond multi-task training, these approaches show better generalization to unseen tasks when prompted with a few examples (Bragg et al., 2021; Min et al., 2022a,b) or language definitions and constraints (Weller et al., 2020; Zhong et al., 2021b; Xu et al., 2022). Instructions serve as a natural interface for multitask learning to support generalization to unseen tasks. PromptSource (Sanh et al., 2022), FLAN (Wei et al., 2022) and NATURAL INSTRUCTIONS (Mishra et al., 2021; Wang et al., 2022b) collected instructions and datasets for a variety of general NLP tasks. GPT3-Instruct model (Ouyang et al., 2022) is tuned on a dataset of rankings of model outputs and was trained using reinforcement learning from human feedback, but it is not publicly available and expensive to train and test. Our work instead is tailored to dialogue tasks and incorporates numerous dialogue datasets, tasks, and benchmarks. We show that these approaches are complementary to instruction tuning on dialogue and boost its performance when used as base models. Madotto et al. (2021) explored prompt-based few-shot learning in dialogue tasks, but without any fine-tuning. Mi et al. (2021a) designed task-specific instructions for TOD tasks and demonstrated good few-shot performance on several tasks. Our work covers a far greater variety of dialogue domains and datasets in comparison.

3 Methodology

In this section, we first discuss instruction tuning setup. We then discuss dialogue tasks, including our taxonomy of dialogue tasks, the schema used for representing the task meta-information, and discuss how dialogue datasets and tasks are mapped into our schema. Finally, we discuss model training and fine-tuning details.

3.1 Instruction Tuning Background

A typical supervision setup for a dialogue task t consists of a set of training instances $d^t_{train} \ni (x_i, y_i)$, where x_i is an input and y_i is a gold output. A model $\mathcal{M}$ is then trained on d^t_{train} and tested on d^t_{test} for the task. In a cross task setup, the model M is instead tested on test instances $d^{\hat{t}}_{test}$ of an unseen task $\hat{t}$. In instruction tuning, the model M is supplied with additional signal or meta information about the task at both training and inference time. The meta information can consist of prompts, task definitions, constraints, and examples. This meta information guides the model M towards the expected output space of the unseen task.

3.2 Task Collection

We adopt the definition of a task from Sanh et al. (2022), where a task is defined as "a general NLP ability that is tested by a group of specific datasets". In INSTRUCTDIAL tasks are created from existing

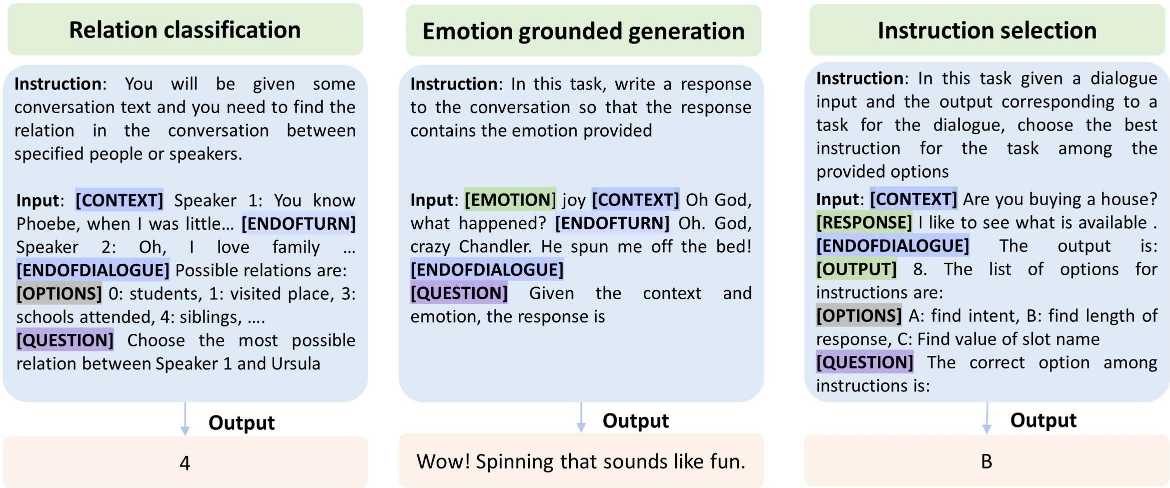

Figure 3: Instruction based input-output samples for three tasks. Each task is formatted as a natural language sequence. Each input contains an instruction, input instance formatted in a textual sequence, optional task-dependent unique inputs (such as emotion tag in emotion grounded generation task and class options in relation classification), and task-specific prompts. The instructions and the input instances are formatted using special tokens such as [CONTEXT] and [QUESTION]. The Instruction selection task is a meta-task as described in Section 3.4

open-access dialogue datasets and follow a specific instruction-based format. As a result, each dataset is incorporated into one or more dialogue tasks. Figure 2 shows the taxonomy of dialogue tasks in INSTRUCTDIAL. The tasks in the taxonomy are grouped together based on task similarity. In Table 9 we provide the list of tasks and datasets used in each task. In our taxonomy, *Classification tasks* consist of NLI, relation extraction, question answering, and intent. *Generation tasks* consist of open-domain and task-oriented response generation, controlled and grounded response generation, slot value generation, dialogue state generation, and summarization. *Evaluation tasks* consist of response selection in addition to relevance and rating prediction tasks. *Edit tasks* are a subset of tasks that involve changing a corrupted dialogue response into a coherent response. Corrupted versions of responses are created through operations such as shuffling, repeating, adding, or removing phrases and sentences in the gold response. *Pre-training tasks* involve filling missing an utterance or finding the index of an incoherent or missing utterance and are unsupervised pretraining tasks for learning dialogue context representations. They include multiple tasks covered in prior pretraining work (Mehri et al., 2019; Zhao et al., 2020b; Whang et al., 2021; Xu et al., 2021b). *Safety Tasks* consist of toxicity detection, non-toxic, and recovery response generation. Finally, *Miscellaneous tasks* are a set of tasks that belong to specialized domains such as giving advice or persuading a user.

3.3 Task Schema and Formatting

All tasks in INSTRUCTDIAL are expressed in a natural language sequence-to-sequence format and consist of an input and output. Each dataset is first preprocessed into a unified format so that the instances across tasks contain common properties. These properties include a dialogue history, a final response, and task-specific annotations such as intent and slots. These annotations and properties can then be leveraged across various task formulations.

The task schema consists of the following:

- *Task Definition*: Description of the task containing information about how to produce an output given an input.
- *Instance Inputs*: Instances from a dataset converted into a sequence.
- *Constraints*: Additional metadata or constraints for a task (e.g., emotion tag for emotion-based generation, classes for classification).
- *Prompt*: Text sequence that connects the instance back to the instruction, expressed as a command or a question.
- *Output*: Output of an instance converted into a sequence.

Figure 3 shows examples of how some sample tasks are formatted using the task schema. For each task, we manually compose 3-10 task definitions and prompts. For every instance, a task definition and a prompt are selected randomly. We do not include examples in the task schema since dialogue contexts are often long and concatenating such long inputs as examples would exceed the maximum allowable input length for most models.

Finally, input instances in a dialogue are formatted using special tokens. The token `[CONTEXT]` signals the start of dialogue content. Dialogue turns are separated by `[ENDOFTURN]` and the end of the dialogue is marked with `[ENDOFDIALOGUE]`. The token `[QUESTION]` marks the start of the prompt text.

Classification Options: In classification tasks, the model is trained to predict an output that belongs to one of several classes. To make the model aware of output classes available for an unseen task, we append a list of classes the model should pick from. We adopt the following two formats for representing the classes: (1) *Name list*: list the class names separated by a class separator token such as a comma, and choose one of the classes from the list as the output, and (2) *Indexed list*: list the classes indexed by either alphabets or numbers (such as 1: class A, 2: class B,...). The model outputs the index corresponding to the predicted class. This representation is useful when the classification options are long in length, such as in the case of response ranking where the model has to output the best response among the provided candidates.

Custom inputs: Some tasks consist of input fields that are unique to the task. For example, emotion grounded generation consists of emotion labels that the model should use for response generation. We append such inputs to the beginning of the instance sequence along with the field label to signal to the model that it should pay attention to the provided input. For example, we prepend "[EMOTION] happy" to the dialogue context in the emotion generation task.

3.4 Meta tasks

A model trained using instructions can learn to perform well on tasks during training by inferring the domain and uniqueness of the dataset instead of paying attention to the instructions. This would lead to poor test-time performance since the test instructions and the data would be unseen to the model. To handle this issue, we (1) train the models with a variety of outputs given the same input data and context through multiple task formulations, and (2) we introduce two meta-tasks: instruction selection and instruction binary, which helps the model learn the association between the instruction, the data, and the task.

In *the Instruction selection task*, the model is asked to select the instruction which corresponds to a given input-output pair. In the *Instruction binary task*, the model is asked to predict "yes" or "no" if the provided instruction can lead to a given output from an input. An example of an instruction selection task is shown in Figure 3.

3.5 None-of-the-above Options

For classification tasks, most tasks assume that the ground truth is always present in the candidate set, which is not a necessary condition for all unseen tasks. We propose adding a NOTA (None-of-the-above) option in the classification tasks during training both as correct answers and as distractors following Feng et al. (2020b). Adding NOTA increases the difficulty of the task since the model has to infer that none of the provided class options are valid outputs, which in turn should improve the model's representation capabilities. The probability p_n to add NOTA in an input instance is a hyperparameter which we set to 0.05 in our experiments. For 50% of the response sets, we replace the ground truth class with the "none of the above" option and for the remaining 50% we replace a randomly chosen class option.

4 Experimental Setup

4.1 Model Details

Our models use an encoder-decoder architecture where input is fed to the encoder and the target output is decoded by the decoder. The models are trained using maximum likelihood training objective. The decoder generates the target token-by-token in an auto-regressive manner. We finetune the following two base models on the tasks from INSTRUCTDIAL:

1. T0-3B (Sanh et al., 2022) a model with 3 billion parameters and initialized from the 3B parameters version of T5 (Lester et al., 2021). T0-3B is trained on a multitask mixture of general non-dialogue tasks such as question answering, sentiment detection, and paraphrase identification.
2. BART0 (Lin et al., 2022), a model with 406 million parameters (8x smaller than T0-3B) based on Bart-large (Lewis et al., 2020), trained on the same task mixture as T0-3B.

We name the BART0 model tuned on INSTRUCTDIAL as **ID-BART0** and T0-3B model tuned on INSTRUCTDIAL as **ID-T03B**. We treat ID-BART0 as our main model for experiments since a 3 Billion parameter model is very large and impractical to use on popular affordable GPUs. On the other hand, ID-BART0 is a more parameter-efficient model and has shown comparable zero-shot performance to T0 (Lin et al., 2022) despite being 8 times smaller. We perform finetuning on these two models since they both are instruction-tuned on general NLP tasks and thus provide a good base for building a instruction tuned model for dialogue.

4.2 Training Details

Data Sampling For training data creation, we first generate instances from all datasets belonging to each task. Since the number of instances per task can be highly imbalanced, we sample a fixed maximum of N number of instances per task. In our main models and experiments, we set $N = 5000$. Each instance in a task is assigned a random task definition and prompt. We truncate the input sequences to 1024 tokens and target output sequences to 256 tokens.

Implementation Details Our models are trained for 3 epochs with a learning rate of 5e-5 with an Adam optimizer (Kingma and Ba, 2015) with linear learning rate decay. We perform checkpoint selection using a validation set created from the train tasks. We use the HuggingFace Transformers library[2] for training and inference implementation and use Deepspeed library[3] for improving training efficiency. We train ID-BART0 on 2 Nvidia 2080Ti GPUs using a batch size of 2 per GPU and an effective batch size of 72 with gradient checkpointing. We train ID-T03B on 2 Nvidia A6000 GPUs using a batch size of 1 per GPU and an effective batch size of 72 with gradient checkpointing. For all classification tasks, we perform greedy decoding, and for all generation tasks, we perform top-p sampling with $p = 0.7$ and temperature set to 0.7. The repetition penalty is set to 1.2.

5 Experiments and Results

We evaluate our models on a variety of zero-shot and few-shot settings. We first establish benchmark results on the main zero-shot tasks set and compare our models with existing models and baselines. We then establish a zero-shot benchmark for the automatic response evaluation task. Next, we perform zero-shot and few-shot experiments on three important dialogue tasks: intent detection, slot value generation, and dialogue state tracking. Finally we discuss limitations and future work for instruction tuning in dialogue.

5.1 Zero-shot Evaluation on Unseen Tasks

In this experiment, we test our models' zero-shot ability on tasks not seen during training. We present the tasks used in the evaluation, establish baselines and then discuss and compare the performance of the baseline and ablated versions of our models.

5.1.1 Tasks for Zero-shot Setting

In our main zero-shot experiments, we perform evaluation on the set of the following 6 tasks not seen during training:

1. *Dialfact classification*: based on the Dialfact dataset (Gupta et al., 2021) where the model must decide if an evidence supports, refutes, or does not have enough information to validate the response.
2. *Relation classification*: predict the relation between two persons in a dialogue.
3. *Answer selection*: predict an answer to a conversational question.
4. *Eval selection*: choose the most relevant response among the provided 4 options. Eval tasks are based on dataset and ratings released in the DSTC 10 Automatic evaluation challenge (Chen et al., 2021b).
5. *Knowledge grounded generation*: generate a response based on provided background knowledge.
6. *Begins with generation*: generate a response that starts with the provided initial phrase.

All 6 tasks are of varying levels of difficulty and were selected to cover both classification and generation settings. To emulate a zero-shot scenario, we remove all relation-based, evaluation type, answer generation, and wiki-based tasks from the training task set. Hence, for all tasks except for begins-with generation and eval selection, the training set does not contain any dataset in the test tasks. The set of tasks used for training is presented in Table 9. We evaluate the full test sets for Dialfact, relation, and answer classification (11808, 1853, and 8386 respectively), and sample 1000 instances for the rest of the tasks for evaluation efficiency.

[2] https://github.com/huggingface/transformers
[3] https://github.com/microsoft/DeepSpeed

Model	ES	AS	RC	DC	BW				KG			
	ACC	ACC	ACC	ACC	ACC	BLEU2	ROUGEL	GRADE	F1	BLEU2	ROUGEL	GRADE
BART0	22.2	58.5	6.3	33.7	4.2	4.9	12.0	45.7	17.4	5.3	13.3	23.9
T0-3B	45.9	60.2	1.3	33.1	14.1	4.1	10.7	55.5	14.2	3.2	10.7	78.0
GPT-3	57.5	56.5	11.5	37.3	16.5	7.2	15.7	57.0	18.5	3.9	11.6	83.8
ID-BART0	66.7	59.5	17.8	35.6	56.3	13.1	26.4	60.2	27.8	11.1	21.4	68.5
ID-T03B	74.4	65.2	6.4	34.5	55.0	12.4	26.5	61.3	22.2	7.2	16.5	69.8
IDB-Few	77.1	69.1	28.0	43.0	72.2	16.7	30.7	60.3	27.9	9.7	20.0	68.0
IDB-Full	90.7	83.3	62.7	77.4	83.7	20.8	33.8	61.0	30.9	11.6	22.8	70.5
IDB-no-base	40.1	52.7	17.1	35.1	53.9	12.0	26.6	57.8	29.8	12.0	22.8	69.6
IDB-noinstr	15.8	32.3	13.1	34.8	59.3	11.8	27.1	61.1	28.6	9.8	20.1	66.3
IDB-no-instr	23.0	43.2	15.1	35.4	50.0	13.0	27.0	61.1	30.1	11.2	20.8	65.7
IDB-no-nota	66.5	57.2	17.2	35.9	56.1	10.9	25.3	58.4	28.0	11.0	21.4	67.6
IDB-no-meta	44.5	52.0	14.1	35.4	52.5	14.1	28.1	61.3	29.6	11.8	22.1	70.5

Table 1: Zero-shot evaluation on unseen tasks. Here ES stands for Eval selection, AS - Answer selection, RC - Relation Classification, DC - Dialfact classification, BW - Begins with and KG - Knowledge grounded generation. Our models ID-BART0 and ID-T03B outperform the baseline models and their ablated versions.

5.1.2 Setup and baselines

We perform inference and evaluation on the 6 unseen tasks described in Section 5.1.1. We compare the following models and baselines:

- *BART0* and *T0-3B* - Models trained on a mixture of non-dialogue general NLP tasks (described in Section 4.1).
- GPT-3 (Brown et al., 2020) - GPT-3 model tested on our test set and instructions
- ID-BART0 and ID-T03B- Our models described in Section 4.1.
- IDB-Few - Few-shot version of ID-BART0 where 100 random instances of the test tasks are mixed with the instances of the train tasks.
- IDB-Full - Version of ID-BART0 where instances of the test tasks are mixed with the instances of train tasks. The number of instances of both train and test tasks is set to $N = 5000$ per task. This baseline serves as the upper bound for our models' performance.

We compare our models with the following ablations:

- IDB-no-base - Version of ID-BART0 that uses Bart-large for instruction tuning on INSTRUCTDIAL instead of using the BART0 as the base model.
- IDB-no-instr - Version of ID-BART0 trained with no definitions and prompts. Task constraints and class options are still specified. We specify the task name instead of instructions to help the model identify the task.
- IDB-no-nota - Version of ID-BART0 trained without None-of-the-above from Section 3.5
- IDB-no-meta - Version of ID-BART0 trained without the meta tasks from Section 3.4

5.1.3 Results and Discussion

We present the results for zero-shot experiments in Table 1. We report the accuracy metric for the Eval selection, Answer selection, Relation classification, and Dialfact classification tasks. For Begins with task, we report Bleu2, RougeL, and accuracy which is defined as the proportion of responses that begins with the initial phrase provided in the instruction. For Knowledge grounded generation we report Bleu2, and RougeL metrics along with F1 as defined in (Dinan et al., 2019c). For the generation tasks we also report the automatic metric GRADE (Huang et al., 2020) that has shown good correlation with human ratings on response coherence. For GPT-3 baseline we report the metrics on 200 randomly sampled instances per task. We report the average score obtained across the instructions and prompts. We notice the following general trends in our results.

Instruction tuning on INSTRUCTDIAL improves performance on unseen dialogue tasks The ID-BART0 and ID-T03B models that are instruction-tuned on INSTRUCTDIAL achieve better performance on all tasks compared to their corresponding base models BART0 and T0-3B. Notably, for the Eval selection, Relation classification and Begins with controlled generation tasks, our models perform about 2 to 4 times better than the base models BART0 and T0-3B. Our model also performs significantly better than GPT-3 for all tasks except for Dialfact classification. In the case of the Answer selection task, the difference in performance is lower between our models and the baseline models compared to the other tasks since the baseline models are also trained on simi-

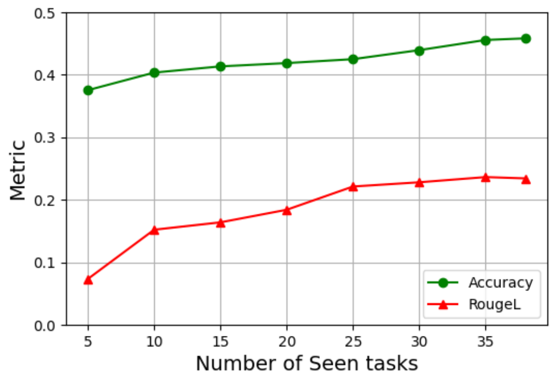

Figure 4: Model's performance improves with the number of seen tasks during training. We report average Accuracy across Eval Selection, Answer Selection, Relation Classification, and Dialfact Classification, and average RougeL scores for Knowledge Grounded Generation and Begins with Generation.

lar extractive and multi-choice question answering tasks. Relation classification and Dialfact classification are hard tasks for all models since there are no similar tasks in the set of tasks used for model training.

Effect of model size Results indicate the larger models are not necessarily better across tasks. While T0-3B and ID-T03B perform better on the Eval selection and Answer Selection tasks and perform equivalently on the Begins with generation task, BART0 and ID-BART0 perform better on the rest of the unseen tasks.

Few-shot training significantly improves performance IDB-Few model that incorporates 100 instances per test task in its training data shows significant improvements in performance compared to its zero-shot counterpart ID-BART0. We see about 12 to 16 percent improvements on the Eval selection, Answer selection, and Dialfact classification tasks, and 30 to 50 percent improvement on the Begins with and Relation classification tasks. These results indicate that multi-task instruction tuning models are good at few-shot generalization to new tasks.

Full-shot training can improve performance across multiple tasks IDB-Full model that incorporates 5000 instances per test task in its training data achieves very high performance across all test tasks. The full-shot performance of ID-BART0 on Dialfact and Relation classification tasks are near state of the art performance on those tasks, even without leveraging the full train datasets.

Pretraining on general NLP tasks helps dialogue instruction tuning IDB-no-base model that uses BART-large as the base model shows a notable drop in performance on Eval selection and Answer selection tasks, while a slight drop in performance across all the other test tasks. Therefore, we conclude that a model pretrained with instruction tuning on general NLP tasks can benefit instruction tuning for dialogue.

Using instructions leads to better generalization IDB-no-instr, the ablated version of ID-BART0 trained without instructions shows substantially worse performance than ID-BART0 across all tasks, especially on Eval selection, Answer selection, and Relation classification tasks. This indicates that training with instructions is crucial for zero-shot performance on unseen tasks.

None-of-the-above and meta tasks are important for improving performance IDB-no-nota, the ablated version of ID-BART0 that does not use the NOTA objective shows a slight drop in performance in the classification task, indicating that though NOTA objective is helpful, it is not crucial for instruction tuning. However, we see a substantial drop in performance on the unseen classification tasks (except for Knowledge grounded generation task) when we removed the meta tasks described in Section 3.4. This shows that the proposed meta tasks helps the model develop better representations and understanding of the natural language instructions.

Training on more seen tasks improves generalization on unseen tasks In Figure 4 we show the impact of varying the number of seen tasks on the performance on unseen tasks. We adopt the train-test task split from section 5.1. We observe that the performance improves sharply up to 20-25 tasks and then further keeps steadily increasing with each new task. This indicates that increasing the number of tasks can lead to better zero-shot generalization and that scaling to more tasks may lead to better instruction-tuned models.

Sensitivity to instruction wording To analyze the sensitivity of our models to the wording of the instructions, we breakdown the evaluation metrics per unique instruction used during inference for the ID-BART0 model. For Eval selection, the accuracy varies from 65.6 to 67.8 across instructions, it ranges from 52.5 to 75.0 for Answer selection, 17.1 to 18.4 for Relation classification, 34.7 to 37.1 for Dialfact classification, 49.8 to 62.3 for Begins with generation, and the F1 score varies from 26.6 to 28.6 for Knowledge grounded generation. The

Model	DSTC6	DSTC7	HUMOD	TU	PZ	DZ	CG	PU	DGU	DGR	FT	EG	FD	Average
MAUDE (2020)	0.115	0.045	0.112	0.136	0.360	0.120	0.304	0.306	0.192	-0.073	-0.11	-0.057	-0.285	0.085
GRADE (2020)	0.121	0.332	0.612	0.176	0.583	0.532	0.571	0.329	0.596	0.254	0.048	0.300	0.106	0.329
USR (2020b)	0.166	0.249	0.34	0.291	0.496	0.363	0.487	0.140	0.353	0.066	0.055	0.268	0.084	0.242
FED (2020a)	-0.082	-0.070	-0.077	-0.090	-0.232	-0.080	-0.137	-0.004	0.025	-0.009	0.173	0.005	0.178	-0.030
FlowScore (2021)	0.095	0.067	-0.049	0.068	0.202	-0.063	-	0.053	0.053	-	-0.043	-	-0.009	0.029
USL-H (2020)	0.180	0.261	0.53	0.319	0.409	0.385	0.452	0.493	0.481	0.09	0.115	0.237	0.202	0.297
QuestEval (2021)	0.089	0.222	0.217	0.104	0.32	0.22	0.344	0.106	0.243	-0.026	0.168	0.195	0.114	0.165
DEB (2020)	0.214	0.351	0.649	0.123	0.579	0.486	0.504	0.351	0.579	0.363	0.044	0.395	0.141	0.346
DynaEval (2021)	0.252	0.066	0.112	-0.013	0.165	0.169	0.202	0.148	0.038	0.122	0.247	0.159	0.555	0.162
DialogRPT (2020)	0.162	0.255	0.198	0.118	0.114	0.067	0.158	-0.036	0.075	0.037	-0.249	0.203	-0.134	0.065
Best in DSTC10	0.616	0.313	0.225	0.455	0.764	0.545	0.570	0.479	0.789	0.644	0.352	0.501	0.774	0.522
Ours	0.542	**0.451**	0.559	0.397	**0.647**	**0.559**	*0.493*	*0.402*	**0.618**	0.249	*0.250*	0.469	*0.257*	0.421

Table 2: Spearman correlation of model predictions with human ratings. Bold, underlined and italicised scores represent the datasets on which our model performs the best, second best and third best respectively. TU, PZ, DZ, CG, PU, DGU, DGR, FT, EG and FD are abbreviations for TopicalChat-USR (Mehri and Eskenazi, 2020b), PersonaChat-Zhao (Zhao et al., 2020a), DailyDialog-Zhao (Zhao et al., 2020a), ConvAI2-GRADE (Huang et al., 2020), PersonaChat-USR (Mehri and Eskenazi, 2020b), DailyDialog-Gupta (Gupta et al., 2019), DailyDialog-GRADE (Huang et al., 2020), FED-Turn (Mehri and Eskenazi, 2020a), Empathetic-GRADE (Huang et al., 2020), and FED-Dial (Mehri and Eskenazi, 2020a).

variance is low for tasks such as Eval selection while high for Answer selection. This shows our model is moderately sensitive to the wording of the instruction wording. Expanding the set of instructions used during training is a possible solution to this issue.

Errors in model outputs We perform qualitative analysis on randomly sampled instances from the outputs of our models and the baselines. For classification tasks, a common error across all models is that they sometimes generate outputs that are out of scope of the list of classes they are asked to choose from. However, while this happens with GPT-3 for 20%, BART0 10% and T0-3B 17.8% of the inputs, for ID-BART0 and ID-T03B this occurs only for 2.5% and 4.8% of the inputs. Other possible but rare types of errors include copying the provided input as output, early truncation of generated responses, and performing an unspecified task. Apart from the unseen task set adopted for our experiments in section 5.1.1, we tried some other seen-unseen task configurations and found that our models and baselines were incapable of performing certain tasks such as Infilling missing utterance, Recovery response generation to a toxic context, and Ends-with response generation in a zero-shot manner. However the models were quickly able to learn these tasks when trained on a few instances of these tasks.

5.2 Zero-shot Automatic Response Evaluation

Development of automatic metrics that show high correlations with human judgements for dialogue evaluation is a challenging and crucial task for the development of dialogue systems. Automated metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) are widely used to evaluate the quality of model generated text. However, these metrics correlate poorly with human judgement ratings of generated text quality (Gupta et al., 2019). In this experiment, we test our model on the zero-shot automatic evaluation task through the Eval Relevance task. In this task, we use the evaluation ratings and datasets released in the DSTC 10 Automatic evaluation challenge (Chen et al., 2021b). The data consists of 65,938 context-response pairs along with corresponding human ratings aggregated from multiple prior works. We use the larger model ID-T03B for this task and present the set of tasks used for training the model in Table 9. Given a dialogue context and a candidate response, we instruct the model to predict "yes" if the response is relevant to the context, and otherwise predict "no". We calculate the probability of "yes" or that the response is relevant to the context as $p(yes) = p(yes)/(p(yes) + p(no))$. We then calculate the Spearman correlation of the model's prediction with human ratings for relevance provided in the DSTC 10 datasets. We present the correlation scores of ID-T03B model in Table 2. We compare our model with reference-free metrics studied in Yeh et al. (2021). The last row in the table is the best system (Team7) in DSTC10 competition (Chen et al., 2021b), however, no description or technical detail of the system is released yet. ID-T03B is ranked the first or second best in the ma-

Model	Accuracy
ConvERT (Casanueva et al., 2020)	83.32
ConvERT + USE (Casanueva et al., 2020)	85.19
Example-Driven (Mehri and Eric, 2021)	85.95
PPTOD$_{base}$ (Su et al., 2022b)	82.81
PPTOD$_{large}$ (Su et al., 2022b)	84.12
ID-BART0 (Ours)	84.30
BART0 (zero-shot)	14.72
ID-BART0 (Ours, zero-shot)	58.02

Table 3: Intent prediction accuracy on the BANKING77 corpus (Casanueva et al., 2020). Models in the first section of the table are trained in a few-shot setting with 10 instances per intent. Models in the second section are tested in a zero-shot setting.

jority of the evaluation datasets. Our model learns coherence from the variety of tasks it is trained on and demonstrates high zero-shot dialogue evaluation capabilities.

5.3 Zero-shot and Few-shot Dialogue Tasks

We test the zero-shot and few-shot abilities of our models on three important dialogue tasks: intent prediction, slot filling, and dialogue state tracking.

5.3.1 Intent Prediction

Intent prediction is the task of predicting an intent class for a given utterance. The large number of potential intent classes precludes zero-shot generalization, due to the limited length of the natural language instructions. As such, we conduct few-shot experiments on the Banking77 benchmark dataset (Casanueva et al., 2020) that contains 77 unique intent classes. Models are trained on 10 instances per intent class. We compare our model ID-BART0 with Convert Models (Casanueva et al., 2020) that are Bert-based dual encoder discriminative models and PPTOD (Su et al., 2022b), a model pre-trained on multiple task-oriented dialogue datasets. For this experiment, ID-BART0 is pretrained on the training task mixture from Section that includes few intent detection datasets except for Banking77 dataset. The results in Table 3 shows the performance of our model relative to prior work. Our model is able to attain competitive performance in the few-shot setting, without necessitating complex task-specific architectures or training methodology. It is notable that ID-BART0 performs better than PPTOD which uses about about two times more parameters and is trained similarly to our model using a Seq2Seq format. We also provide benchmark zero-shot results. While BART0 model struggles in the zero-shot setting, ID-BART0

Model	F1
CONVEX (HENDERSON AND VULIĆ, 2020)	5.2
COACH+TR (LIU ET AL., 2020)	10.7
GENSF (MEHRI AND ESKENAZI, 2021)	19.5
ID-BART0 (Ours)	56.4

Table 4: Zero-shot slot filling results on the Restaurant8k corpus.

Domain	GENSF	ID-BART0 (Ours)
Buses	90.5	97.8
Events	91.2	94.3
Homes	93.7	96.5
Rental Cars	86.7	94.2

Table 5: F1 scores for few-shot slot filling on the DSTC8 corpus.

shows greatly improved performance.

5.3.2 Slot Filling

Slot filling is the problem of detecting slot values in a given utterance. We carry out zero-shot experiments on the Restaurant8k corpus (Coope et al., 2020a) and few-shot experiments on the DSTC8 dataset (Rastogi et al., 2020a), demonstrating significant performance gains over prior work. The set of tasks used for training the model are presented in Table 9. For the zero-shot experiments, the training set includes several slot filling datasets except for the Restaurant8k dataset. The few-shot experiments on the DSTC8 datasets span four domains - buses, events, homes, rental cars and involves training on 25% of the training dataset.

Table 4 shows that our approach attains a 36.9 point improvement in zero-shot slot filling. This result especially highlights the efficacy of instruction tuning at leveraging large-scale pretrained language models to generalize to unseen tasks. Similarly, we see significant improvement in the few-shot setting on the DSTC8 benchmark in Table 5.

5.3.3 Dialogue State Tracking

We also evaluate our model on task-oriented dialogue. In particular, we examine dialogue state tracking. Using the same experimental setup as PPTOD (Su et al., 2022a), we conduct few-shot experiments on MultiWOZ 2.0 (Budzianowski et al., 2018). Similar to PPTOD, our ID-BART0 model is first trained on 7 datasets: KVRET (Eric et al., 2017), WOZ (Mrkšić et al., 2017), CamRest676 (Wen et al., 2017), MSR-E2E (Li et al., 2018), Frames (El Asri et al., 2017), TaskMaster (Byrne et al., 2019), Schema-Guided (Rastogi et al., 2020b). We then train on 1% and 5% splits of

Model	1%	5%
PPTOD$_{base}$	29.7	40.2
ID-BART0 (Ours)	29.2	38.1

Table 6: Joint goal accuracy for dialogue state tracking in few-shot setting on 1% and 5% data of Multiwoz.

MultiWOZ for 40 epochs with a learning rate of $5e-5$. In Table 6 we present few-shot dialogue state tracking results on the MultiWOZ test set. We find that our model obtains 29.2 and 38.1 joint goal accuracy on the 1% and 5% training data splits, respectively. Our results demonstrate that our model performs well on few-shot dialogue state tracking, and achieves competitive results against other baseline models.

6 Conclusion

We propose INSTRUCTDIAL, an instruction tuning framework for dialogue, which consists of a repository of a variety of dialogue tasks unified into text-to-text format and created from openly available dialogue datasets. We also propose two meta-tasks to encourage the model to pay attention to the instructions. Our empirical results show that models trained on INSTRUCTDIAL achieve good zero-shot performance on unseen tasks, such as dialogue evaluation, as well as good few-shot performance on dialogue tasks such as intent prediction, slot filling, and dialogue state tracking. We perform ablation studies showing the impact of using an instruction tuned base model, model size and type, the effect of using instructions, increasing the number of tasks, and incorporation of our proposed meta tasks. Our experiments reveal that instruction tuning does not benefit all unseen test tasks and that there is further room for improvement in aspects such as developing invariance to instruction wording and task interference. We hope that INSTRUCTDIAL will facilitate further progress on instruction-tuning systems for dialogue tasks.

7 Limitations and Future Work

Our work is the first to explore instruction tuning for dialogue and establishes baseline performance for a variety of dialogue tasks. However, there is room for improvements in the following aspects: 1) Unlike a few prior works, the instructions and prompts used in this work are not crowdsourced and are limited in number. Furthermore, our instructions and tasks are only specified in the English language. Future work may look into either crowdsourcing or automatic creation methods for augmenting the set of instructions in terms of both language diversity as well as quantity. 2) Instruction tuning does not show significant improvements in zero-shot setting on a few tasks such as relation classification in our experiments. Future work can look into investigating why certain tasks are more challenging than others for zero-shot generalization. 3) We observed a few instances of task interference in our experiments. For example, the set of train tasks used for zero-shot automatic response evaluation experiment as mentioned in table 9 is different and smaller from the set of tasks used in our main experiments in Section 5.1.1. We found that incorporating a few additional tasks lead to a reduction in the performance on zero-shot automatic evaluation. Furthermore, training on multiple tasks can lead to task forgetting. To address these issues, future work can take inspiration from work related to negative task interference (Wang et al., 2020a; Larson and Leach, 2022), transferability (Vu et al., 2020; Wu et al., 2020b; Xing et al., 2022) and lifelong learning (Wang et al., 2020b). 4) Our models are sensitive to the wording of the instruction in zero-shot settings as discussed in Section 5.1.3. Improving insensitivity to prompts and instructions is an important research topic for future work. 5) Our work does not explore in-context few-shot learning through examples as well as the composition of multiple tasks through instructions. Both these aspects warrant further investigations.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng

Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525, Online. Association for Computational Linguistics.

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. Flex: Unifying evaluation for few-shot nlp. In *Advances in Neural Information Processing Systems (NeurIPS 2021)*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22, Hong Kong. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, 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*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021a. 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.

Zhang Chen, João Sadoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. 2021b. Automatic evaluation and moderation of open-domain dialogue systems. *arXiv preprint arXiv:2111.02110*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Sam Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. 2020a. Span-convert: Few-shot span extraction for dialog with pretrained conversational representations. *arXiv preprint arXiv:2005.08866*.

Samuel Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. 2020b. 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, Online. Association for Computational Linguistics.

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*, pages 12–16.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019a. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. 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 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019b. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019c. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.

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, Saarbrücken, Germany. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020a. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Yulan Feng, Shikib Mehri, Maxine Eskenazi, and Tiancheng Zhao. 2020b. "none of the above": Measure uncertainty in dialog response retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2013–2020, Online. Association for Computational Linguistics.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at dstc7.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *Proceedings of 7th IEEE Workshop on Spoken Language Technology*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391, Stockholm, Sweden. Association for Computational Linguistics.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Dialfact: A benchmark for fact-checking in dialogue. *arXiv preprint arXiv:2110.08222*.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online. Association for Computational Linguistics.

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*.

Matthew Henderson and Ivan Vulić. 2020. Convex: Data-efficient and few-shot slot labeling. *arXiv preprint arXiv:2010.11791*.

Chiori Hori and Takaaki Hori. 2017. End-to-end conversation modeling track in dstc6. *arXiv:1706.07440*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Stefan Larson and Kevin Leach. 2022. Redwood: Using collision detection to grow a large-scale intent classification dataset. *arXiv preprint arXiv:2204.05483*.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 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, Hong Kong, China. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

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

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. 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 128–138, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Kangmin Tan, Chris Miller, Beiwen Tian, and Xiang Ren. 2022. Unsupervised cross-task generalization via retrieval augmentation. *ArXiv*, abs/2204.07937.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Qi Liu, Lei Yu, Laura Rimell, and Phil Blunsom. 2022. Pretraining the noisy channel model for task-oriented dialogue. *Transactions of the Association for Computational Linguistics*, 9(0):657–674.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021b. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pages 165–183. Springer.

Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2020. Coach: A coarse-to-fine approach for cross-domain slot filling. *arXiv preprint arXiv:2004.11727*.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

Shikib Mehri and Mihail Eric. 2021. Example-driven intent prediction with observers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2979–2992, Online. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2021. Gensf: Simultaneous adaptation of generative pre-trained models and slot filling. *arXiv preprint arXiv:2106.07055*.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining methods for dialog context representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3836–3845, Florence, Italy. Association for Computational Linguistics.

Erinc Merdivan, Deepika Singh, Sten Hanke, Johannes Kropf, Andreas Holzinger, and Matthieu Geist. 2020. Human annotated dialogues dataset for natural conversational agents. *Applied Sciences*, 10(3):762.

Fei Mi, Yitong Li, Yasheng Wang, Xin Jiang, and Qun Liu. 2021a. Cins: Comprehensive instruction for few-shot learning in task-oriented dialog systems. *ArXiv*, abs/2109.04645.

Fei Mi, Wanhao Zhou, Lingjing Kong, Fengyu Cai, Minlie Huang, and Boi Faltings. 2021b. Self-training improves pre-training for few-shot learning in task-oriented dialog systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1887–1898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022a. MetaICL: Learning to learn in context. In *NAACL-HLT*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Natural instructions: Benchmarking generalization to new tasks from natural language instructions. *In Annual Meeting of the Association for Computational Linguistics (ACL).*

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, Florence, Italy. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. 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 1699–1713, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. In *Transactions of the Association for Computational Linguistics*.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIME-DIAL: Temporal commonsense reasoning in dialog. 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 7066–7076, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Dinesh Raghu, Shantanu Agarwal, Sachindra Joshi, and Mausam. 2021. End-to-end learning of flowchart grounded task-oriented dialogs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4348–4366, Online and Punta Cana, Dominican Republic. 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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020a. Schema-guided dialogue state tracking task at dstc8. *arXiv preprint arXiv:2002.01359*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020b. 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.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations (ICLR)*.

Timo Schick and Hinrich Schütze. 2021. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Karin Sevegnani, David M. Howcroft, Ioannis Konstas, and Verena Rieser. 2021. OTTers: One-turn topic transitions for open-domain dialogue. 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 2492–2504, Online. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022a. Multi-task pre-training for plug-and-play task-oriented dialogue system.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022b. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland. Association for Computational Linguistics.

Megan Ung, Jing Xu, and Y-Lan Boureau. 2021. Saferdialogues: Taking feedback gracefully after conversational safety failures.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022a. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, et al. 2022b. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv*.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020a. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Zirui Wang, Sanket Vaibhav Mehta, Barnabas Poczos, and Jaime Carbonell. 2020b. Efficient meta lifelong-learning with limited memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 535–548, Online. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts?

Jason Wei, Maarten Paul Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew Mingbo Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E. Peters. 2020. Learning from task descriptions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1361–1375, Online. Association for Computational Linguistics.

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

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14041–14049.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020a. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Chien-Sheng Wu, Andrea Madotto, Wenhao Liu, Pascale Fung, and Caiming Xiong. 2021. Qaconv: Question answering on informative conversations. *arXiv preprint arXiv:2105.06912*.

Sen Wu, Hongyang R Zhang, and Christopher Ré. 2020b. Understanding and improving information transfer in multi-task learning. *arXiv preprint arXiv:2005.00944*.

Yujie Xing, Jinglun Cai, Nils Barlaug, Peng Liu, and Jon Atle Gulla. 2022. Balancing multi-domain corpora learning for open-domain response generation. *arXiv preprint arXiv:2205.02570*.

Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. 2022. Zero-prompt: Scaling prompt-based pretraining to 1,000 tasks improves zero-shot generalization. *arXiv preprint arXiv:2201.06910*.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021a. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online. Association for Computational Linguistics.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021b. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14158–14166.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14230–14238.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940, Online. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Elizabeth Clark, Lianhui Qin, Ali Farhadi, and Yejin Choi. 2021. TuringAdvice: A generative and dynamic evaluation of language use. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4856–4880, Online. Association for Computational Linguistics.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. 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 5676–5689, Online. Association for Computational Linguistics.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020a. Designing precise and robust dialogue response evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Online. Association for Computational Linguistics.

Yufan Zhao, Can Xu, and Wei Wu. 2020b. Learning a simple and effective model for multi-turn response generation with auxiliary tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3472–3483, Online. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021a. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021b. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2856–2878, Punta Cana, Dominican Republic. Association for Computational Linguistics.

8 Appendix

In Table 7 we present the list of tasks with sample inputs for each task. In Table 8 we present the list of Tasks with datasets used in each task. In Table 9 we provide the configurations of experiments, that is, the tasks used for training for each experiment.

Task Type	Task Name	Example
Intent	Intent Classification	[RES] Make a reservation for 4 [EOD]. The possible intents are: [OPT] BookRestaurant | ShareETA [Q] The best option is
	Intent Present	[RES] list the three earliest flights [EOD]. The possible options are: [OPT] yes | no [Q]. Is the intent flight correct?. "
NLI	DialFact	[CTX] What year did they start making pasta? [RES] I think pasta was first made somewhere in Europe many centuries ago. [EOD] The possible classes are: [OPT] refutes | supports | not enough info [Q]. Choose the most possible class
	DNLI	[CTX] i am named after a cartoon fox . [RES] i have a dog . [EOD]. The possible classes are: positive | negative | neutral [Q]. The predicted class is
	Contradiction	[CTX] lol are they fast drying ? [EOT] Kind of slow lol. [RES] I know they dry fast. [EOD]. The possible classes are: uncontradicted | contradicted [Q]. What is the class given the context and the response
Safety Classification	Toxicity Classification	[CTX] Hello [EOT] hello ... [EOD] [REP] not interesting [Q] Is the response toxic? Answer choices [OPT] yes | no
Relation Extraction	Relation Classification	[CTX]It's like this, me, no jokes. [EOT] All right ... [EOD]. The possible relations are: [OPT] per:siblings ... [Q]. The relation between A and B is
	Relation Present	[CTX] Hello, Mark? ... [EOT] That is so made up! [EOD] . Does the relation per:alternate exist between A and B? Answer [OPT] yes | no.
Evaluation	Relevance	[CTX] to holden my dad ... [EOD] [REP] you can send us your email address. [SEP] Is the response contextual? Answer [OPT] yes | no.
	Selection	[CTX] this is sprint great service URL [EOD] The best response is [OPT] you can send us | please ...
	Rating	[CTX] this is sprint great service URL [EOD] Please give a rating ranging from 1 to 5 to the following response: please dm us your account
Slot	Slot Classifcation	[RES] what do you have tomorrow after 5 o'clock from atlanta to san francisco [EOD] [Q] What is the value of slot: city_name in the response
	Slot Present	[RES] Yes. That sounds great. Can I scheduled ... [EOD]. The possible options are: [OPT] yes | no [Q]. The slot visit date is present in the response?
	Slot Value Generation	[CTX] I need tickets to [EOT] Great! [RES] You've got 2 tickets [EOD] [Q]. What is the value of slot: starttime in the response
Safety Generation	Non-Toxic Feedback	[CTX] I have never met [EOT] another group is ... [EOD] [Q] Given the conversation, a non toxic response is
	Recovery Resp. Generation	[CTX] I have never met [EOT] another group is ... [EOD] [Q] Given the conversation, a non toxic recovery response is
Grounded Generation	Emotion	[EMO] anger [CTX] I won! [EOD] [Q] Given the context and emotion, the response is
	DB based	[STATE] hotelparking: yes [DB] Type: guest house [CTX] there are ... [EOD] [Q] Given the context, db, and state, the response is
	Document-grounded	[WIKI] you must report .. [CLASS] That is the case ... [CTX] Hello ... [EOD] [Q] Given the context and doc, the response is
	Graph Based	[GRAPH] the subject is, relation: [CTX] do you like iron man [EOD] [Q] Given the context and triplets, the response is
	Persona	[P] i'm 60 years old ... [CTX] Hello! How is your ... [EOD] Given the context and persona, the response is
	Schema Based	[SCHEMA] terminal: false, label: open circuit [CTX] My car is ... [EOD] [Q] Given this context and schema, the response is
	Knowledge-Grounded	[DOC] demetri martin was accepted into harvard law , but left out of boredom to pursue a career in comedy [CTX] do you know who demetri martin is ? [EOD] Given this context and knowledge, the response is
QA and Commensense	Answer Generation	[DOC] Jessica went to sit in her rocking chair ... [Q] Who had a Birthday? Jessica. How old would she be?
	Answer Selection	[DOC] Jessica went to sit in her r ... [OPT] 80 | park ... [Q] Who had a Birthday? Jessica. How old would she be?
	Question Generation	[DOC] Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80 [Q] what should we ask about this conversation
	Target Guided	[Target] i love chocolate. [CTX] i love walking in the park. [Q] Generate a text which connects the context with the target sentence."
Controlled Generation	Begins With	[INIT] I tell ya [CTX] can I ask you something? ... [EOD] [Q] Given this context generate a response which starts with the given initial sentence:
	Ends with	[FINAL] checks ? [CTX] Are you through with your meal ... [EOD] [Q] Given this context and final phrase, the response is
	Keyword Based	[KEY] lot of memory, desktop computer and memory [CTX] Can I help you ... [EOD] [Q] Here is a response which contains the given keywords
	N Words	[CTX] Do you know Manchester United F.C ... [EOD] [Q] Given this context, the response with 3 number of words is
Dialog State Generation	Dialog State Generation	[CTX] I need help finding an apartment [EOT] what area are you hoping ... [EOD] [Q] What is the belief state?
Edit Generation	Shuffling	
	Adding	[RES] hi, you have to report [CTX] Many DMV ... [EOD] [Q] Given this context and response provided, the edited response is
	Removing	
Pretrain Tasks	Fill Missing Utterance	[CTX] Do you know Manchester United F.C.? ... [EOD] [Q] Given this context generate a response coherent to the context
	Find Incoherence Utterance	[CTX] Do you know Manchester United F.C.? ... [EOD] [Q] Given this context generate a response coherent to the context
	Find Missing Utterance	[CTX] Do you know Manchester United F.C.? [EOT] [MASK] ... [EOD] [Q] Here is the missing utterance that can take place of [MASK]
	Find Swapped Utterance	[CTX] Do you know Manchester United F.C.? [EOD] [Q] Given this context the swapped indices of responses are
Response Generation	Open Domain Task-oriented	[CTX] Do you know Manchester United F.C.? ... [EOD] [Q] Given this context generate a response coherent to the context"
Summarization	Summary Generation	[CTX] Person2 OK. [EOT] Person1: Well, how old are you? ... [EOD] [Q] Given this dialog context, its summary is the following:
Misc	Act Classification	[CTX] Hi, I am looking for a nice German restaurant [EOD] The possible acts are: [OPT] request | inform [Q] The dialog act is
	Advice Present	[CTX] Anyone take mental ... [EOD] [RES] Back at my old job ... [Q] Does the response provide advice for the issue? Choices [OPT] yes | no
	Advice Generation	[CTX] Anyone take mental health days from work? ... [EOD] [Q] The response is
	Deal Present	[CTX] I like the basketball and the hat ... [EOT] deal [EOD] [Q] Was an agreement reached? Choices [OPT] yes | no
	Emotion Tagging	[CTX] Hey, so did you have fun with Joey ... [EOD] The possible emotions are [OPT] disgust ... [Q] The emotions in the dialog are
	Persuasion Present	[CTX] Hello How are you ...[EOD] [RES] Are you involved with charities [Q] Is task-related-inquiry used in the response? Choices [OPT] yes | no
	Persuasion Strategy	[CTX] how can i help [EOD] The possible strategies are: [OPT] request | inform [Q] The strategy is
	Persuasion Generation	[STRATEGY] proposition-of-donation [CTX] how can i help? [EOD] [Q] The response is
	Count Response Words	[CTX] Do you know Manchester United F.C.? ... [EOD] [Q] Given this context Here is length of the response in the context"

Table 7: List of tasks with sample inputs for each task. The left column describes the general task type. The middle column lists the specific task. The right column displays an example formatted using a randomly selected task definition and prompt for the task. [CTx] is short for [CONTEXT], [Q] is short for [QUESTION], [EOT] is short for [ENDOFTURN] and [EOD] is short for [ENDOFDIALOGUE]

Task Type	Task Name	Datasets
Intent	Intent Classification	ATIS (Hemphill et al., 1990) SNIPS (Coucke et al., 2018) CLINIC150 (Larson et al., 2019)
	Intent Present	HWU64 (Liu et al., 2021b) Banking77 (Casanueva et al., 2020)
NLI	DialFact	DialFact (Gupta et al., 2021)
	DNLI	Decode (Nie et al., 2021) Dialogue NLI (Welleck et al., 2019)
	Contradiction	
Safety Classification	Toxicity Classification	ToxiChat (Baheti et al., 2021) BAD (Xu et al., 2021a) Build it Break it Fix it (Dinan et al., 2019a)
Relation Extraction	Relation Classification	DialogRE (Yu et al., 2020)
	Relation Present	
Evaluation	Relevance	DSTC6 (Hori and Hori, 2017) DSTC7 (Galley et al., 2019) Persona-Chatlog (See et al., 2019)
	Selection	USR (Mehri and Eskenazi, 2020b) FED (Mehri and Eskenazi, 2020a) DailyDialog (?Zhao et al., 2020a)
	Rating	PersonaChat (Zhao et al., 2020a) GRADE (Huang et al., 2020) HUMOD (Merdivan et al., 2020)
Slot	Slot Classifcation	RESTAURANTS-8K (Coope et al., 2020b) DSTC8-SGD (Rastogi et al., 2020b)
	Slot Present	ATIS (Hemphill et al., 1990) SNIPS (Coucke et al., 2018)
	Slot Value Generation	TaskMaster (Byrne et al., 2019) MSRE2E (Li et al., 2018)
Safety Generation	Non-Toxic Feedback	SaFeRDialogues (Ung et al., 2021)
	Recovery Response Generation	
Grounded Generation	Emotion	EmpatheticDialogues (Rashkin et al., 2019) GoEmotions (Demszky et al., 2020) EmotionLines (Hsu et al., 2018)
	DB based	MultiWOZ (Budzianowski et al., 2018)
	Document-grounded	doc2dial (Feng et al., 2020a)
	Graph Based	OpenDialKG (Moon et al., 2019)
	Persona	ConvAI (Dinan et al., 2019b) PersonaChat (Zhang et al., 2018)
	Schema Based	FloDial (Raghu et al., 2021)
	Knowledge-Grounded	TopicalChat (Gopalakrishnan et al., 2019) WoW (Dinan et al., 2019c)
QA and Commensense	Answer Generation	CIDEr (Vedantam et al., 2015) TIMEDIAL (Qin et al., 2021) MuTual (Cui et al., 2020)
	Answer Selection	QAConv (Wu et al., 2021) CoQA (Reddy et al., 2019) QuAC (Choi et al., 2018)
	Question Generation	QAConv (Wu et al., 2021)
	Target Guided	OTTers (Sevegnani et al., 2021)
Controlled Generation	Begins With	EmpatheticDialogues (Rashkin et al., 2019) DailyDialog (Li et al., 2017) ConvAI (Dinan et al., 2019b)
	Ends with	
	Keyword Based	TuringAdvice (Zellers et al., 2021) EmotionLines (Hsu et al., 2018) WoW (Dinan et al., 2019c)
	N Words	DailyDialog (Li et al., 2017) WoW (Dinan et al., 2019c) EmpatheticDialogues (Rashkin et al., 2019)
Dialog State Generation	Dialog State Generation	MultiWOZ (Budzianowski et al., 2018) KVRET (Eric et al., 2017) WOZ (Mrkšić et al., 2017) CamRest676 (Wen et al., 2017) MSR-E2E (Li et al., 2018) Frames (El Asri et al., 2017) TaskMaster (Byrne et al., 2019) Schema-Guided (Rastogi et al., 2020b)
Edit Generation	Shuffling	TopicalChat (Gopalakrishnan et al., 2019) EmotionLines (Hsu et al., 2018) EmpatheticDialogues (Rashkin et al., 2019)
	Adding	WoW (Dinan et al., 2019c) Persuation (Wang et al., 2019) CaSiNo (Chawla et al., 2021) DialogSum (Chen et al., 2021a)
	Removing	DailyDialog (Li et al., 2017) ConvAI (Dinan et al., 2019b) EmotionLines (Hsu et al., 2018)
Pretrain Tasks	Fill Missing Utterance	DailyDialog (Li et al., 2017) WoW (Dinan et al., 2019c) EmpatheticDialogues (Rashkin et al., 2019) OpenDialKG (Moon et al., 2019)
	Find Incoherence Utterance	
	Find Missing Utterance	
	Find Swapped Utterance	
Response Generation	Open Domain	DailyDialog (Li et al., 2017) ConvAI (Dinan et al., 2019b) WoW (Dinan et al., 2019c) EmpatheticDialogues (Rashkin et al., 2019) OpenDialKG (Moon et al., 2019)
	Task-oriented	MultiWOZ (Budzianowski et al., 2018)
Summarization	Summary Generation	DialSum (Goo and Chen, 2018) QMSum (Zhong et al., 2021a) SAMSum (Gliwa et al., 2019)
Misc	Act Classification	MSRE2E (Li et al., 2018) DailyDialog (Li et al., 2017) MultiWOZ (Budzianowski et al., 2018)
	Advice Present	TuringAdvice (Zellers et al., 2021)
	Advice Generation	
	Deal Present	Deal (Lewis et al., 2017)
	Emotion Tagging	GoEmotions (Demszky et al., 2020) EmotionLines (Hsu et al., 2018) DailyDialog (Li et al., 2017)
	Persuasion Present	Persuation (Wang et al., 2019) CaSiNo (Chawla et al., 2021)
	Persuasion Strategy	
	Persuasion Generation	
	Count Response Words	DailyDialog (Li et al., 2017) WoW (Dinan et al., 2019c) EmpatheticDialogues (Rashkin et al., 2019)

Table 8: List of Tasks with datasets used in each task. The left column describes the general task type. The middle column lists the specific task. The right column shows all datasets used for a specific task type.

Experiment	Base model(s)	Tasks
Main zero-shot tasks	ID-BART0, ID-T0	act classification act generation advice generation advice present answer generation count response words db based generation deal present document grounded generation edit generation emotion generation emotion tagging endswith controlled generation fill missing utterance find incoherent utterance find missing utterance graph based generation intent classification intent present (no intent banking dataset) keyword controlled generation nli classification nontoxic feedback generation persona grounded generation persuasion generation persuasion present persuasion strategy question generation recovery generation response generation response generation with n words schema based generation slot present slot value generation summarization target controlled generation toxic classification
Evaluation	ID-BART0	act classification act generation advice present answer generation answer selection beginswith controlled generation belief state generation db based generation deal present document grounded generation emotion generation emotion tagging endswith controlled generation graph based generation intent classification intent present keyword controlled generation knowledge grounded generation nli classification persona grounded generation persuasion generation persuasion present question generation relation classification relation present response generation schema based generation slot present slot value generation summarization target controlled generation
Dialog State Generation	ID-BART0	act classification act generation advice generation advice present answer generation answer selection beginswith controlled generation count response words db based generation deal present dialfact classification dialog state generation (no multi-woz) document grounded generation edit generation emotion generation emotion tagging endswith controlled generation fill missing utterance find incoherent utterance find missing utterance find swapped utterance gensf slot tagging graph based generation intent classification intent present keyword controlled generation knowledge grounded generation nli classification nontoxic feedback generation persona grounded generation persuasion generation persuasion present persuasion strategy question generation recovery generation relation classification relation present response generation response generation with n words schema based generation slot present slot value generation summarization target controlled generation toxic classification
Slot Filling	ID-BART0	act classification act generation answer generation answer selection beginswith controlled generation belief state generation count response words db based generation deal present dialfact classification document grounded generation edit generation emotion generation emotion tagging endswith controlled generation eval binary eval ranking eval rating fill missing utterance find incoherent utterance find missing utterance find swapped utterance intent classification intent present keyword controlled generation knowledge grounded generation nli classification nontoxic feedback generation persona grounded generation persuasion generation persuasion present persuasion strategy question generation recovery generation relation classification relation present response generation response generation with n words schema based generation slot present slot value generation summarization target controlled generation toxic classification

Table 9: List of experiments and their base models. The tasks listed in the right column are all the tasks a base model was trained with for their corresponding experiment.