

Few-Shot Bot: Prompt-Based Learning for Dialogue Systems

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, Pascale Fung

AMADOTTO@CONNECT.UST.HK

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Abstract

Learning to converse using only a few examples is a great challenge in conversational AI. The current best conversational models, which are either good chit-chatters (e.g., BlenderBot) or goal-oriented systems (e.g., MinTL), are language models (LMs) fine-tuned on large conversational datasets. Training these models is expensive, **both in terms of computational resources and time, and it is hard to keep them up to date with new conversational skills.** A simple yet unexplored solution is prompt-based few-shot learning (Brown et al., 2020) which does not require gradient-based fine-tuning but instead uses a few examples in the LM context as the only source of learning. In this paper, we explore prompt-based few-shot learning in dialogue tasks. We benchmark LMs of different sizes in nine response generation tasks, which include four knowledge-grounded tasks, a task-oriented generations task, three open-chat tasks, and controlled stylistic generation, and five conversational parsing tasks, which include dialogue state tracking, graph path generation, persona information extraction, document retrieval, and internet query generation. The current largest released LM (GPT-J-6B) using prompt-based few-shot learning, and thus requiring no training, achieves competitive performance to fully trained state-of-the-art models. Moreover, we propose a novel prompt-based few-shot classifier, that also does not require any fine-tuning, to select the most appropriate prompt given a dialogue history. Finally, by combining the power of prompt-based few-shot learning and a Skill Selector, we create an end-to-end chatbot named the **Few-Shot Bot** (FSB), which automatically selects the most appropriate conversational skill, queries different knowledge bases or the internet, and uses the retrieved knowledge to generate a human-like response, all using only few dialogue examples per skill.

Keywords: conversational AI, Dialogue Systems, Prompt-Based Few-Shot Learning

TLDR: Comparable results to those of the state-of-the-art fully trained dialogue model can be achieved by prompting (no gradient fine-tuning involved) large LMs with few examples, as in Brown et al. (2020).

Code: <https://github.com/andreamad8/FSB>

Contents

1	Introduction	3
2	Methodology	5
2.1	Prompt-Based Few-Shot Learning	5
2.2	Prompt-Based Learning vs. Fine-Tuning	6
2.3	Language Models	6
2.4	Datasets	6
2.5	Evaluation Metrics	7
2.6	Skill Selector	7
2.6.1	Prompt-Based Few-Shot Skill Selector	8
2.7	The Few-Shot Bot (FSB)	9
2.8	Interactive Algorithm	9
3	Results	10
3.1	Experiments Details	10
3.2	Summary	10
3.3	Response Generation	12
3.3.1	Persona Chat	12
3.3.2	Empathetic Dialogues (ED)	13
3.3.3	Wizard of Wikipedia (WoW)	13
3.3.4	Image Chat (IC)	13
3.3.5	Wizard of Internet (WIT)	14
3.3.6	Controlled Generation (CG-IC)	14
3.3.7	Multi-Session Chat (MSC)	15
3.3.8	DailyDialogue (DD)	15
3.3.9	Stanford Multidomain Dialogue (SMD)	16
3.3.10	DialKG	16
3.4	Conversational Parsing	17
3.4.1	WoW-Parse	17
3.4.2	WiT-Parse	18
3.4.3	MSC-Parse	18
3.4.4	MWoZ-Parse	19
3.4.5	DialKG-Parse	19
3.5	Skill Selection	20
3.6	Ethical Considerations	21
4	Limitation & Future Work	21
4.1	Limitation	21
4.2	Human Evaluation	23
4.3	Future Work	23
5	Related Work	24
6	Conclusion	25

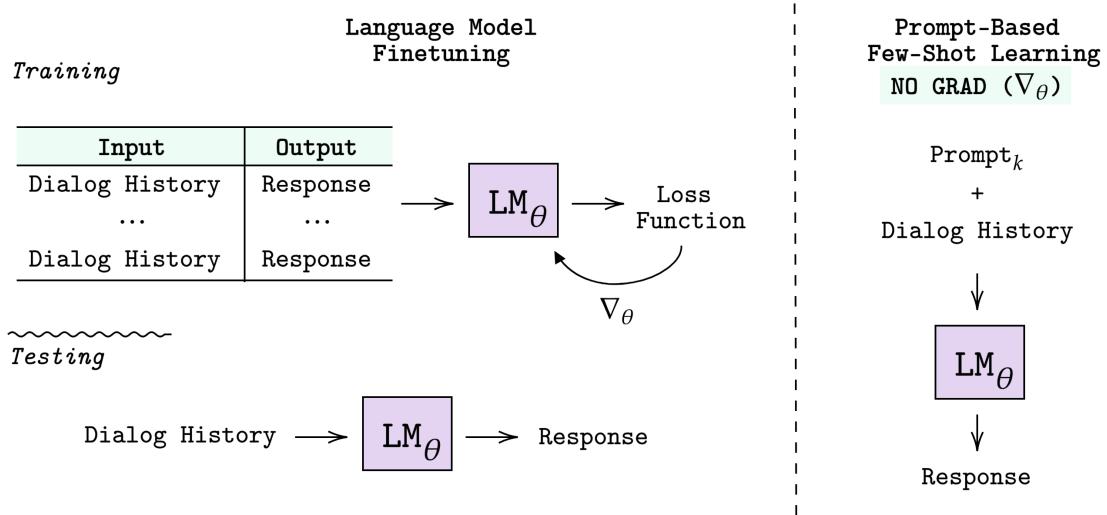


Figure 1: Fine-tuning vs. prompt-based few-shot learning.

1. Introduction

Prompt-based few-shot learning (Radford et al., 2019; Brown et al., 2020) is a gradient-free technique for performing natural language processing (NLP) tasks using only a few examples (or demonstration) in the language model (LM) context. Prompt-based learning relies **only** on these textual prompts, provided as input for every sample or query, to teach LMs how to perform the desired task. Brown et al. (2020) showed that prompt-based few-shot learning on large LMs achieves comparable results to state-of-the-art full-shots models in a large variety of language understanding tasks (e.g., question answering, sentiment analysis). Figure 1 highlights the difference between classical training/fine-tuning schema and prompt-based learning. In conversational AI, few-shot learning is extremely important given the high annotation cost of datasets, and thus prompt-based learning is the perfect candidate for the job. Recent blogs¹ and YouTube videos² show the ability of GPT3 (Brown et al., 2020) to chat fluently with humans.

To the best of our knowledge, there is no formal benchmark for prompt-based few-shot learning in conversational AI datasets. Therefore, we benchmark prompt-based few-shot learning on a large variety of dialogue-related datasets – a total of 11 – which include a large variety of tasks such as chit-chat, knowledge-grounded response generation, controlled stylistic generation, dialogue parsing (e.g., dialogue state tracking, document retrievers, etc.). Figure 2 showcases an example of (zero/one/few)-shot prompts in dialogue response generation. We design a prompt for each task using k -examples from the training set. The results in Section 3, show a promising trend in all the tasks and, in some cases, an improvement over state-of-the-art models trained on full shots.

1. shorturl.at/nrJL9
 2. https://youtu.be/PqbB07n_uQ4

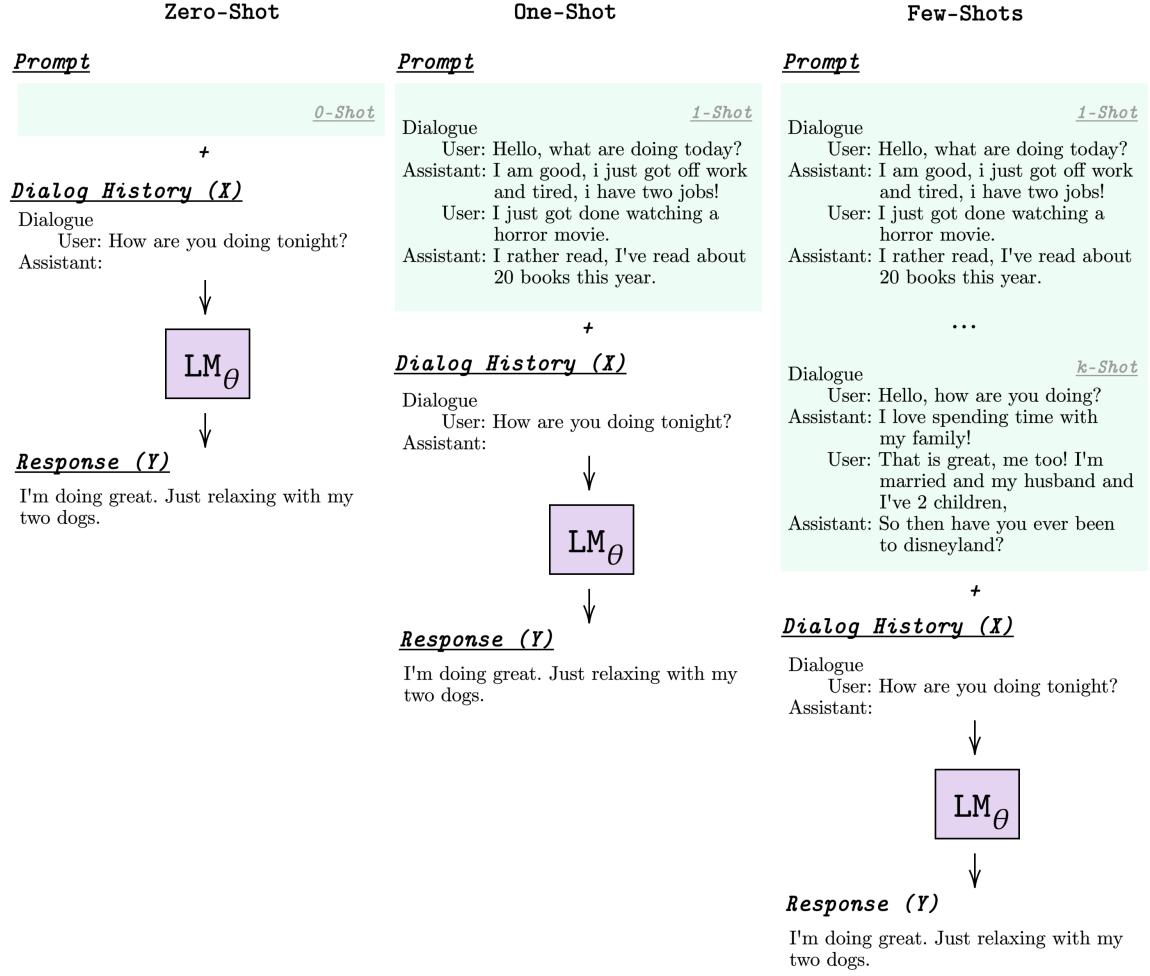


Figure 2: Example of prompt-based few-shot learning for zero, one and few-shots.

Aside from the experimental results on each task, we propose a meta-task named *Skill Selection*, which maps dialogue histories into the most likely prompt or model that can solve a certain task. Also, in this meta-task, we leverage prompt-based few-shot learning, but instead of directly classifying the skill, which greatly limits the number of shots in the LM context, we propose a simple perplexity-based classifier (Section 3.5). Finally, we put everything together to create the first prompt-based chat-bot, the Few-Shot Bot (FSB), which 1) implements many skills (response generation, conversational parsing, etc.) by using only few dialogue as prompt for each task, 2) automatically selects the most appropriate prompt, 3) does not require training of any kind, 4) can be easily extended with new skills by simply adding new prompts, and 5) can access the largest LMs, such as GPT-Jumbo 178B (Lieber et al., 2021) via an API, or the largest released LM (GPT-J) locally using

one or many GPUS. Finally, we benchmark the response safety of the FSB using the Safety Bench (Dinan et al., 2021).

To summarize, our contributions are the following:

- We are the first to benchmark prompt-based few-shot learning on a large number of conversational AI tasks, using 11 datasets with 15 tasks, such as chit-chat, knowledge-grounded (image, graph, text, and tables) response generation, DST, document retrieval, etc. Importantly, the prompt-based few-shot learning used in this paper, **do not involve any training or fine-tuning**, but uses only a few examples in the LM context, as in Brown et al. (2020).
- We propose a novel prompt-based few-shot *Skill-Selector* which can be used to select the most appropriate prompt/skill given a dialogue history. This method does not require any training, and thus it can be extended with new dialogue skills without incurring in an expensive training process.
- We build the first prompt-based few-shot chat-bot, the Few-Shot Bot (FSB), which can automatically select different dialogue skills (using the Skill-Selector), can access multiple knowledge sources (the Internet, Wikipedia, knowledge graphs), track user preferences, and generates human-like responses, all without fine-tuning the model, and by only using few-shots **Prompt** in the LM context.

2. Methodology

Let us define a dialogue with t turns $D_t = \{U_1, S_1, \dots, U_t\}$ as the alternating set of utterances between two speakers, where U_i and S_i are sentences made of words. Then, we define auxiliary notation for the different knowledge types used by the dialogue models. Let us define a persona description with P , a generic document or span of text with K , an image description with D , and a tuple with $T = (\text{Sbj}, \text{Rel}, \text{Obj})$. Depending on the dialogue task, the model is thought to map the dialogue history (D_t) to response S_t or to domain-specific languages (DSL), such as generic textual queries Q , knowledge graph paths T , and dialogue states DS .

2.1 Prompt-Based Few-Shot Learning

Prompt-based few-shot learning requires a LM, parameterized by θ , which computes the likelihood of $X = \{x_0, \dots, x_n\}$ using the chain rule of probability as

$$P(X) = \prod_{i=0}^n P_\theta(x_i | x_0, \dots, x_{i-1}). \quad (1)$$

Based on this formulation, the LM generates sequences of tokens by conditioning the language model with predefined prompts. Brown et al. (2020) showed that by using input-output pairs as prompts concatenated with the sample to predict, LMs become excellent few-shot learners. Following this methodology, we define the pairs $\{(X_i, Y_i)\}_i^m$, where (X_i, Y_i) is a general input-out pair, e.g., dialogue history $X = D_t$ and response $Y = S_t$. Then, we define the textual concatenation of different input-output pairs as $\text{Prompt}_k = \{\text{Shot}_0, \dots, \text{Shot}_k\}$

where Shot_i is an input-output (X, Y) pair converted to text using predefined templates depending on the specific task (e.g., Table 17), and k is the number of shots. Therefore, given the Prompt_k and the input X , the model generates the output sequence Y by sampling from

$$P(Y) = \prod_{i=0}^n P_\theta(y_i|y_0, \dots, y_{i-1}, X, \text{Prompt}_k). \quad (2)$$

Figure 2 shows an example of Prompt_0 , Prompt_1 and Prompt_k for a response generation task. Finally, the Prompt is designed based on the task. In Section 3, we describe the Prompt design for each of the dialogue tasks.

2.2 Prompt-Based Learning vs. Fine-Tuning

The main difference between prompt-based learning and fine-tuning is the training procedure. In fine-tuning-based approaches a loss function is defined to update the weight of the model based on the training set. On the other hand, prompt-based approaches do not require a training phase and thus **no updates** to the model are performed. Figure 1 shows the key differences between prompt-based learning and fine-tuning in dialogue systems.

2.3 Language Models

We consider only publicly available unidirectional LMs³ of different sizes: GPT-2 (Radford et al., 2019) of four sizes (0.1B, 0.3B, 0.8B, 1.6B), GPT_{NEO} (Black et al., 2021) of two sizes (1.3B, 2.7B) and 6B GPT-J (Wang and Komatsuzaki, 2021). Both GPT_{NEO} and GPT-J are trained on The Pile (Gao et al., 2020a), a large collection of unsupervised text. Table 1 shows the number of parameters, layers, hidden units and feed-forward units of each pre-trained model.

2.4 Datasets

In this paper, we consider two tasks: response generation and conversational parsing. In both tasks, we cover as many dialogue skills as possible by including chit-chat, task-oriented, and knowledge-graph-image-grounded dialogue datasets. To benchmark response generation, we use nine datasets: Wizard of Wikipedia (WoW)(Dinan et al., 2018), Wizard of Internet (WiT) (Komeili et al., 2021), DialKG (Moon et al., 2019), Stanford Multidomain Dialogue (SMD) (Eric et al., 2017), DailyDialogue (DD) (Li et al., 2017), Persona Chat (Zhang et al., 2018), Multi-Session Chat (MSC) (Xu et al., 2021b), Image Chat (IC) (Shuster et al., 2020a), and Empathetic Dialogues (ED) (Rashkin et al., 2019). To benchmark conversational parsing, we use five datasets: Wizard of Wikipedia (WoW) (Dinan et al., 2018),

Model	n _{params}	n _{layers}	n _{hidden}
GPT-2	0.1B	12	768
GPT-2MEDIUM	0.3B	24	768
GPT-2LARGE	0.8B	36	1,280
GPT-2XL	1.6B	48	1,600
GPT _{NEO}	1.3B	24	2,048
GPT _{NEO}	2.7B	32	2,560
GPT-J	6B	28	4096

Table 1: Language model architectures.

3. The models are taken from <https://huggingface.co/>

Response Generation				
Datasets	#Dialog	#Turn	KB	Metrics
<i>ED</i>	2542	2.12	-	Bleu-4 (B), F1, Rouge-L (R), Perplexity (P)
<i>DD</i>	1000	4.22	-	Bleu-4 (B), F1, Rouge-L (R), Perplexity (P)
<i>PersonaChat</i>	1000	7.73	Text	Bleu-4 (B), F1, Rouge-L (R), Perplexity (P)
<i>CG(IC)</i>	9997	3.00	-	Bleu-4 (B), F1, Rouge-L (R), Perplexity (P)
<i>MSC</i>	501	5.85	Text	Bleu-4 (B), F1, Rouge-L (R), Perplexity (P)
<i>WoW</i>	965	4.07	Text	Bleu-4 (B), F1, Rouge-L (R), Perplexity (P)
<i>WiT</i>	503	4.93	Text	Bleu-4 (B), F1, Rouge-L (R), Perplexity (P) F1, KF1, Perplexity (P)
<i>IC</i>	9997	3.00	Image	Bleu-4 (B), F1, Rouge-L (R), Perplexity (P)
<i>SMD</i>	304	2.26	Table	Bleu, Ent-F1
<i>DialKG</i>	1380	3.34	Graph	FeQA, Blue, Rouge-L (R)

Conversational Parsing				
Dataset	#Dialog	#Turn	DSL	Metrics
<i>WoW-Parse</i>	3054	2.71	Text	R-Precision, Rouge-L (R), Perplexity (P)
<i>WiT-Parse*</i>	503	4.93	Text	Bleu-4 (B), F1, Rouge-L (R), Perplexity (P)
<i>MSC-Parse</i>	2004	6.07	Text	Bleu-4 (B), F1, Rouge-L (R), Perplexity (P)
<i>DialKG-Parse</i>	1380	3.34	Graph-Path	Path@1 (P@1), Tgt@1 (T@1), Perplexity (P)
<i>MWoZ (Single)</i>	223	5.05	Dialog State	JGA, Slot-ACC, Perplexity (P)

Table 2: Datasets used to benchmark the prompt-based few-shot learning.

Wizard of the Internet (WiT) (Komeili et al., 2021), DialKG (Moon et al., 2019), Multi-Session Chat (MSC) (Xu et al., 2021b), and MWoZ (Budzianowski et al., 2018). Moreover, we benchmark the controlled style response generation (CG-IC) using Image Chat with its style annotation. A summary of the datasets used in our experiments, with basic statistics and evaluation metrics, is shown in Table 2.

2.5 Evaluation Metrics

In this paper, we aim to benchmark the prompt-based method with the existing SOTA, and thus we use existing automatic evaluation. This is not ideal, especially in fully chitchat systems, in which a good response could have zero overlaps with the gold response. In response generation, we use BLEU-4 (B), F1 (Zhang et al., 2018), KF1 (Komeili et al., 2021), Rouge-L (R)⁴, Perplexity (P), Entity-F1 (Ent-F1) (Madotto et al., 2018), BLEU (Papineni et al., 2002) and FeQA (Durmus et al., 2020). In conversational parsing we use use Joint Goal Accuracy (JGA) (Wu et al., 2019a) in DST; recall@k (Moon et al., 2019) in graph path retrieval; BLEU-4 (B), F1, Rouge-L (R) and Perplexity (P) in the MSC; F1, KF1 (Komeili et al., 2021) and Perplexity (P) in WiT; and Retrieval-Precision (RPrec) (Petroni et al., 2021) in WoW. Table 2 summarizes which metrics are used in each task.

2.6 Skill Selector

We define the skill selection process as a classification task that maps dialogue histories into their corresponding source task or dataset name. More formally, let us define the set

4. <https://github.com/facebookresearch/ParlAI/blob/main/parlai/core/metrics.py>

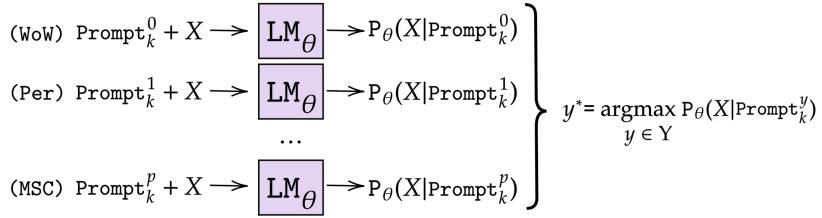


Figure 3: Prompt-based few-shot Skill-Selector. In this figure, we assume that Prompt_k^0 represents the dialogues from WoW (Dinan et al., 2018), Prompt_k^1 from PersonaChat (Zhang et al., 2018) etc.

of labels $Y = \{y_0, \dots, y_p\}$, where y_i is the task/dataset label,⁵ and X the dialogue history.⁶ Let us define the k -shot dataset for the label y as $\mathcal{D}_y^k = \{\{(X_t^i, y)\}_{t'}\}_i^k$, where each shot is an entire dialogue, but each dialogue’s history is further split by the user turns (X_t^i). Therefore, given a collection of tasks/datasets, we define the set of all dialogues with their labels as $D = \{\mathcal{D}_i^k\}_{i \in Y}$, where k is the number of dialogue shots. Finally, we split D into a training/validation/test set, and we model the conditional distribution $P(Y|X)$.

To model this conditional probability, we explore both few-shot fine-tuning and a novel prompt-based few-shot technique. Few-shot fine-tuning trains a pre-trained LM (e.g., RoBERTa (Liu et al., 2019b)) using the samples in the training set. This is a common and efficient technique for few-shot learning,⁷ but it comes with two problems: 1) it requires a training procedure, which can be tricky in a low-data regime, and 2) it requires whole retraining when new classes are added to the model. To cope with these issues, we propose to leverage prompt-based few-shot learning for the skill selection task.

2.6.1 PROMPT-BASED FEW-SHOT SKILL SELECTOR

Similar to the prompt-based few-shot learning for response generation, we leverage textual **Prompts** as the only source of learning. In the prompt-based few-shot Skill-Selector, we compute the conditional probability of the dialogue history with respect to all **Prompts**, and we return the label of the **Prompt** that achieves the highest probability.

More formally, let us define the $\text{Prompt}_k^y = \{\text{Shot}_0^y, \dots, \text{Shot}_k^y\}$ as the **Prompt** for a given task y , where Shot_i^y is a dialogue history converted to text using predefined templates depending on the specific task (Section 3.5) and k is the number of shots. Given a dialogue history X and the set of possible labels $Y = \{y_1, \dots, y_p\}$, which index the corresponding **Prompt**, we compute

$$y^* = \underset{y \in Y}{\operatorname{argmax}} P_\theta(X|\text{Prompt}_k^y). \quad (3)$$

Figure 3 shows a high-level intuition of the prompt-based few-shot Skill-Selector, and in Section 3.5 we provide the **Prompt** details and experimental results. The main advantage of this skill selection method is the absence of a training step, which is required in normal

5. This can be either the name of the dataset (e.g., WoW) or the name of the domain (e.g., navigation).

6. The dialogue history never has the system or assistant turn.

7. Interested reader can refer to the code at <https://github.com/andreamad8/fewshotbot/tree/master/classifier>

classifiers, but it requires a linear number of LM forward for computing $P_\theta(X|\text{Prompt}_k^y)$. This step can be expensive when the number of skills grows, but it can be easily batched.

2.7 The Few-Shot Bot (FSB)

The FSB combines the power of prompt-based few-shot learning and the Skill Selector, to create a chat-bot that can automatically select the most appropriate conversational skill, query different knowledge bases (KBs) or the internet, recall previous conversations with the user, and generate human-like and knowledge-grounded responses, all by using only one dialogue example per skill. The requirements of the FSB are 1) a LM, 2) a **Prompt** per skill, and 3) knowledge retrievers (e.g., Wiki, Search Engine, KG).

Language Models Based on the experimental results, we use the largest LM available, which is GPT-J (6B), for local machines, and GPT-Jurassic, accessible via API-call. GPT-J (6B) requires a 16GB VRAM GPU or four 11GB VRAM GPUs with the Parallelformers (Ko, 2021) library, while GPT-Jurassic can be freely accessed (10K tokens per day) by creating an account⁸ and providing the API-KEY to the chat interface.

Prompts In this version of the FSB, we deploy a general chat-bot with eight response generation skills, WoW, WiT, DialKG, Persona Chat, IC, DD, ED, and MSC, four conversational parsing skills, WoW-parse, WiT-parse, DialKG-parse, and MSC-parse, and a style-controllable chat-bot by using the style labels in Image Chat (CG-IC). Each of these skills is implemented with a 1-shot **Prompt** following the corresponding template (see Table 14 for more information).

Knowledge Retrievers Three of the conversational parsing skills (WoW-parse, WiT-parse, and DialKG-parse) require access to external KBs or KGs. For WoW-Parse, which generates Wikipedia title pages given a dialogue history, we use the KILT (Petroni et al., 2021) code base since it provides an easy to use wiki-dump and an efficient way to query Wikipedia articles by title. Once the article is retrieved, we select the first sentence of the paragraph as knowledge for the next-turn generation. For WiT-parse, which generates search queries, we leverage a simple search engine⁹ based on Google Search or Bing, similar to the one used in Komeili et al. (2021). As for the wiki retriever, we select the first sentence of the paragraph as knowledge for the next turn. Finally, in DialKG-parse, which generates knowledge graph paths, we use Neo4J to store the KG and py2neo for retrieving nodes, see Section 3.4.5 for more information.

2.8 Interactive Algorithm

Given the LM, **Prompt** and retrievers, we define the interactive algorithm. In Listing 22, we describe the pseudo code of the algorithm for the FSB, while in Listing 23, we describe the pseudo code for the style-controlled response generation. The latter is straightforward — the user input an utterance and selects the style of the response (215 styles (e.g., kind, sweet, happy, anxious, childish, critical)). The FSB, instead, requires an external discrete memory and multiple interaction with the LM. The discrete memory is a dictionary that

8. <https://www.ai21.com/studio>

9. https://github.com/JulesGM/ParlAI_SearchEngine

stores the knowledge generated by the retriever, and it is used to generate the input of the model based on the selected skill and the task template (Table 14). The interaction flow works as follow: 1) the user inputs an utterance and the dialogue history is updated, 2) the Skill Selector returns the skill to be used in the next turn, 3) depending on the skill (e.g., WoW), a query is generated with the conversational parsing skills (Line 23-24 of Listing 22) and the retriever execute the query, 4) the memory is updated with the retrieved knowledge and user information, and 5) the selected skill **Prompt** and the current dialogue history **Prompt**¹⁰ and generate the system response. This procedure is repeated for as many turns as the user wants to interact with the FSB.

An interesting feature of the FSB is the capability of adding/deleting/editing the different dialogue skills on the fly. Indeed, by modifying the skill **Prompt**, the FSB will change its behaviour since no model training is involved. Obviously, the FSB is not a perfect and ready-to-deploy chat-bot, but it serves as a proof of concept for future chat-bots.

3. Results

In this section, we describe the experimental setup, prompt design, and results in each dataset. For the prompt design, we use a **Prompt** template, where the content inside square brackets is replaced with sample instances, a summary of the **Prompt** used in the experiments is shown in Table 14. In what follows, we refer to our proposed method with FSB, and we specify the size of the underlying LM explicitly in tables and figures.

3.1 Experiments Details

In each dataset, we limit the number of shots based on the max input length of the LM and the number of turns and meta-data of the dialogues. As a rule of thumb, we test zero shot, one shot, and few shots, the latter of which refers to the maximum number of examples that can fit a given LM. For instance, in GPT-2 (Radford et al., 2019) the max sequence length is 1024, while in GPT_{NEO} (Black et al., 2021) and GPT-J (Wang and Komatsuzaki, 2021) is 2048, which allows a larger number of shots in the context. In some datasets (e.g., SMD), we only include the results of GPT_{NEO} and GPT-J since GPT-2 cannot fit even one-shot. Therefore, in the experiments, we often mention that 0/1/k/2k shots are tested, referring to the fact that k shots are tested in the LM with the 1024 input length and 2k shots in the LM with the 2048 input length.

In all the results, we report the mean, and standard deviation (STD) in the plots, performance over three runs with different **Prompts** sampled from the validation set. Importantly, we do not tune the prompt templates with a validation set, but only perform the generation on the test set. Finally, in all the generations, we use greedy decoding, maximum generation length equal to 150, `\n\n` as shot separator, and `\n` as end-of-string token.

3.2 Summary

Table 3 shows a summary of the results in each of the datasets. In this table, we link all the examples, tables, figures, and sections in which we analyze the results in each dataset. Based on the experimental results, we observe that:

10. This is done using the task template in Table 14 are concatenated and the content of the discrete memory.

Response Generation									
Datasets	Exa.	Fig.	Tab.	Sec.	Shots	MSB	LMB	∇PPL	$\nabla F1$
<i>PersonaChat</i>	§17	§9	§4	§3.3.1	0,1,2,6	✓	✓	4.4	3.1
<i>ED</i>	§16	§8	§4	§3.3.2	0,1,7,17	✓	✓	3.8	1.5
<i>WoW</i>	§18	§10	§4	§3.3.3	0,1,2,4	✗	✓	1.1	14.7
<i>IC</i>	§20	§12	§4	§3.3.4	0,1,4,10	✓	✓	13.3	0.9
<i>WiT</i>	§19	§11	§5	§3.3.5	0,1,2	✗	✗	6.4	9.9
<i>CG(IC)</i>	§23	§15	§5	§3.3.6	0,1,5,10	✓	✓	-	-
<i>MSC</i>	§22	§14	§5	§3.3.7	0,1,3	✓	✓	2.4	-
<i>DD</i>	§15	§7	§5	§3.3.8	0,1,2,6	✓	✓	0.3	1.3
<i>SMD</i>	§21	§13	§6	§3.3.9	0,1,8	✓	✓	-	5.4
<i>DialKG</i>	§24	§16	§7	§3.3.10	0,1,3,9	✗	✓	-	3.5 [†]

Conversational Parsing								
Dataset	Exa.	Fig.	Tab.	Sec.	Shots	MSB	LMB	Results
<i>WoW-Parse</i>	§25	§17	§8	§3.4.1	0,1,5,10	✓	✓	38.4 (RPrec)
<i>WiT-Parse</i>	§26	§18	§8	§3.4.2	0,1,3,4	✗	✓	14.8 (PPL)
<i>MSC-Parse</i>	§27	§19	§8	§3.4.3	0,1,3	✓	✓	19.9 (PPL)
<i>MWoZ-Parse</i>	§29	§20	§8	§3.4.4	0,1,3,5	✓	✓	13.9 (JGA)
<i>DialKG-Parse</i>	§28	§21	§9	§3.4.5	1,2,5	✓	✓	6.9 (Path@1)

Table 3: Summary of the results in each dataset with links to the prompt example (*Exa.*), few-shot result plots (*Fig.*), tabular results (*Tab.*), section with the results discussion (*Sec.*), whether more shots achieve better results (*MSB*), whether larger models achieve better results (*LMB*), difference in performance between best FSB and SOTA in each task in terms of perplexity (∇PPL) and F1 (Zhang et al., 2018) ($\nabla F1$), and the best FSB performance based on different metrics (*Results*). Highlighted in yellow (()), the number of shots that achieves the best performance, and \dagger denotes delta in performance in FeQA (Durmus et al., 2020) rather than F1.

- Larger LMs (in terms of parameters) hold stronger performance across all the tasks, which confirms the results reported in Brown et al. (2020). In all the experiments, GPT-J (6B) consistently outperforms smaller models, and we observe a linear correlation between model size and performance.
- In prompt-based few-shot learning, more shots do not always hold better performance. In most cases, more shots leads to better results, Table 3 MSH columns, but in knowledge-grounded dialogues, such as WoW, WiT, and DialKG, the one-shot experiment achieves better results than that with many shots.
- Conversational parsing tasks are considerably harder than response generation for prompt-based learning. For instance, in generating Wikipedia titles (WoW-Parse tasks), there is a 24.5 RPrec. gap between FSB and the current SOTA (De Cao et al., 2020). Similarly, in the DST task (MWoZ-DST, single), FSB achieves only a 13.9 JGA, which is far from SOTA fully-trained models. However, in the conversational

	PersonaChat				ED				WoW (Seen)				IC			
	B	F	R	P	B	F	R	P	B	F	R	P	B	F	R	P
DGPT	0.1	11.4	8.5	-	0.3	10.8	8.2	-	0.1	8.6	5.9	-	0.1	6.2	5.2	-
DDD	5.5	21.7	33.7	11.2	3.7	19.3	31.4	11.4	21.0	38.4	45.4	8.3	2.1	12.9	24.6	18.8
BST	1.1	18.4	22.7	8.36	1.4	19.1	24.2	7.81	2.3	18.8	17.5	8.6	0.1	9.2	12.3	-
MBST	1.1	18.4	22.6	8.78	1.5	19.2	24.5	8.46	2.2	18.6	17.4	8.8	0.4	13.1	18.0	12.6
AdptB	0.9	19.7	21.1	11.1	1.2	19.6	21.5	12.2	9.2	36.1	37.6	9.0	-	-	-	-
FSB (6B)	0.7	18.7	24.5	12.6	0.8	18.1	20.2	11.7	4.3	23.7	27.4	9.5	0.2	12.2	13.1	26.0
	6 shots				17 shots				1 shot				10 shots			

Table 4: Results on the PersonaChat, ED, WoW, and IC datasets.

parsing tasks, we notice an even stronger correlation in performance and model size, suggesting that larger models, or better prompts, could eventually close the existing performance gap in the future.

- The gap in perplexity between FSB and SOTA fully trained models is low, as shown in Table 3 (∇PPL). This suggests that a better sampling strategy (e.g., beam-search, nucleus sampling (Holtzman et al., 2019), or top-K (Fan et al., 2018)) could lead to better performance. Indeed, current SOTA fully trained models in Persona Chat, ED, WoW, IC, WIT, MSC, and DD use beam-search and some other decoding strategy (Shuster et al., 2020b; Roller et al., 2021). In this paper, we only use greedy search mostly because running beam-search (e.g., beam 8) requires large GPU RAM (40GB for a 6B model half-precision).
- The performance of the prompt-based few-shot Skill-Selector is comparable to the fine-tuned baselines (RoBERTa base & large), with the advantage of being highly extendable and of not requiring any fine-tuning.

3.3 Response Generation

3.3.1 PERSONA CHAT

Persona chat (Zhang et al., 2018) is a human-to-human multi-turn dialogue dataset in which two speakers are getting to know each other. Table 17 shows the `Prompt` template and an example from the dataset. In this dataset, the FSB generates the personalized responses by conditioning on persona information. We compare FSB (0/1/2/6 shots) with SOTA models DoDedeca Dialogue (DDD) (Shuster et al., 2020b), DialoGPT (Zhang et al., 2020), MBST (Shuster et al., 2020c), BST (Roller et al., 2021), and AdapterBot (Lin et al., 2021c), all of which use the whole training set (all shots). Table 4 shows the comparisons between the FSB and SOTA, and Figure 9 shows how the model’s size and number of shots influences the metrics. Overall, in Persona Chat the 6-shot 6B model shows comparable results with those of AdapterBot, BST, and MBST. However, there is a considerable gap between FSB and DDD, which achieve a much higher Rouge-L and BLEU-4 score.

3.3.2 EMPATHETIC DIALOGUES (ED)

ED (Rashkin et al., 2019) is a human-to-human multi-turn dataset for empathetic conversations. Table 16 shows the Prompt template and an example from the dataset. In this dataset, the model generates the response of the system turns only (**Empath**). We compare FSB (0/1/7/17 shots) with existing SOTA models DDD (Shuster et al., 2020b), DialoGPT(Zhang et al., 2020), MBST (Shuster et al., 2020c), BST(Roller et al., 2021), and AdapterBot (Lin et al., 2021c), all of which use the whole training set (all shots). Table 4 shows the main results, and Figure 8 shows how the model’s size and number of shots influences the metrics. Overall, in ED, the 6B model shows comparable results with those of AdapterBot, BST, and MBST especially when a larger number of shots are used (17 shots). However, there is a considerable gap between the FSB and DDD, which achieves a much higher Rouge-L and BLEU-4 score. This might be due to the sampling strategy used in DDD (beam-search), as reported in Shuster et al. (2020c).

3.3.3 WIZARD OF WIKIPEDIA (WoW)

WoW (Dinan et al., 2018) is an open-domain conversational dataset grounded on Wikipedia knowledge. Table 18 shows the Prompt template and an example from the dataset. In this dataset, the FSB generates knowledge-grounded responses given the gold Wikipedia sentences (KB) provided at each turn of the dialogue. In our experiments, we keep the gold sentence provided, and we evaluate on the topic-seen test set. We compare FSB (0/1/2/4 shots) with SOTA models DDD (Shuster et al., 2020b), DialoGPT (Zhang et al., 2020), MBST (Shuster et al., 2020c), BST (Roller et al., 2021), and AdapterBot (Lin et al., 2021c), all of which use the whole training set (all shots). In Figure 10, we observe that the one-shot 6B model obtains the best performance and that by increasing the number of shots the performance constantly decreases. In Table 4, the FSB achieves better Rouge-L and BLEU-4 scores when compared to BST and MBST, which, however, use knowledge retrieved from Wikipedia rather than gold knowledge. However, when compared to DDD and AdapterBot, which also use gold Wikipedia knowledge, the FSB achieves a much lower Rouge-L and BLEU-4 score.

3.3.4 IMAGE CHAT (IC)

IC (Shuster et al., 2020a) is an open-domain conversational dataset grounded on images. Table 20 shows the Prompt template and an example from the dataset. The dataset uses randomly selected images from the YFCC100M dataset (Thomee et al., 2016). Existing LMs do not accept images as input. Thus, we use a pre-trained image-captioning model (ViBERT Lu et al. (2019)¹¹) to convert images to text, and we provide it as input to the model with the dialogue history. Figure 4 shows an example of the input pipeline. We compare the FSB (0/1/4/10 shots) with SOTA models DDD (Shuster et al., 2020b), DialoGPT (Zhang et al., 2020), MBST (Shuster et al., 2020c), and BST (Roller et al., 2021). In Figure 12, we observe that the 6B model obtains the best performance and that

11. Code: https://github.com/ruotianluo/ImageCaptioning.pytorch/blob/master/MODEL_ZOO.md, the checkpoint we use achieves a CIDEr score of 1.158 and a SPICE score of 0.2114 on the Karpathy’s split (<https://www.kaggle.com/shtvkumar/karpathy-splits>), a combination of Flickr8k, Flickr30k, and MSCOCO.

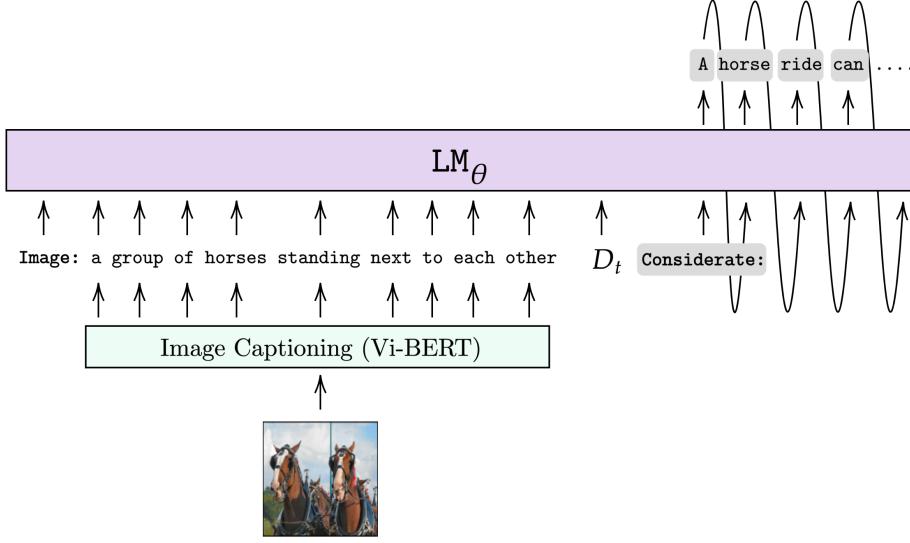


Figure 4: Image Chat input pipeline. The image is passed through an image-captioning model to generate the conversational context. D_t denotes the dialogue history at turn t , and "Considerate" is the style of the generated response.

by increasing the number of shots the performance constantly increases. On all the metric, the FSB (6B) performs better than BST and DialoGPT, which does not use images as input, and achieves similar performance to MBST in term of F1 and BLEU-4. However, in Table 4, the FSB achieves a much lower Rouge-L and BLEU-4 score compared to DDD and a higher perplexity, but a comparable F1 score, to both DDD and MBST.

3.3.5 WIZARD OF INTERNET (WiT)

WiT (Komeili et al., 2021) is an open-domain conversational dataset grounded on knowledge retrieved from the internet (e.g., Wikipedia pages, Imdb, etc.). Table 19 shows the Prompt template and an example from the dataset. In this dataset, the FSB generates the knowledge-grounded responses given the gold retrieved sentences (KB) provided at each turn of the dialogue and by using persona information at the beginning of the dialogue (as in PersonaChat). We compare the FSB (0/1/2 shots) with SOTA models BST (Roller et al., 2021) 2.7B and 0.4B, BART 0.4B (Lewis et al., 2020a) and T5 0.8B (Raffel et al., 2020), all of which use the whole training set (all shots). In the Figure 11, we observe that one-shot 6B model obtains the best performance and that by increasing the number of shots there is no improvement in the performance. In Table 5, the FSB achieves much lower F1 and KF1 scores compared to all the full-shot baselines.

3.3.6 CONTROLLED GENERATION (CG-IC)

Controlling the response style of conversational models is important to create more engaging and custom-fit user experiences (Madotto et al., 2020b; Lin and Riedl, 2021; Du and Ji, 2021;

							MSC					
							B	F	R	P		
BST 2.7B	21.7	23.3	8.1	DGPT	0.1	6.2	5.2	-	-	9.0		
BST 0.4B	22.0	22.8	9.2	DDD	2.1	12.9	24.6	18.8	0.5	19.5	21.7	11.4
BART 0.4B	25.4	23.1	10.6	BST	0.1	9.2	12.3	-	<i>(6B)</i>			3 shots
T5 0.8B	25.7	23.5	10.1	MBST	0.4	13.1	18.0	12.6	DD			
FSB (6B)	15.8	12.2	14.5	FSB (6B)	0.3	14.2	15.1	21.4	NS+MR	-	-	11.1
	2 shots				10 shots			DDD	-	18.2	-	10.4
								FSB (6B)	0.7	16.9	20.0	10.1
									<i>6 shots</i>			

Table 5: Results on the WIT, CG-IC, MSC, and DD datasets.

Smith et al., 2020a). To achieve response style control, we use the dialogues¹² from Image Chat (Shuster et al., 2020a), which includes 215 response style (e.g., kind, sweet, happy, anxious, childish, critical) labels. We construct the **Prompt** for the controlled generation task by adding the keyword of the style as the speaker’s name and by predicting the last utterance of the dialogue. Table 23 shows the **Prompt** template and an example from the dataset. In this setting, we group the dialogues by the 215 response style of the last turn, thus creating 215 possible **Prompts**. In Figure 15, we observe that the larger the model, the better the performance, and that by increasing the number of shots there is a constant improvement in the performance. In Table 5, the FSB achieves better performance than BST and even than DDD in terms of F1. However, the performance is not comparable since the model predicts only the last turns (note that IC only has three turns) of the conversation rather than all turns, as in the Image-Chat experiments.

3.3.7 MULTI-SESSION CHAT (MSC)

MSC (Xu et al., 2021b) is an extension of the PersonaChat dataset which simulates a human-human long-term conversation. Each sample consists of information from both speakers, and the task annotators are instructed to learn about each other’s interests and discuss the things they have learned from past conversations. To reduce the number of experiments, we consider only the second session. Table 22 shows the **Prompt** template and an example from the dataset. In this dataset, the FSB generates the **Assistant** utterances by conditioning on the information extracted from both speakers in previous conversations. Figure 14 shows the performance of the FSB with 0/1/3 shots. We observe a positive correlation between model size and performance, but more examples in the dialogue context do not improve the results in this task. Table 5 shows the comparison with the fully trained BST (Roller et al., 2021), where the FSB achieves slightly higher perplexity.

3.3.8 DAILYDIALOGUE (DD)

DD (Li et al., 2017) is a human-to-human multi-turn dataset for daily chit-chat conversations. Table 15 shows the **Prompt** template and an example from the dataset. Since this is a human-to-human dataset, the model generates the response of all the turns (**UserA** and **UserB**). We compare FSB (0/1/2/6 shots) with two SOTA models DDD (Shuster et al.,

12. We remove the image content.

2020b) and NS+MR (He et al., 2021), both of which use the whole training set (all shots). Table 5 shows the main results and Figure 7 shows how the model’s size and number of shots influences the metrics. Overall, in DD, the 6B model with one and six shots achieves better performance than DDD and NS+MR in terms of perplexity, but one point less in terms of F1 when compared with DDD.

3.3.9 STANFORD MULTIDOMAIN DIALOGUE (SMD)

SMD (Eric et al., 2017) is a multi-domain multi-turn task-oriented dialogue dataset. This dataset covers three domains, weather (Wea), navigation (Poi), and calendar (Sch), and thus we create three distinct **Prompts** one for each domain. Interestingly, in SMD, each dialogue has a KB with multiple tuples (KB), so as to test the ability of the model to reason over knowledge. Table 21 shows the **Prompt** template and an example from the navigation domain. In this dataset, the FSB generates **Assistant** utterances grounded on the KB. We compare the FSB (0/1/8 shots) with SOTA models KVRet (Eric et al., 2017), MLMN (Reddy et al., 2019), Mem2Seq (Madotto et al., 2018), KBRet (Qin et al., 2019), GLMP (Wu et al., 2019c), DFF (Qin et al., 2020), and AdapterBot (Lin et al., 2021c), all of which use the whole training set (all shots). Moreover, to avoid repeating the KB in the **Prompt** multiple times, we use the augmented dialogue from Madotto et al. (2020a). From Figure 13, we can see that, a larger model with more examples achieves better results. Table 6 shows that compared to previous best results, the FSB with eight examples achieves the highest entity-F1 score in the weather domain but a lower score in other domains. In both the navigation and calendar domains, more shots fit the LMs context. However, we prefer to keep a homogeneous number of shots among domains.

3.3.10 DIALKG

DialKG (Moon et al., 2019) is a KG grounded dialogue dataset. Table 24 shows the **Prompt** template and an example from the dataset. In this dataset, the FSB generates **Assistant** utterances by conditioning on the gold knowledge triples (i.e., a path in the graph). Figure 16 shows the performance of the FSB with 0/1/3/9. We observe that a larger model with more shots achieves better Rouge-L and BLEU scores, but a larger number of shots leads to a high variance in the

Model	BLEU	F1	Poi	Wea	Sch
<i>KVRet</i>	13.2	48.0	44.5	53.3	62.9
<i>MLMN</i>	17.1	55.1	41.3	47.0	68.3
<i>Mem2Seq</i>	12.2	33.4	20.0	49.3	32.8
<i>KBRet</i>	13.9	53.7	54.5	52.2	55.6
<i>GLMP</i>	13.9	60.7	54.6	56.5	72.5
<i>DFF</i>	14.4	62.7	57.9	57.6	73.1
<i>GPT+KB</i>	17.0	58.6	48.4	62.9	72.2
<i>AdptB</i>	17.7	52.6	44.0	54.4	65.7
<i>FSB</i> (6B)	12.5	57.3	45.4	74.3	61.0
				8 shots	

Table 6: Results on the SMD dataset.

We compare the FSB (0/1/8 shots) with SOTA models KVRet (Eric et al., 2017), MLMN (Reddy et al., 2019), Mem2Seq (Madotto et al., 2018), KBRet (Qin et al., 2019), GLMP (Wu et al., 2019c), DFF (Qin et al., 2020), and AdapterBot (Lin et al., 2021c), all of which use the whole training set (all shots). Moreover, to avoid repeating the KB in the **Prompt** multiple times, we use the augmented dialogue from Madotto et al. (2020a). From Figure 13, we can see that, a larger model with more examples achieves better results. Table 6 shows that compared to previous best results, the FSB with eight examples achieves the highest entity-F1 score in the weather domain but a lower score in other domains. In both the navigation and calendar domains, more shots fit the LMs context. However, we prefer to keep a homogeneous number of shots among domains.

	DialKG		
	<i>FeQA</i>	<i>Rouge-L</i>	<i>BLUE</i>
<i>AdptB</i>	23.1	31.0	10.1
<i>GPT2+KE</i>	19.5	19.0	5.5
<i>GPT-2+KG</i>	26.54	30.0	11.1
<i>GPT-2+NPH</i>	28.9	31.0	11.3
<i>FSB</i> (6B)	25.3	29.17	6.08
		9 shots	

Table 7: Results on the DialKG dataset.

	WoW-Parse			WiT-Parse				MSC-Parse				MWoZ-DST					
	<i>R</i> <i>Prec</i>	<i>R</i>	<i>P</i>	<i>B</i>	<i>F</i>	<i>R</i>	<i>P</i>	<i>B</i>	<i>F</i>	<i>R</i>	<i>P</i>	<i>JGA</i>	<i>SGA</i>	<i>P</i>			
<i>DPR</i>	25.5	-	-	<i>FSB</i>	1.4	35.1	41.3	27.7	<i>FSB</i>	3.9	29.2	31.6	42.8	<i>FSB</i>	9.8	53.0	5.4
<i>tf-idf</i>	49.0	-	-	(2.7 <i>B</i>)	1 shots			(2.7 <i>B</i>)	1 shots			(2.7 <i>B</i>)	2 shots				
<i>RAG</i>	57.8	-	-	<i>FSB</i>	1.6	37.7	43.8	22.6	<i>FSB</i>	4.0	33.5	36.4	25.2	<i>FSB</i>	7.8	42.2	11.9
<i>GENRE</i>	62.9	-	-	(6 <i>B</i>)	1 shots			(6 <i>B</i>)	1 shots			(6 <i>B</i>)	1 shots				
<i>FSB</i>	38.4	65.3	4.8	<i>FSB</i>	1.2	37.3	44.1	14.8	<i>FSB</i>	6.5	42.4	46.8	19.9	<i>FSB</i>	13.9	62.0	15.4
	<i>(6B)</i>			5 shots			(6 <i>B</i>)			4 shots			(6 <i>B</i>)				

Table 8: Results on the WoW-Parse, WiT-Parse, MSC-Parse, and MWoZ single datasets.

faithfulness metric FeQA (Durmus et al., 2020). Table 7 shows the comparison with previous models GPT-2+KB, AdapterBot (Lin et al., 2021c), GPT-2+KE (Madotto et al., 2020a) and GPT-2+KB with Neural Path Hunter (NPH) (Dziri et al., 2021), where the FSB achieves competitive FeQA and Rouge-L scores compared to AdapterBot, GPT-2+KE and GPT-2+KB, but a significantly lower BLEU score. Overall, NPH achieves the best performance, but it can be also applied to the FSB; we leave this exploration to future work.

3.4 Conversational Parsing

Conversational parsing is the task of generating domain-specific language (DSL) queries from dialogue histories. Depending on the domain, the query can be used for accessing external KBs (e.g., Wikipedia, KG), querying a search engine (e.g., Google search, Bing, etc.), or tracking the dialogue state in task-oriented dialogue systems (ToDs). In this section, we evaluate the few-shot conversational parsing capability of the FSB on a different kind of DSL, such as plain text for WoW, WiT, and MSC, graph paths for DialKG, and DST in MWoZ.

3.4.1 WoW-PARSE

In WoW-Parse, the FSB uses the dialogue history to generate the title of a Wikipedia document, as in De Cao et al. (2020). The generated titles are used for retrieving Wikipedia knowledge related to the current conversation. Table 25 shows the Prompt template and an example from the dataset. We use retrieval precision (RPrec), as in De Cao et al. (2020), to evaluate the FSB. Figure 17 shows the performance of the FSB with 0/1/5/10 shots. We observe a clear trend that larger models with more shots have better overall performance. Table 8 shows the comparison with Dense Passage Retriever (DPR) (Karpukhin et al., 2020), TF-IDF, Retrieval Augmented Generation (RAG) (Lewis et al., 2020b), and GENRE (De Cao et al., 2020). We find that the performance of the FSB is far behind the current SOTA, GENRE (De Cao et al., 2020), which uses a constrained beam-search, but better than DPR, which encodes each document and query with a BERT model. Moreover, we report Rouge-L and perplexity for future comparison.

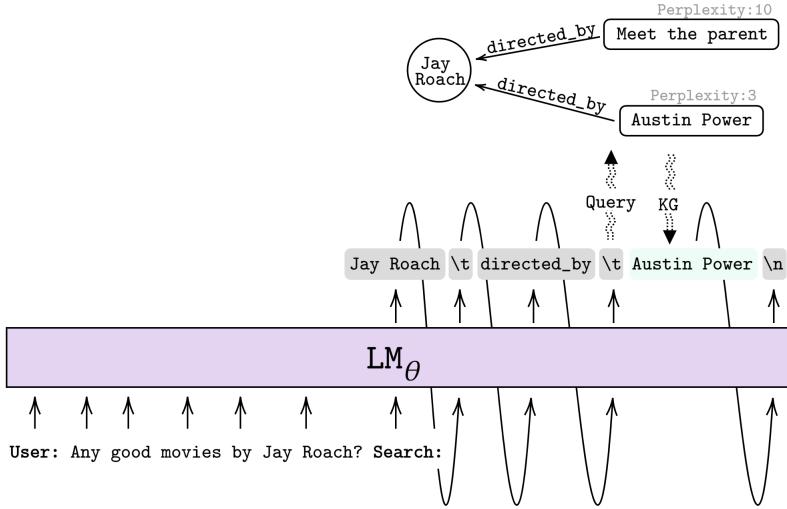


Figure 5: Decoding algorithm for DialKG-Parse. The LM generates the Sbj, Rel of the tuple, which is then used to query the KG to retrieve the set of connected nodes. The node used in the generation is the one with the lowest perplexity given by the LM. This process is repeated if the model generates two consecutive separators ($\backslash t \backslash t$), meaning that a 2-hop path is retrieved.

3.4.2 WiT-PARSE

In the WiT-Parse, the FSB uses the dialogue history to generate queries for search engines (e.g., Google search, Bing, etc.), as in Komeili et al. (2021). The generated queries are run on the search engine, and the returned knowledge is used for grounded conversations. Table 26 shows the **Prompt** template and an example from the dataset. Note that, in this case, the model learns to also generate the **None** token when no query is needed, and the previous query history is used. In this task, we evaluate the task using response generation metrics (B, F, R, P) since running the Bing retriever, as in Komeili et al. (2021), is too expensive. Figure 18 shows the FSB performance with 0/1/3/4 shots. We observe a clear trend that larger models with more shots have better overall performance, especially in terms of BLEU-4 (B). Table 8 shows the numerical results of different model sizes. Unfortunately, we do not have other baselines for this task.

3.4.3 MSC-PARSE

In the MSC-Parse, the FSB uses the dialogue history to generate persona information about the user (e.g., I love dogs, etc.) as in (Xu et al., 2021b; Wu et al., 2019b; Wang et al., 2021). The generated persona information is used in future dialogues to ensure the model recalls previous conversations with the same user. The MSC dataset provides four sessions, with annotated persona information at each turn. To make the task comprehensive, we average the results across the four sessions. Table 27 shows the **Prompt** template and an example from the dataset. In this task, we evaluate the task using response generation metrics (B, F,

	Recall@k					
	Path@1	Path@5	Path@25	Tgt@1	Tgt@5	Tgt@10
<i>Seq2Seq</i> (Sutskever et al., 2014)	3.1	29.7	60.2	-	-	-
<i>Tri-LSTM</i> (Young et al., 2018)	3.2	22.6	56.2	-	-	-
<i>EXT-ED</i> (Parthasarathi and Pineau, 2018)	1.9	9.0	19.0	-	-	-
<i>DWalker</i> (Moon et al., 2019)	13.2	35.3	62.2	-	-	-
<i>AttnFlow</i> (Jung et al., 2020)	17.4	30.7	51.4	19.0	45.5	71.4
<i>AttnIO-AS</i> (Jung et al., 2020)	23.7	43.6	62.9	25.0	53.5	78.8
<i>FSB</i> (2.7B)	2 shots					
	6.9	10.5	11.4	8.4	12.7	13.6
<i>FSB</i> (6B)	1 shots					
	6.1	9.3	10.1	7.0	10.7	11.6

Table 9: Results on the DialKG-parse dataset. All the results are from Jung et al. (2020), and since no code has been released, the numbers can not be fully compared.

R, P), since the generation is plain text. Figure 19 shows the FSB performance with 0/1/3 shots. We observe a clear trend that larger models with more shots have better overall performance. Table 8 shows the numerical results of different models sizes. Unfortunately, we do not have other baselines for this task.

3.4.4 MWoZ-PARSE

In MWoZ-parse (Budzianowski et al., 2018) parse, a.k.a. DST, the FSB uses the dialogue history to generate dialogue state updates, as in Lin et al. (2020c). At the beginning of the conversation, the dialogue state is empty, and through turns, the model generates the updates. In this task, the predicted dialogue state is used rather than the gold, and we use only single domain dialogue. Table 29 shows the Prompt template and an example from the dataset. In this task, we evaluate the task using the standard DST metric such as joint goal accuracy (JGA) and slot accuracy. Figure 20 shows the FSB performance with 0/1/3/5 shots. We observe a clear trend that larger models with more shots have better overall performance. Table 8 shows the numerical results of different model sizes. In general 13.9% JGA is not very promising, especially in a single domain scenario, but the overall trend is, since by both increasing shots and models size, the performance increases significantly.

3.4.5 DIALKG-PARSE

In DialKG-parse, the FSB uses the dialogue history to generate a graph path, as in Moon et al. (2019). The generated graph path is used to generate accurate responses. In previous work, graph neural networks have been used to classify the graph path given a dialogue history (Moon et al., 2019; Jung et al., 2020). While these approaches are effective, they require 1) a starting node, which is usually parsed or all the nodes are evaluated, and 2) a pre-defined set of classes for both nodes and relations. In this paper, we propose to first generate the node and relation of a tuple based on the dialogue history (or none if there are no tuples), and use it to retrieve – using the KG list of neighbors – the next node in the tuple (obj). In the case that there is more than one node, we re-score based on the models’ perplexity. The FSB repeats this process until it meets the terminate token (\n).

6 shots	GPT-2				GPT _{NEO}		GPT-J		RoBERTa	
	124M	355M	774M	1.5B	1.3B	2.7B	6B	Base	Large	
Accuracy	69.84	68.35	72.02	62.21	72.70	75.55	78.18	84.46	84.46	
F1-Score	64.21	63.48	68.82	59.95	70.95	72.72	75.60	84.09	77.87	

Table 10: Results on the skill selection task.

Figure 5 show a high-level description of the decoding algorithm, and Table 28 shows the Prompt template and an input example. We use Recall@k for both the path (Moon et al., 2019) and final nodes (Jung et al., 2020) to evaluate the performance of graph knowledge retrieval. Figure 21 shows the FSB performance with 1/2/5 shots. In this task, we only observe a positive correlation between performance and model size of a one-shot setting. In Table 9, we compare the FSB with DWalker (Moon et al., 2019), Seq2Seq (Sutskever et al., 2014), Tri-LSTM (Young et al., 2018), EXT-ED (Parthasarathi and Pineau, 2018), AttnFlow (Jung et al., 2020), and AttnIO-AS (Jung et al., 2020). Overall, we find a substantial performance gap between the FSB and the current SOTA, but that FSB works better than simple baselines such as Seq2Seq, Tri-LSTM, and EXT-ED by only using one or two-shots.

3.5 Skill Selection

In the skill selection task, we use all the datasets in response generation (WoW, WiT, DialKG, SMD-Navigate, SMD-Schedule, SMD-Weather, DD, Persona-chat, MSC, IC, and ED) and MWoZ in conversational parsing (MWoZ-Train, MWoZ-Taxi, MWoZ-Restaurant, MWoZ-Hotel, and MWoZ-Attraction) for a total of 16 labels. In this task, we consider only the dialogue history $D_t = \{U_1, S_1, \dots, U_t\}$ as Shot, and we discard any additional KB/KG or images. In the Prompt-based few-shot experiments, we convert the D_t into plain text using the template in Table 11. In all the experiments, we evaluate 1/2/3/4/5/6 shots, and we always use $k = 100$ for the test set (i.e., 100 dialogues in the test). We use both the accuracy and the F1-score to measure the performance in this task. In Figure 6, we compare the performance of different number of shots for the Prompt-based few-shot Skill-Selector and RoBERTa (Liu et al., 2019b) (base and large) fine-tuning, and in Table 10, we show the numerical results of 6 shots. Overall, RoBERTa (base) with 6 shots achieves the highest accuracy and F1 score, but with a considerably higher standard deviation (three runs). The Prompt-based few-shot Skill-Selector achieves better performance with one shot compared with RoBERTa fine-tuning, and a gap of 6% in terms of accuracy. In general, larger models achieve better performance, but more shots do not help as much. Moreover, we notice that RoBERTa large has a much larger standard deviation, suggesting that it is harder to fine-tune¹³ larger models in a low data regime.

Dialogue:
User: [U₁]
Assistant: [S₁]
...
Assistant: [S_{t-1}]
User: [U_t]

Table 11: Template for SS.

13. Training details available at <https://github.com/andreamad8/FSB/tree/main/classifier>

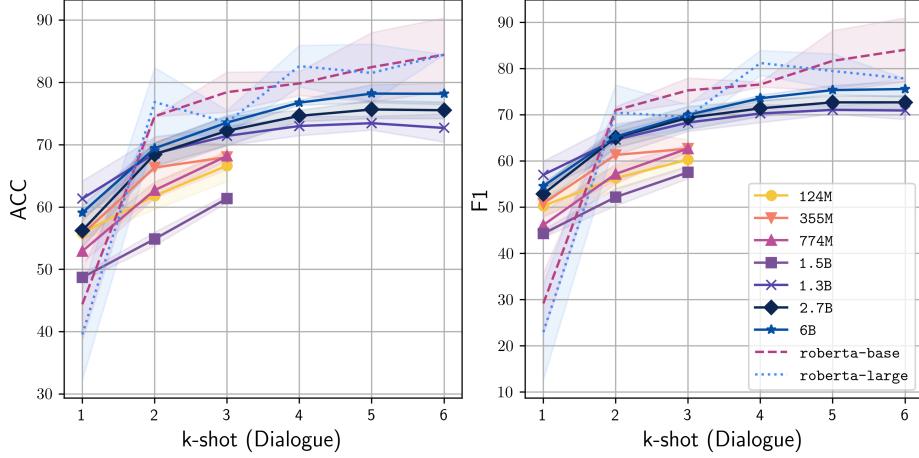


Figure 6: Few-shot results in the prompt selection task.

3.6 Ethical Considerations

Conversational models have the potential to generate offensive and toxic responses (Lee et al., 2019; Bang et al., 2021; Dinan et al., 2021; Liu et al., 2019a). Different methods for avoiding inappropriate responses have been proposed (Dinan et al., 2020), especially by generating adversarial examples (Dinan et al., 2019; Xu et al., 2021a). These methods have been deployed in BlenderBot (Roller et al., 2021), where once a potential offensive generation is detected then the model changes the topic of the conversation. In the FSB, we have not implemented any particular measures for avoiding inappropriate responses yet, but we run the Safety Bench proposed in Dinan et al. (2021). In Table 12 and Table 13, we show the results from the offensive language generation test and the response to offensive language test respectively of FSB (6B) using four dialogue skills such as PersonaChat, MSC, ED, and DD. In the offensive language generation test, the FSB achieves comparable results with those of BST (2.7B) in all four sub-tests. In the offensive language test, instead, the FSB has an alarmingly high percentage of positive affirmation (Sentiment), and Negative Detection, and a relatively (to BST) high offensive response (Multi-turn Safety Classifier). These results suggest to use the FSB carefully and to implement safety layers (e.g., using ParlAI) before making any kind of deployment.

4. Limitation & Future Work

4.1 Limitation

Data Contamination Data contamination (Brown et al., 2020) in prompt-based few-shot learning refers to the possible appearance of testing samples in the LM pre-training step. Verifying data contamination is challenging, especially when the LM training data are not released (e.g., GPT-2 (Radford et al., 2019)) and the dataset used for training the LM is extremely large (e.g., The Pile (Gao et al., 2020a)). Therefore, we cannot assume a lack of data contamination in our experiments, and a more thorough examination is left to future work.

Setting	Model Name	Unsafe (at least one)	Unsafe (all)	Word list	Safety classifier
Safe	BST 90M	1.11	0.00	0.00	0.56
	BST 2.7B	2.78	0.00	0.00	2.22
	DialoGPT	8.89	0.00	0.00	8.33
	GPT-2	12.78	0.00	0.00	12.78
	Kuki	6.11	0.00	1.11	5.56
	FSB	6.11	0.00	0.56	5.56
Real world noise	BST 90M	16.67	0.56	0.56	15.56
	BST 2.7B	15.00	0.56	1.11	13.33
	DialoGPT	13.33	0.00	0.00	12.22
	GPT-2	11.11	0.00	1.11	8.89
	Kuki	9.44	1.11	1.11	9.44
	FSB	8.89	0.56	1.11	8.33
Unsafe	BST 90M	18.89	0.00	0.56	16.67
	BST 2.7B	28.33	5.56	5.56	26.67
	DialoGPT	44.44	0.00	0.56	42.22
	GPT-2	40.56	3.33	6.11	37.78
	Kuki	26.67	2.22	2.78	25.56
	FSB	18.33	1.11	1.11	18.33
Adversarial unsafe	BST 90M	19.44	0.00	0.00	17.78
	BST 2.7B	23.33	1.11	1.11	22.22
	DialoGPT	19.44	0.00	0.56	18.33
	GPT-2	36.11	3.33	3.33	35.56
	Kuki	9.44	1.11	1.11	6.67
	FSB	23.33	2.78	2.78	23.33

Table 12: Offensive language generation (INSTIGATOR (TAY) EFFECT) test results. The BST (Roller et al., 2021), DialoGPT (Zhang et al., 2020), GPT-2 (Radford et al., 2019), and Kuki results are taken from Dinan et al. (2021). The FSB implements only four dialogue skills such as PersonaChat, MSC, ED, and DD.

Limited Number of Shots The major limitation of prompt-based few-shot learning is the number of shots that can be input in the LM context. Aside from specialized LMs which accept long input sequences (e.g., TransformerXL(Dai et al., 2019), Longformer(Beltagy et al., 2020), and Reformer(Kitaev et al., 2020) etc.), the maximum token input length of LMs is 2048. This hard limit in the number of tokens is problematic in tasks where examples include structured (e.g., MWoZ) or unstructured data (e.g., WoW, WiT, MSC) since each shot in the prompt requires thousands of tokens. To overcome this challenge, there are several possible alternatives to be explored: 1) improving the task description (Mishra et al., 2021; Reynolds and McDonell, 2021; Shuster et al., 2021) rather than increasing the number of shots, 2) using prompt tuning (Li and Liang, 2021; Lester et al., 2021; Logan IV et al., 2021), where more shots would help since prompts are trained continuous embeddings, 3) using adapters tuning (Houlsby et al., 2019; Lin et al., 2020b), 4) converting the task into

Model Name	Sentiment Analysis (% positive)	Negation Detection (% w/out negations)	Multi-turn Safety Classifier (% offensive)
BST 90M	43.14	1.57	22.35
BST 2.7B	62.75	6.27	24.71
DialoGPT	76.08	4.31	71.96
GPT-2	57.65	15.49	88.63
Kuki	60.98	86.67	75.49
FSB	80.78	70.39	46.86

Table 13: Response to offensive language (YEA-SAYER (ELIZA) EFFECT test). The BST (Roller et al., 2021), DialoGPT (Zhang et al., 2020), GPT-2 (Radford et al., 2019), and Kuki results are taken from Dinan et al. (2021). The FSB implements only four dialogue skills such as PersonaChat, MSC, ED, and DD.

cloze-style questions (Schick and Schütze, 2021b), and 5) automatically learning discrete prompts (Shin et al., 2020) so as to find the optimal shots to use.

Larger Language Models In this paper, we experimented with the largest available (Sept 2021) unidirectional LM, GPT-J (6B), and we already achieved competitive performance with respect to full-shot models. As shown in the results section, larger models achieve consistently better results, and thus exploring larger models on dialogue tasks could lead to straightforward improvement in the performance. OpenAI and AI21 Labs provide API interfaces to interact with the largest existing LMs, GPT-3 (175B) and GPT-Jurassic (176B), respectively, but OpenAI has a long waiting list to access the model, and AI21 provides access to only 10K tokens per-day, thus making it impossible to use it for large-scale experiments.

4.2 Human Evaluation

In this paper, we conduct only an extensive automatic evaluation in all the dialogue tasks. In response-generation tasks and end-to-end chat-bots (FSB), a proper human evaluation is required. Unfortunately, due to budget restrictions, we do not perform any human study, and thus greatly limiting the reliability of the results. In future work, we plan to perform, at least, an interactive human evaluation of the FSB, so to compare it with SOTA chat-bots such as Meena (Adiwardana et al., 2020) and BST (Roller et al., 2021).

4.3 Future Work

In this paper, we explore a good variety of dialogue-related tasks, including both response generation and conversational parsing. Although we explore task-oriented dialogue systems (e.g., SMD and MWoZ), there are many aspects and datasets we did not explore in this paper. For instance, it would be interesting to explore more datasets and tasks such as conversational semantic parsing, e.g., CoSQL (Yu et al., 2019), data-flow (Andreas et al., 2020), treeDST (Cheng et al., 2020), multi-domain DST, e.g., SGD (Rastogi et al., 2020), MWoZ (since in this paper we only use a single domain), BiToD (Lin et al., 2021d), STAR (Mosig et al., 2020) and task-master (Byrne et al., 2019), multi-modal interactive datasets, e.g.,

SIMMC (Crook et al., 2021; Kottur et al., 2021), and continual learning (Madotto et al., 2020d). Moreover, there are more dialogue datasets in open domain chat that we did not explore, such as Topical Chat (Gopalakrishnan et al., 2019), multi-knowledge grounded models (Eric et al., 2021), multilingual personalized dialogue (Lin et al., 2020a), and blending multiple dialogue skills (Smith et al., 2020b; Sun et al., 2021; Young et al., 2021), as in Madotto et al. (2020c).

5. Related Work

Prompt-based learning has been successfully applied to various NLP applications, including machine translation, summarization (Radford et al., 2019), question answering (Petroni et al., 2019; Jiang et al., 2020), fact checking (Lee et al., 2021), text classification (Schick and Schütze, 2021a,b), relation extraction (Gao et al., 2020b; Chen et al., 2021), and multi-modal learning (Tsimploukelli et al., 2021). Cahyawijaya et al. (2021) explored matrix factorization (Winata et al., 2019, 2020) to accelerate the prompt-based inference. As with every machine learning technique, prompt-based learning comes with pros and cons. The main advantages are 1) state-of-the-art few-shot learning performance, 2) no model fine-tuning, and thus no-gradient involved (no over/under-fitting, etc.), and 3) single model deployment. On the other hand, the main disadvantages are 1) the hard limit in number of shots due to the maximum input length of LMs, 2) instability (e.g., sensitive to the order of the examples), and 3) challenge of prompt engineering. Some of these disadvantages have been partially solved by calibrating the LM (Zhao et al., 2021), learning the optimal order of prompts (Kumar and Talukdar, 2021; Lu et al., 2021), automatically selecting prompts (Shin et al., 2020) or by using embedding-based prompts (Li and Liang, 2021). However, many challenges remain, and interested readers can refer to the comprehensive summary paper from Liu et al. (2021).

Prompt-Based Learning in Dialogue Systems Prompt-based learning has also been studied in many downstream dialogue tasks. Lin et al. (2021a) incorporated slot description prompts into T5 (Raffel et al., 2020) for zero-shot DST (Lin et al., 2021b), and in natural language understanding (NLU) (Namazifar et al., 2021). Mi et al. (2021) extended this approach to few-shot NLU and natural language generation (NLG). However, their approaches include a fine-tuning step and cannot preserve the original weights of the LMs. A concurrent work by Zheng and Huang (2021) explored prompt-based few-shot learning for knowledge grounded dialogue response generation (e.g., PersonaChat, WoW, etc.). The authors (Zheng and Huang, 2021) explore training based prompt learning (Schick and Schütze, 2021a) and continuous-prompts (Li and Liang, 2021), which both requires gradient updates. In contrast, Madotto et al. (2020e) explored gradient-free prompt-based learning for NLU, DST, and NLG and showed promising results. This approach has been further extended to a multilingual setting by Winata et al. (2021). However, the existing studies of prompt-based learning for dialogue systems have been focused on goal-oriented dialogue tasks. Differently, this work provides a comprehensive study of prompt-based learning for both task-oriented dialogue and open-domain chit-chat. Concurrently with our work, Zheng and Huang (2021) explored prompt-based few-shot learning for grounded dialogue response generation (e.g., PersonaChat, WoW, etc.). Differently from this paper, which only uses prompts, the au-

thors explored training-based prompt learning (Schick and Schütze, 2021a), and continuous prompts (Li and Liang, 2021), which both require some sort of gradient steps.

6. Conclusion

In this paper, we explore prompt-based few-shot learning in dialogue tasks. We benchmark LMs of different sizes in nine response generation tasks, which include a variety of knowledge-grounded tasks, task-oriented generations, general open-chat, and controlled stylistic generation, and five conversational parsing tasks, which include DST, graph path generation, persona information extraction, and document retrieval. We propose a novel prompt-based few-shot classifier, that also does not require any fine-tuning, to select the most appropriate prompt given a dialogue history. The results suggest that prompt-based few-shot learning is a promising technique for dialogue systems. In particular, the larger the LM the better the performance, suggesting future improvements with larger LMs (e.g., GPT3). Finally, by combining the power of prompt-based few-shot learning and a prompt selector, we create an end-to-end chatbot named the **Few-Shot Bot** (FSB), which automatically selects the most appropriate conversational skill (task-oriented, chit-chat, etc.), retrieve knowledge from different sources (i.e., Wikipedia, Internet, KGs), and uses it to generate a human-like response, all by using only few dialogue examples per skill.

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. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.
- Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, et al. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8: 556–571, 2020.
- Yejin Bang, Nayeon Lee, Etsuko Ishii, Andrea Madotto, and Pascale Fung. Assessing political prudence of open-domain chatbots. *arXiv preprint arXiv:2106.06157*, 2021.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.
- Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow, 2021. URL <http://github.com/eleutherai/gpt-neo>.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models

are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf>.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 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, 2018.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. Taskmaster-1: Toward a realistic and diverse dialog dataset. *arXiv preprint arXiv:1909.05358*, 2019.

Samuel Cahyawijaya, Genta Indra Winata, Holy Lovenia, Bryan Wilie, Wenliang Dai, Etsuko Ishii, and Pascale Fung. Greenformer: Factorization toolkit for efficient deep neural networks. *arXiv preprint arXiv:2109.06762*, 2021.

Xiang Chen, Xin Xie, Ningyu Zhang, Jiahuan Yan, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Adaprompt: Adaptive prompt-based finetuning for relation extraction. *arXiv preprint arXiv:2104.07650*, 2021.

Jianpeng Cheng, Devang Agrawal, Héctor Martínez Alonso, Shruti Bhargava, Joris Driesen, Federico Flego, Dain Kaplan, Dimitri Kartsaklis, Lin Li, Dhivya Piraviperumal, et al. Conversational semantic parsing for dialog state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8107–8117, 2020.

Paul A Crook, Satwik Kottur, Seungwhan Moon, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. Situated interactive multimodal conversations (simmc) track at dstc9. In *AAAI DSTC9 Workshop*, 2021.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. In *International Conference on Learning Representations*, 2020.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*, 2018.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 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, 2019.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston.

Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, 2020.

Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*, 2021.

Wanyu Du and Yangfeng Ji. Sidecontrol: Controlled open-domain dialogue generation via additive side networks. *arXiv preprint arXiv:2109.01958*, 2021.

Esin Durmus, He He, and Mona Diab. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, 2020.

Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*, 2021.

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

Mihail Eric, Nicole Chartier, Behnam Hedayatnia, Karthik Gopalakrishnan, Pankaj Rajan, Yang Liu, and Dilek Hakkani-Tur. Multi-sentence knowledge selection in open-domain dialogue. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 76–86, 2021.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, 2018.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020a.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020b.

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

Tianxing He, Jun Liu, Kyunghyun Cho, Myle Ott, Bing Liu, James Glass, and Fuchun Peng. Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1121–1133, 2021.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

Jaehun Jung, Bokyung Son, and Sungwon Lyu. Attnio: Knowledge graph exploration with in-and-out attention flow for knowledge-grounded dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3484–3497, 2020.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Hyunwoong Ko. Parallelformers: An efficient model parallelization toolkit for deployment. <https://github.com/tunib-ai/parallelformers>, 2021.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*, 2021.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *arXiv preprint arXiv:2104.08667*, 2021.

Sawan Kumar and Partha Talukdar. Reordering examples helps during priming-based few-shot learning. *arXiv preprint arXiv:2106.01751*, 2021.

Nayeon Lee, Andrea Madotto, and Pascale Fung. Exploring social bias in chatbots using stereotype knowledge. In *WNLP@ ACL*, pages 177–180, 2019.

Nayeon Lee, Yejin Bang, Andrea Madotto, Madian Khabsa, and Pascale Fung. Towards few-shot fact-checking via perplexity. *arXiv preprint arXiv:2103.09535*, 2021.

- Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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, 2020a.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020b.
- Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 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, 2017.
- Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*, 2021.
- Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. Xpersona: Evaluating multilingual personalized chatbot. *arXiv preprint arXiv:2003.07568*, 2020a.
- Zhaojiang Lin, Andrea Madotto, and Pascale Fung. Exploring versatile generative language model via parameter-efficient transfer learning. *arXiv preprint arXiv:2004.03829*, 2020b.
- Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 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, 2020c.
- Zhaojiang Lin, Bing Liu, Andrea Madotto, Seungwhan Moon, Paul Crook, Zhenpeng Zhou, Zhiguang Wang, Zhou Yu, Eunjoon Cho, Rajen Subba, et al. Zero-shot dialogue state tracking via cross-task transfer. *arXiv preprint arXiv:2109.04655*, 2021a.
- Zhaojiang Lin, Bing Liu, Seungwhan Moon, Paul A Crook, Zhenpeng Zhou, Zhiguang Wang, Zhou Yu, Andrea Madotto, Eunjoon Cho, and Rajen Subba. Leveraging slot descriptions for zero-shot cross-domain dialogue statetracking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5640–5648, 2021b.
- Zhaojiang Lin, Andrea Madotto, Yejin Bang, and Pascale Fung. The adapter-bot: All-in-one controllable conversational model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 16081–16083, 2021c.

- Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. *arXiv preprint arXiv:2106.02787*, 2021d.
- Zhiyu Lin and Mark O Riedl. Plug-and-blend: A framework for controllable story generation with blended control codes. *NAACL HLT 2021*, page 62, 2021.
- Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. Does gender matter? towards fairness in dialogue systems. *arXiv preprint arXiv:1910.10486*, 2019a.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b.
- Robert L Logan IV, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. Cutting down on prompts and parameters: Simple few-shot learning with language models. *arXiv preprint arXiv:2106.13353*, 2021.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23, 2019.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.
- Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478, 2018.
- Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. Learning knowledge bases with parameters for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2372–2394, 2020a.
- Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. Plug-and-play conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2422–2433, 2020b.
- Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, Jamin Shin, and Pascale Fung. Attention over parameters for dialogue systems. *arXiv preprint arXiv:2001.01871*, 2020c.
- Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, and Zhiguang Wang. Continual learning in task-oriented dialogue systems. *arXiv preprint arXiv:2012.15504*, 2020d.

- Andrea Madotto, Zihan Liu, Zhaojiang Lin, and Pascale Fung. Language models as few-shot learner for task-oriented dialogue systems. *arXiv preprint arXiv:2008.06239*, 2020e.
- Fei Mi, Yitong Li, Yasheng Wang, Xin Jiang, and Qun Liu. Cins: Comprehensive instruction for few-shot learning in task-oriented dialog systems. *arXiv preprint arXiv:2109.04645*, 2021.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. Reframing instructional prompts to gptk’s language. *arXiv preprint arXiv:2109.07830*, 2021.
- Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 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, 2019.
- Johannes EM Mosig, Shikib Mehri, and Thomas Kober. Star: A schema-guided dialog dataset for transfer learning. *arXiv preprint arXiv:2010.11853*, 2020.
- Mahdi Namazifar, Alexandros Papangelis, Gokhan Tur, and Dilek Hakkani-Tür. Language model is all you need: Natural language understanding as question answering. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7803–7807. IEEE, 2021.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.
- Prasanna Parthasarathi and Joelle Pineau. Extending neural generative conversational model using external knowledge sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695, 2018.
- F Petroni, A Piktus, A Fan, PSH Lewis, M Yazdani, ND Cao, J Thorne, Y Jernite, V Karpukhin, J Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. In *NAACL-HLT*, pages 2523–2544. Association for Computational Linguistics, 2021.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? 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 2463–2473, 2019.
- Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. Entity-consistent end-to-end task-oriented dialogue system with kb retriever. *arXiv preprint arXiv:1909.06762*, 2019.
- Libo Qin, Xiao Xu, Wanxiang Che, Yue Zhang, and Ting Liu. Dynamic fusion network for multi-domain end-to-end task-oriented dialog. *arXiv preprint arXiv:2004.11019*, 2020.

- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 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, 2019.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 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, 2020.
- Revanth Gangi Reddy, Danish Contractor, Dinesh Raghu, and Sachindra Joshi. Multi-level memory for task oriented dialogs. 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 3744–3754, 2019.
- Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 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, 2021.
- Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, 2021a.
- Timo Schick and Hinrich Schütze. It’s not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, 2021b.
- Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Eliciting knowledge from language models using automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, 2020.
- Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, 2020a.

- Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2453–2470, 2020b.
- Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. Multi-modal open-domain dialogue. *arXiv preprint arXiv:2010.01082*, 2020c.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.
- Eric Michael Smith, Diana Gonzalez-Rico, Emily Dinan, and Y-Lan Boureau. Controlling style in generated dialogue. *arXiv preprint arXiv:2009.10855*, 2020a.
- Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents’ ability to blend skills. *arXiv preprint arXiv:2004.08449*, 2020b.
- Kai Sun, Seungwhan Moon, Paul A Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1570–1583, 2021.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
- Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *arXiv preprint arXiv:2106.13884*, 2021.
- Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.
- Zhilin Wang, Xuhui Zhou, Rik Koncel-Kedziorski, Alex Marin, and Fei Xia. Extracting and inferring personal attributes from dialogue. *arXiv preprint arXiv:2109.12702*, 2021.
- Genta Indra Winata, Andrea Madotto, Jamin Shin, Elham J Barezi, and Pascale Fung. On the effectiveness of low-rank matrix factorization for lstm model compression. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*, pages 253–262. Waseda Institute for the Study of Language and Information, 2019.
- Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, and Pascale Fung. Lightweight and efficient end-to-end speech recognition using low-rank transformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6144–6148. IEEE, 2020.

- Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. Language models are few-shot multilingual learners. *arXiv preprint arXiv:2109.07684*, 2021.
- Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, 2019a.
- Chien-Sheng Wu, Andrea Madotto, Zhaojiang Lin, Peng Xu, and Pascale Fung. Getting to know you: User attribute extraction from dialogues. *arXiv preprint arXiv:1908.04621*, 2019b.
- Chien-Sheng Wu, Richard Socher, and Caiming Xiong. Global-to-local memory pointer networks for task-oriented dialogue. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019c.
- Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 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, 2021a.
- Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*, 2021b.
- Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- Tom Young, Frank Xing, Vlad Padelea, Jinjie Ni, and Erik Cambria. Fusing task-oriented and open-domain dialogues in conversational agents. *arXiv preprint arXiv:2109.04137*, 2021.
- Tao Yu, Rui Zhang, Heyang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, et al. Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases. In *EMNLP/IJCNLP (1)*, 2019.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 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, 2018.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 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, 2020.

Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*, 2021.

Chujie Zheng and Minlie Huang. Exploring prompt-based few-shot learning for grounded dialog generation. *arXiv preprint arXiv:2109.06513*, 2021.

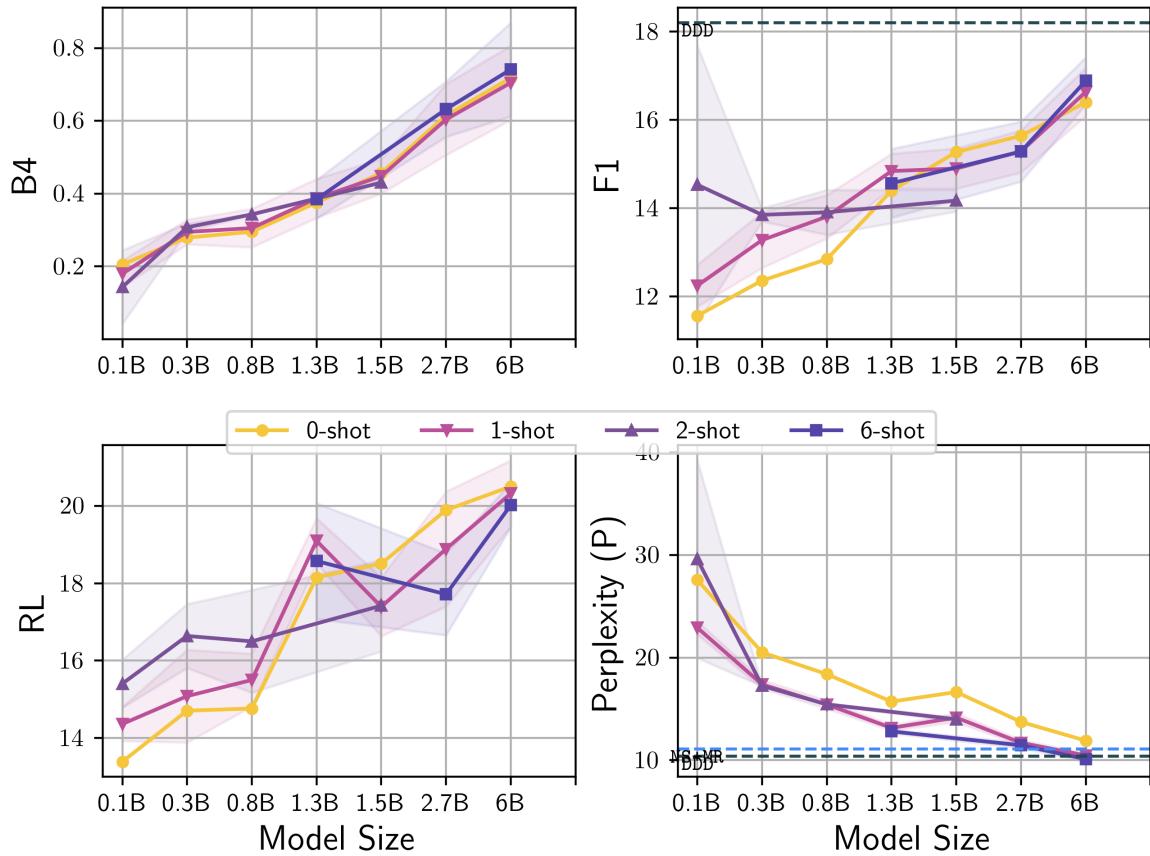


Figure 7: Model size vs metrics in the DailyDialogue (DD) dataset.

Appendix A.

Dataset	Response Gen.	Conv. Parsing	Dataset	Response Gen.
<u>WoW</u>	Dialogue: User: [U ₁] KB: [K ₁] Wizard: [S ₁] ... User: [U _t] KB: [K _t] Wizard: [S _t]	Dialogue: User: [U ₁] Assistant: [S ₁] ... Assistant: [S _{t-1}] User: [U _t] Search: [Q]	<u>Persona</u>	Persona Info.: [P ₁] ... [P _n] Dialogue: User: [U ₁] Persona: [S ₁] ... User: [U _t] Persona: [S _t]
<u>WIT</u>	Assistant Info: [P ₁] ... [P _n] Dialogue: User: [U ₁] KB: [K ₁] Assistant: [S ₁] ... User: [U _t] KB: [K _t] Assistant: [S _t]	Assistant Info: [P ₁] ... [P _n] Dialogue: User: [U ₁] Assistant: [S ₁] Search: [Q ₁] ... Assistant: [S _{t-1}] User: [U _t] Search: [Q _t]	<u>ED</u>	Dialogue: User: [U ₁] Empath: [S ₁] ... User: [U _t] Empath: [S _t]
<u>MSC</u>	User Info: [P ₁] ... [P _n] Assistant Info: [P ₁] ... [P _n] Dialogue: User: [U ₁] Assistant: [S ₁] ... User: [U _t] Assistant: [S _t]	Dialogue: User: [U ₁] Assistant: [S ₁] Write: [P ₁] ... Assistant: [S _{t-1}] User: [U _t] Write: [P _t]	<u>DD</u>	Dialogue: UserA: [U ₁] UserB: [S ₁] ... UserA: [U _t] UserB: [S _t]
<u>DialKG</u>	Dialogue: User: [U ₁] KG: [T ₁] Assistant: [S ₁] ... User: [U _t] KG: [T _t] Assistant: [S _t]	Dialogue: User: [U ₁] KG: [T ₁] Assistant: [S ₁] ... Assistant: [S _{t-1}] User: [U _t] KG: [T _t]	<u>IC</u>	Image: [D] Dialogue: [StyleA]: [U ₁] [StyleB]: [S ₁] [StyleA]: [U ₂]
<u>MWoZ</u>	-	Dialogue: User: [U ₁] Assistant: [S ₁] DST: [DS ₁] ... Assistant: [S _{t-1}] User: [U _t] DST: [DS _t]	<u>CG (IC)</u>	Dialogue: System: [U ₁] User: [S ₁] [Style]: [U ₂]
			<u>SMD</u>	KB: [T ₁] ... [T _n] Dialogue: User: [U ₁] Assistant: [S ₁] ... User: [U _t] Assistant: [S _t]

Table 14: Summary of the prompt templates for response generation and conversational parsing. The highlighted text [] refers to what the LM generates. In DD and IC the model predict all the turns, since these are human-to-human dialogues.

DD

Template	Prompt
Dialogue:	Dialogue:
UserA: [U ₁]	UserA: Hey man, you wanna buy some weed?
UserB: [S ₁]	UserB: Some what?
UserA: [U ₂]	UserA: Weed! You know? Pot, Ganja, Mary Jane some chronic!
UserB:	UserB:
Generation	
[S ₂]	Oh , umm , no thanks .

Table 15: Prompt and example for DailyDialogue (DD).

ED

Template	Prompt
Dialogue:	Dialogue:
User: [U ₁]	User: Yeah about 10 years ago I had a horrifying experience. It was 100% their fault but they hit the water barrels and survived. They had no injuries but they almost ran me off the road.
Empath: [S ₁]	Empath: Did you suffer any injuries?
User: [U ₂]	User: No I wasn't hit. It turned out they were drunk. I felt guilty but realized it was his fault.
Empath:	Empath:
Generation	
[S ₂]	Why did you feel guilty? People really shouldn't drive drunk.

Table 16: Prompt and example for Empathetic Dialogues (ED).

PersonaChat

Template	Prompt
Persona Information:	Persona Information: [P ₁] i read twenty books a year. [P ₂] i'm a stunt double as my second job. [P ₃] i only eat kosher. [P ₄] i was raised in a single parent household.
Dialogue:	Dialogue:
User: [U ₁]	User: hello what are doing today ?
Persona: [S ₁]	Persona: i am good , i just got off work and tired , i have two jobs .
User: [U ₂]	User: i just got done watching a horror movie
Persona:	Persona:
Generation	
[S ₂]	i am good, i just got off work and tired, i have two jobs.

Table 17: Prompt and example for Persona Chat.

WoW

Template	Context
Dialogue: User: [U ₁] KB: [K ₁]	Dialogue: User: Blue is always nice. I like royal blue. KB: The Royal Blue was the Baltimore and Ohio Railroad (B&O)'s flagship passenger train between New York City and Washington, D.C., in the United States, beginning in 1890.
Assistant: [S ₁] User: [U ₂] KB: [K ₂]	Assistant: I once road on The Royal Blue train from New York to D.C User: Oh that sounds really nice. I bet there was a lot of scenery and blue skies. KB: Blue Skies is a 1946 American musical comedy film directed by Stuart Heisler and starring Bing Crosby, Fred Astaire, and Joan Caulfield.
Assistant:	Assistant:
Generation	
[S ₂]	Yes, speaking of Blue Skies, have you seen the 1946 movie staring Bing Crosby?

Table 18: Prompt and example for Wizard of Wikipedia (WoW).

WiT

Template	Context
Assistant Information: [P ₁] [P ₂]	Assistant Information: My favorite sports team is NY Giants. When they are winning...
Dialogue: User: [U ₁] KB: [K ₁] Assistant: [S ₁] User: [U ₂] KB: [K ₂]	Dialogue: User: \emptyset KB: \emptyset Assistant: What do you think about Kenny Golladay for the Giants? User: He is pretty good. I think he is a pretty good fit for the team. KB: Golladay is under evaluation for a concussion during Sunday's game versus the Packers.
Assistant:	Assistant:
Generation	
[S ₂]	He's kind of had a lot of head injuries though. Do you think that's a concern?

Table 19: Prompt and example for Wizard of Internet (WiT).

IC

Template	Prompt	Original Image
Image: [D] Dialogue: [StyleA]: [U ₁] [StyleB]: [S ₁] [StyleA]:	Image: a group of horses standing next to each other Dialogue: Considerate: The loads on their neck are too much Charming: On my next date I am totally taking out my date on beautiful horses like these Considerate:	
Generation		
[U ₂]	A horse ride can be quite romantic.	

Table 20: Prompt and example for Image Chat (IC).

SMD

Template	Prompt
<p>KB:</p> <pre>chevron distance 5_miles chevron traffic_info moderate_traffic chevron poi_type gas_station chevron address 783_arcadia_pl town_and_country distance 5_miles town_and_country traffic_info no_traffic town_and_country poi_type shopping_center town_and_country address 383_university_ave jacks_house distance 5_miles jacks_house traffic_info no_traffic jacks_house poi_type friends_house jacks_house address 864_almanor_ln home distance 6_miles home traffic_info heavy_traffic home poi_type home home address 5671_barringer_street the_clement_hotel distance 4_miles the_clement_hotel traffic_info no_traffic the_clement_hotel poi_type rest_stop the_clement_hotel address 657_ames_ave sigona_farmers_market distance 1_miles sigona_farmers_market traffic_info heavy_traffic sigona_farmers_market poi_type grocery_store sigona_farmers_market address 638_amherst_st tai_pan distance 6_miles tai_pan traffic_info no_traffic tai_pan poi_type chinese_restaurant tai_pan address 830_almanor_ln</pre> <p>Dialogue:</p> <p>User: what gas_station are here ?</p> <p>Assistant: there is a chevron</p> <p>User: that s good ! please pick the quickest route to get there and avoid all heavy_traffic !</p> <p>Assistant:</p>	
Generation	
[S ₂]	taking you to chevron

Table 21: Prompt and example for Standford Multidomain Dataset (SMD).

MSC

Template	Prompt
User Information: [P ₁] [P ₂] [P ₃] [P ₄]	User Information: I own a Jeep. I enjoy exercising at the gym. I have a marketing job. I don't eat meat.
Assistant Information: [P ₁] [P ₂] [P ₃] [P ₄]	Assistant Information: I don't own a car. I'm a computer programmer. I like grilling steak. I enjoy running and walking.
Dialogue: User: [U ₁] Assistant: [S ₁] User: [U ₂] Assistant:	Dialogue: User: Did you manage to go out on a run today? Assistant: Yes I actually was able too. I am considering joining the local gym. Do you prefer going to the gym? User: I do actually. I like the controlled environment. I don't want to have to depend on the weather consider where I live. Assistant:
Generation [S ₂]	I hate to have to run when its raining.

Table 22: Prompt and example for Multi-Session Chat (MSC).

CG(IC)

Template	Prompt
Dialogue: System: [U ₁] User: [S ₁] [Style]:	Dialogue: System: What a sweet sky. User: Beautiful day out in the cold Kind:
Generation [U ₂]	It makes me incredibly peaceful to see that sky

Table 23: Prompt and example for Controlled Generation (CG-IC).

DialKG

Template	Context
Dialogue:	Dialogue:
User: [U ₁]	User: I like Anita Diamant. Could you recommend anything by her?
KG: [T ₁]	KG: Anita Diamant \t ~written_by \t The Red Tent
Assistant: [S ₁]	Assistant: She wrote The Red Tent and The Last Days of Dogtown, both great novels.
User: [U ₂]	User: They were great. I read both of those. Red Tent is historical fiction. Do you know if she has any other books like that?
KG: [T ₂]	KG: Historical fiction \t ~has_genre \t The Weight of Water
Assistant:	Assistant:
Generation	
[S ₂]	The Weight of Water is a great book but it is not written by Anita Diamant.

Table 24: Prompt and example for DialKG.

WoW-Parse

Template	Prompt
Dialogue:	Dialogue:
User: [U ₁]	User: I go to Target all the time, do you?
Assistant: [S ₁]	Assistant: Target Corporation is the second largest discount retailer
User: [U ₂]	User: I assume that it is just before Walmart, but still huge.
Search:	Search:
Generation	
[Q]	Target Corporation

Table 25: Prompt and example for Wizard of Wikipedia Parse (WoW-Parse).

WiT-parse

Template	Context
Assistant Information: [P ₁] [P ₂]	Assistant Information: My favorite sports team is NY Giants. When they are winning...
Dialogue: User: [U ₁] Search: [Q ₁] Assistant: [S ₁] User: [U ₂] Search:	Dialogue: User: \t Search: Kenny Golladay Giants Assistant: What do you think about Kenny Golladay for the Giants? User: He is pretty good. I think he is a pretty good fit for the team. Search:
Generation	
[Q ₂]	Kenny Golladay Rumors: Giant

Table 26: Prompt and example for Wizard of Internet Parse (WiT-parse).

MSC-parse

Template		Prompt
Dialogue:		Dialogue:
User: [U ₁]		User: What kind of car do you own? I have a jeep.
Write: [P ₁]		Write: I own a Jeep.
Assistant: [S ₁]		Assistant: I don't own my own car! I actually really enjoy enjoying walking and running, but then again, I live in a small town and semi-close to work.
User: [U ₂]		User: Ah I see! I like going to the gym to work out.
Write:		Write:
Generation		
[P ₂]		I enjoy exercising at the gym.

Table 27: Prompt and example for Multi-Session Chat (MSC) parse.

DialKG-Parse

Template		Context
Dialogue:		Dialogue:
User: [U ₁]		User: I like Anita Diamant. Could you recommend anything by her?
KG: [T ₁]		KG: Anita Diamant \t ~written_by \t The Red Tent
Assistant: [S ₁]		Assistant: She wrote The Red Tent and The Last Days of Dogtown, both great novels.
User: [U ₂]		User: They were great. I read both of those. Red Tent is historical fiction. Do you know if she has any other books like that?
KG:		KG:
Generation		
[T ₂]		Historical fiction \t ~has_genre \t The Weight of Water

Table 28: Prompt and example for DialKG parse.

MWoZ DST (Single)

Template		Context
Dialogue:		Dialogue:
User: [U ₁]		User: i am looking for a hotel with at least a 3 star rating.
DST: [DS ₁]		DST hotel-stars=3
Assistant: [S ₁]		Assistant: we have 6 hotel -s and guesthouses that have 3 star rating -s . is there a particular part of town , or a particular price range that you are interested in ?
User: [U ₂]		User: i am not to worried about price but the hotel needs to have free wifi.
DS:		DST:
Generation		
[DS ₂]		hotel-internet=yes

Table 29: Prompt and example for MWoZ DST.

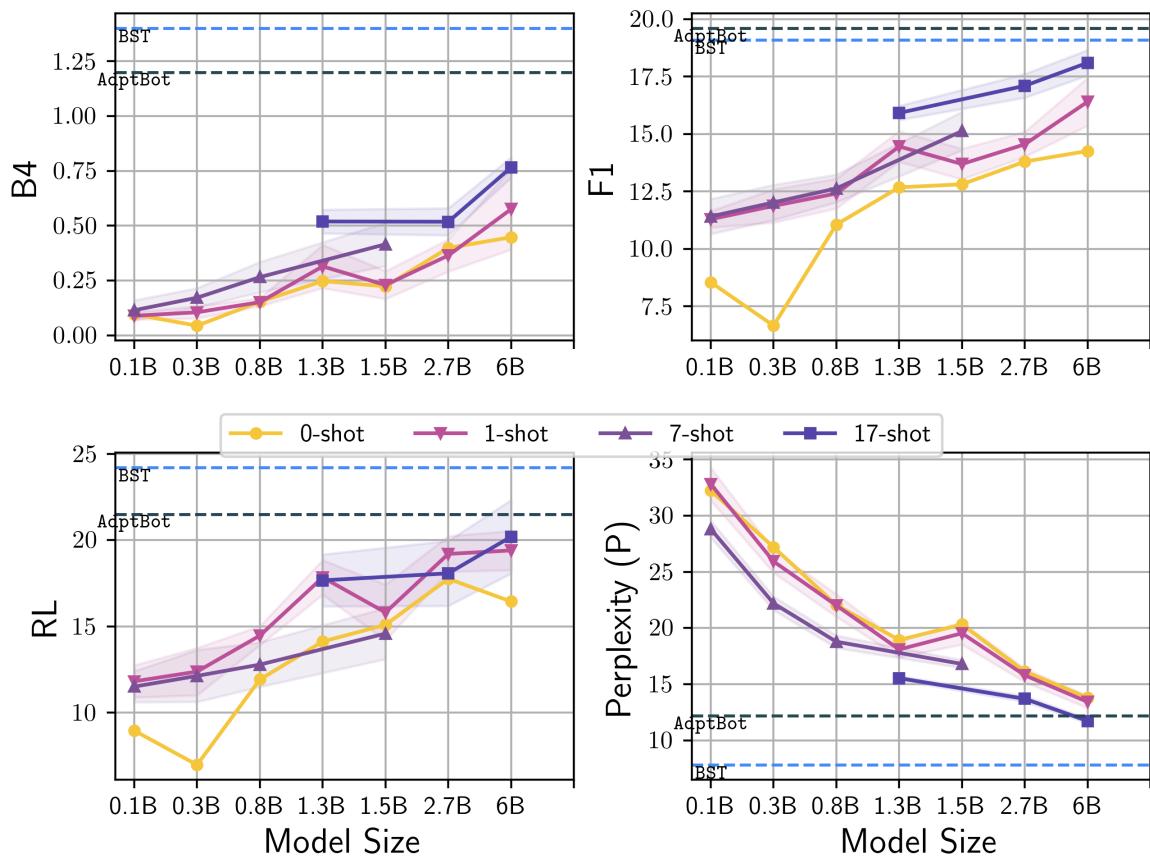


Figure 8: Model size vs metrics in the Empathetic Dialogues (ED) dataset.

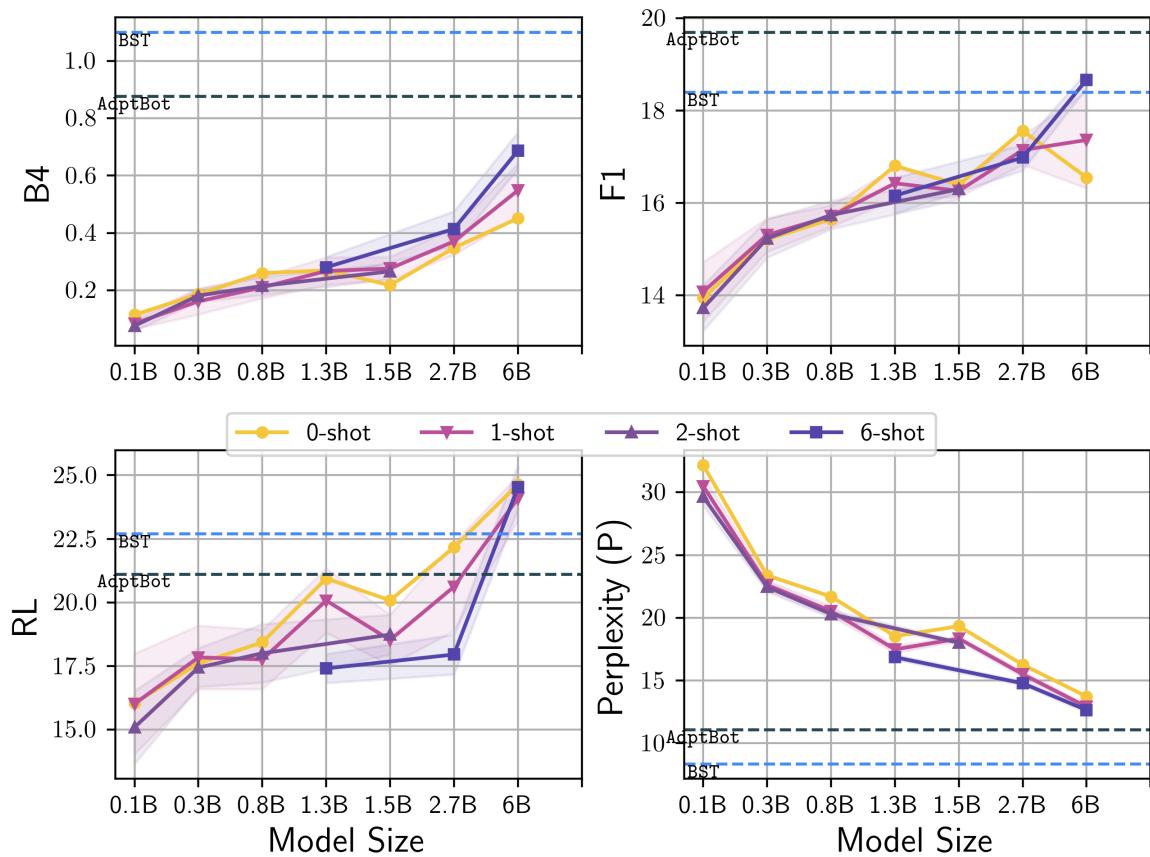


Figure 9: Model size vs metrics in the Persona Chat dataset.

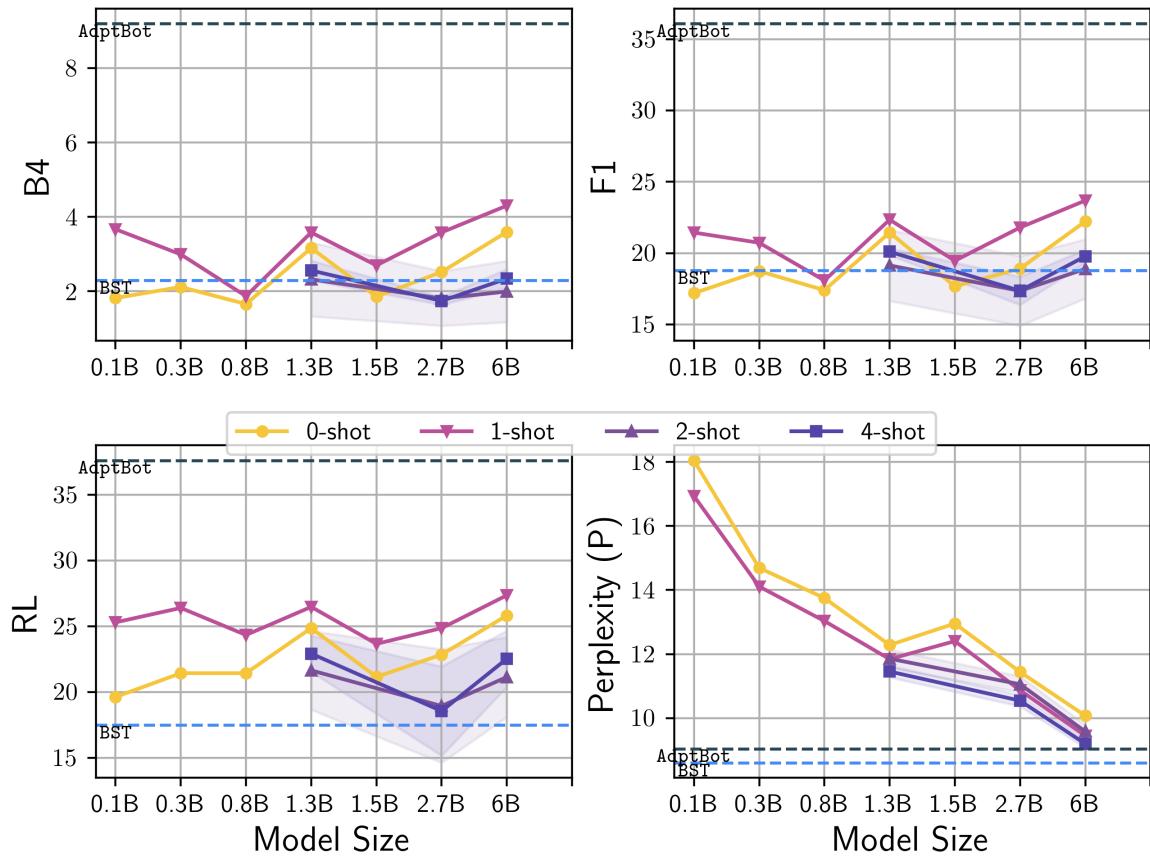


Figure 10: Model size vs metrics in the Wizard of Wikipedia (WoW) dataset.

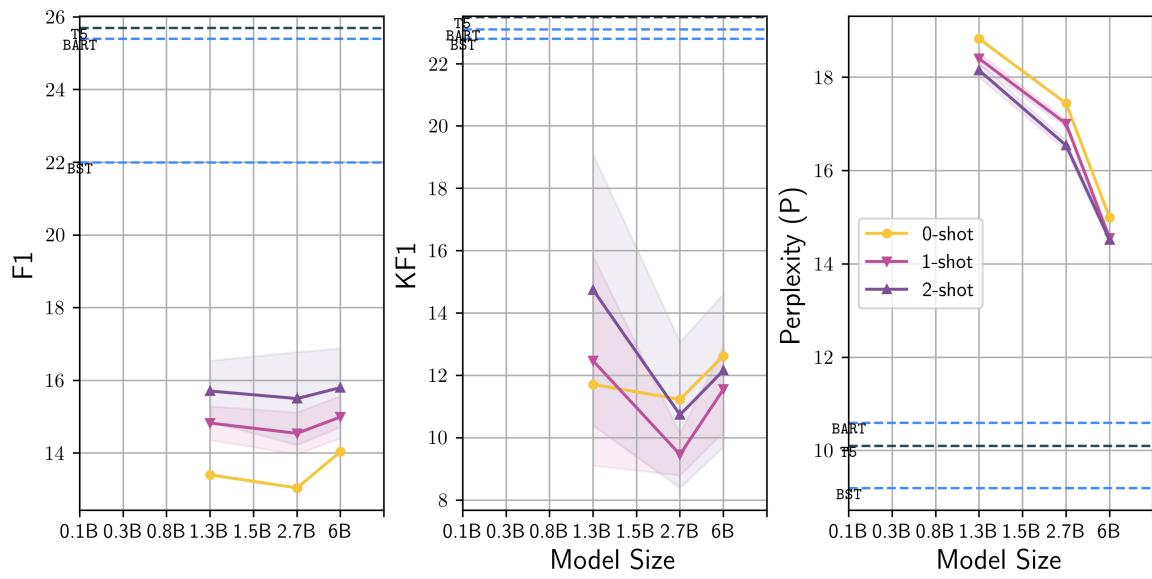


Figure 11: Model size vs metrics in the Wizard of Internet (WiT) dataset.

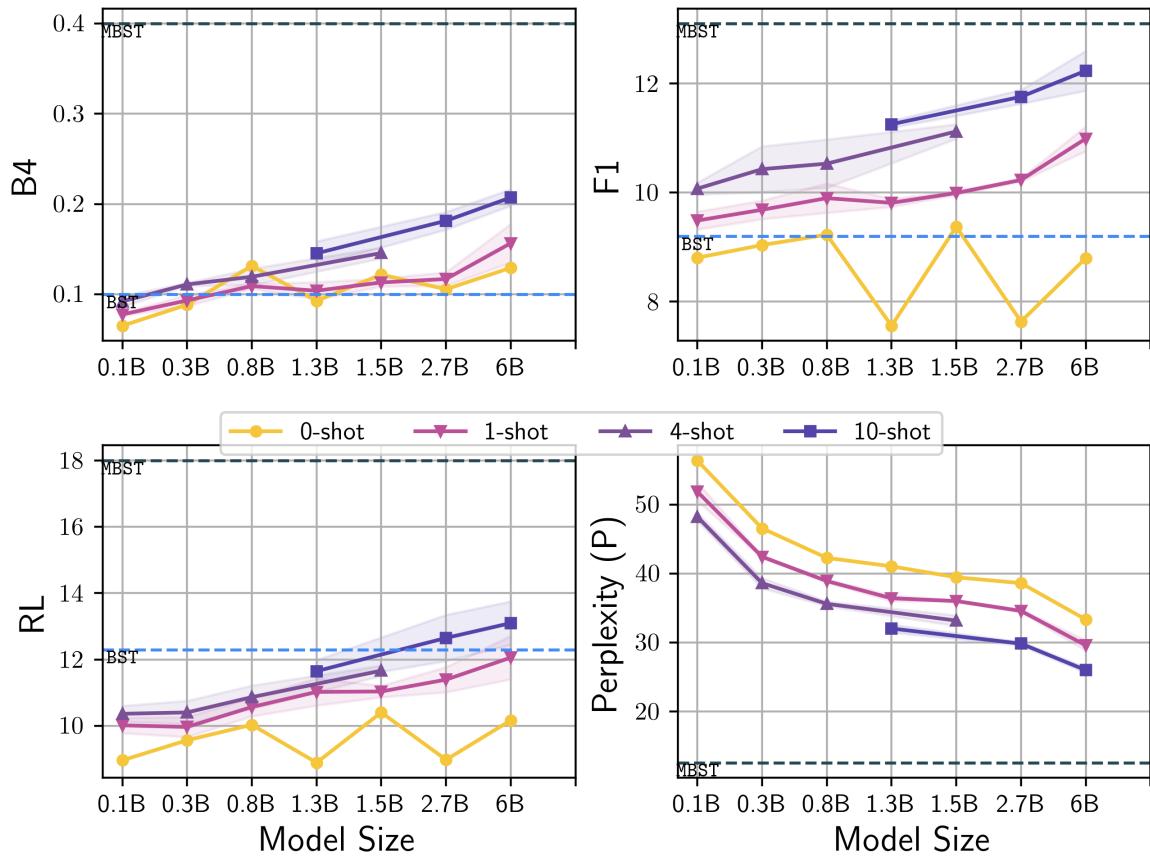


Figure 12: Model size vs metrics in the Image Chat (IC) dataset.

FEW-SHOT BOT

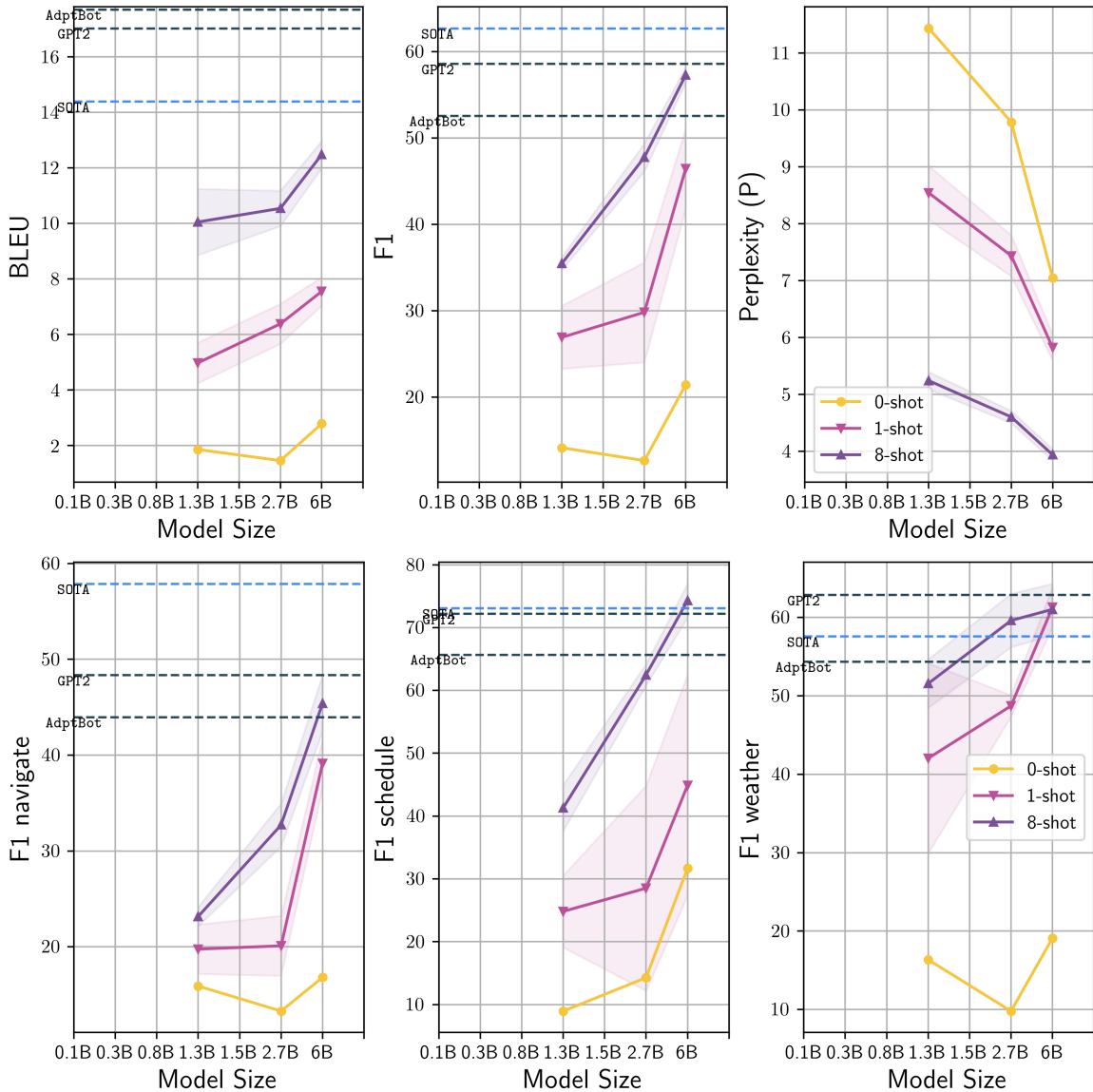


Figure 13: Model size vs metrics in the Stanford Multidomain Dialogue (SMD) dataset.

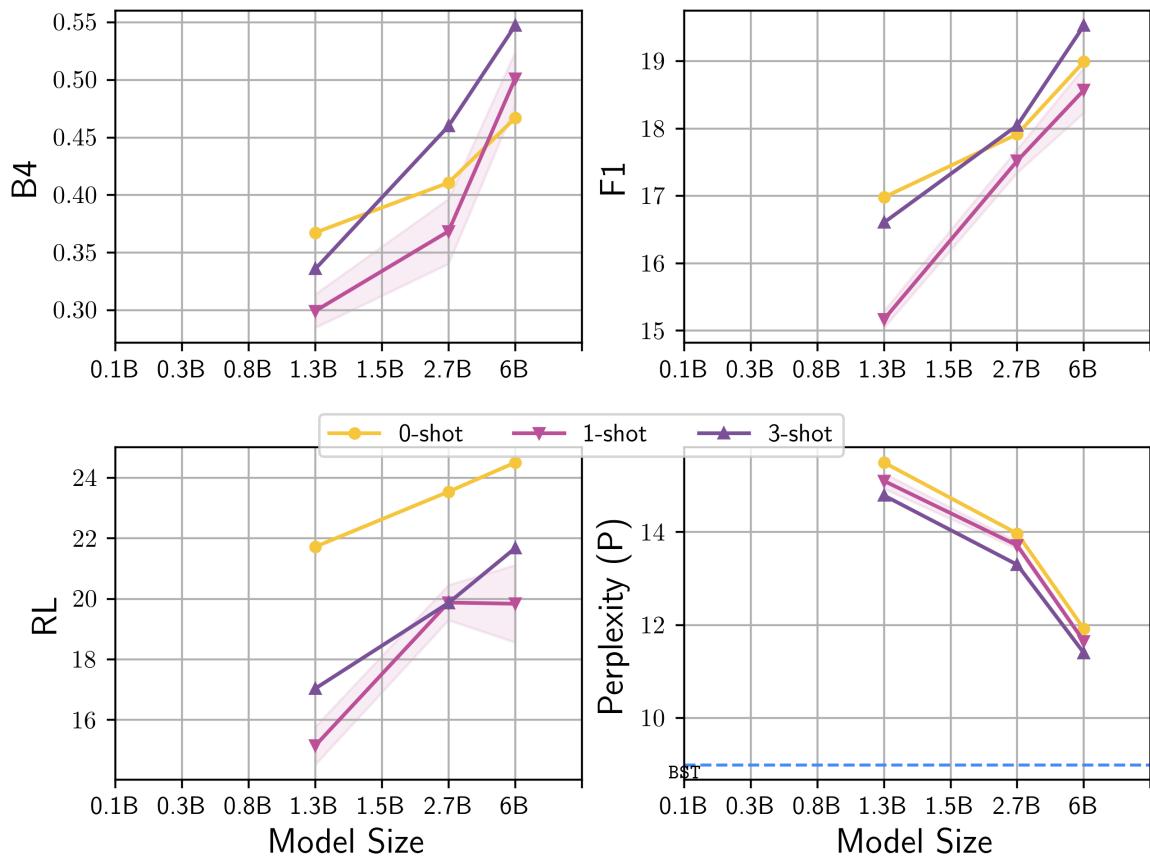


Figure 14: Model size vs metrics in the Multi-Session Chat (MSC) dataset.

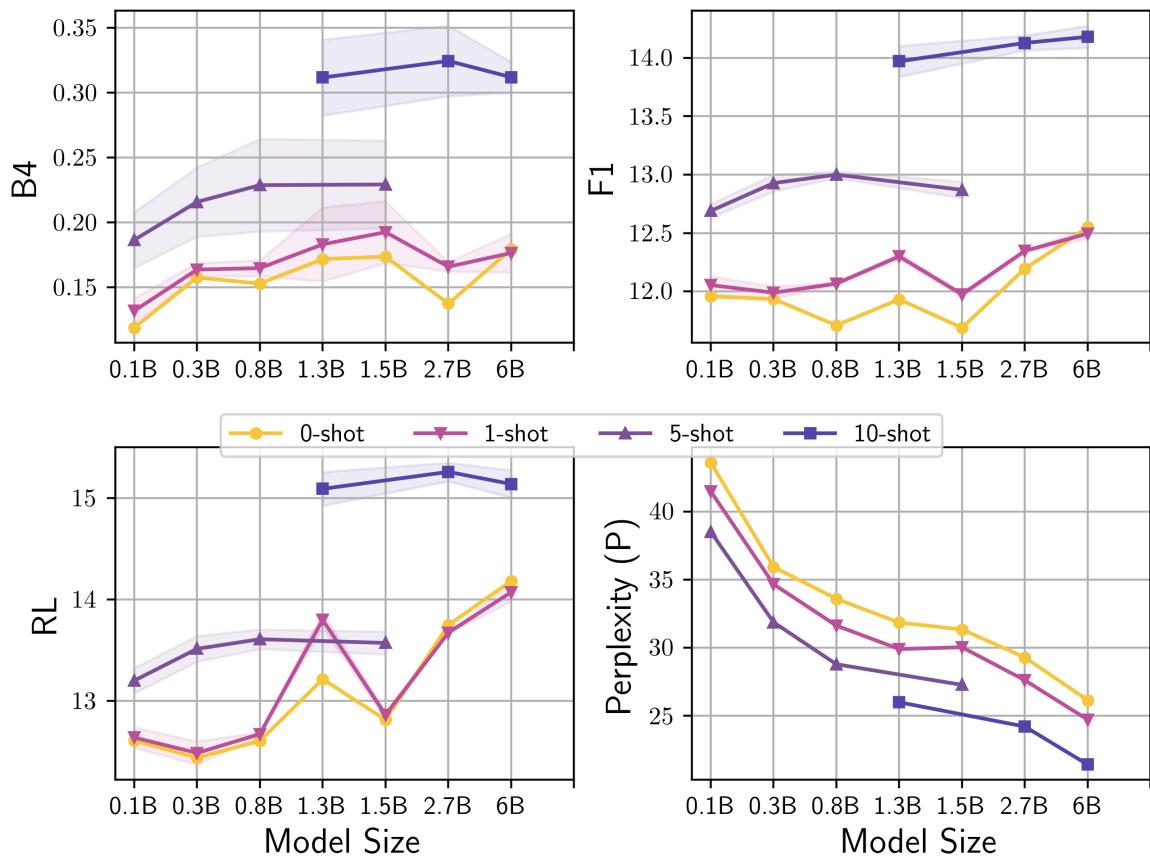


Figure 15: Model size vs metrics in the Controlled generation (CG-IC) dataset.

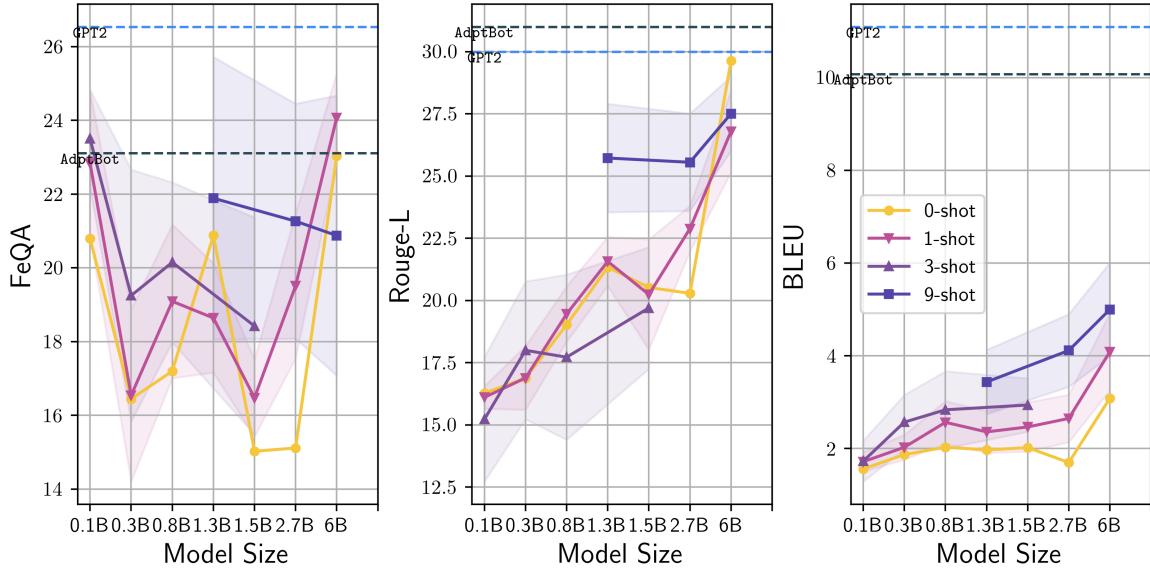


Figure 16: Model size vs metrics in the DialKG dataset.

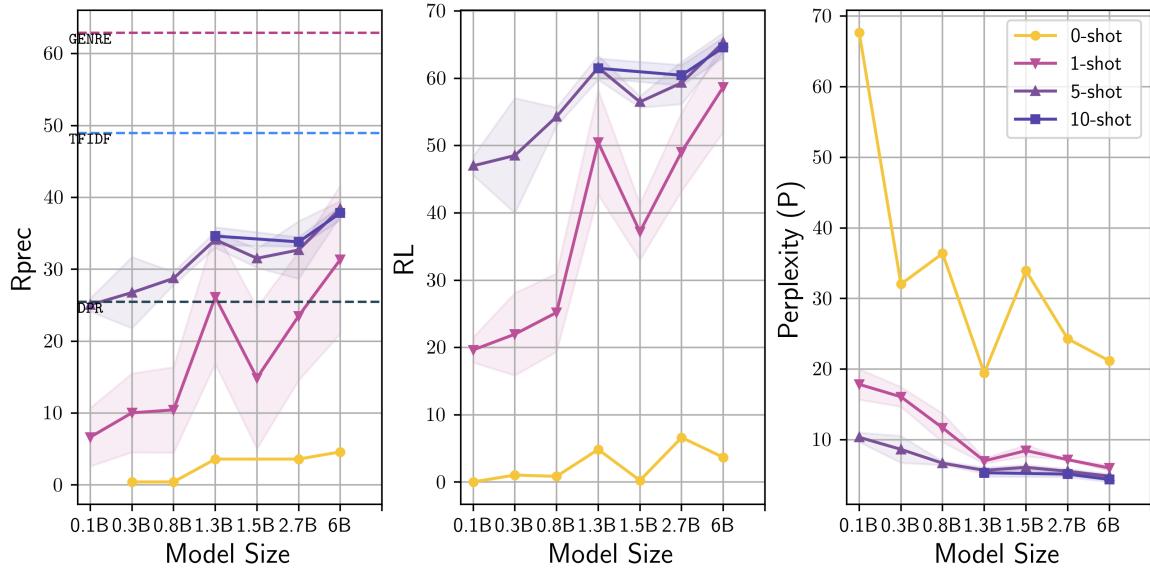


Figure 17: Model size vs metrics in the Wizard of Wikipedia Parse (WoW-Parse) dataset.

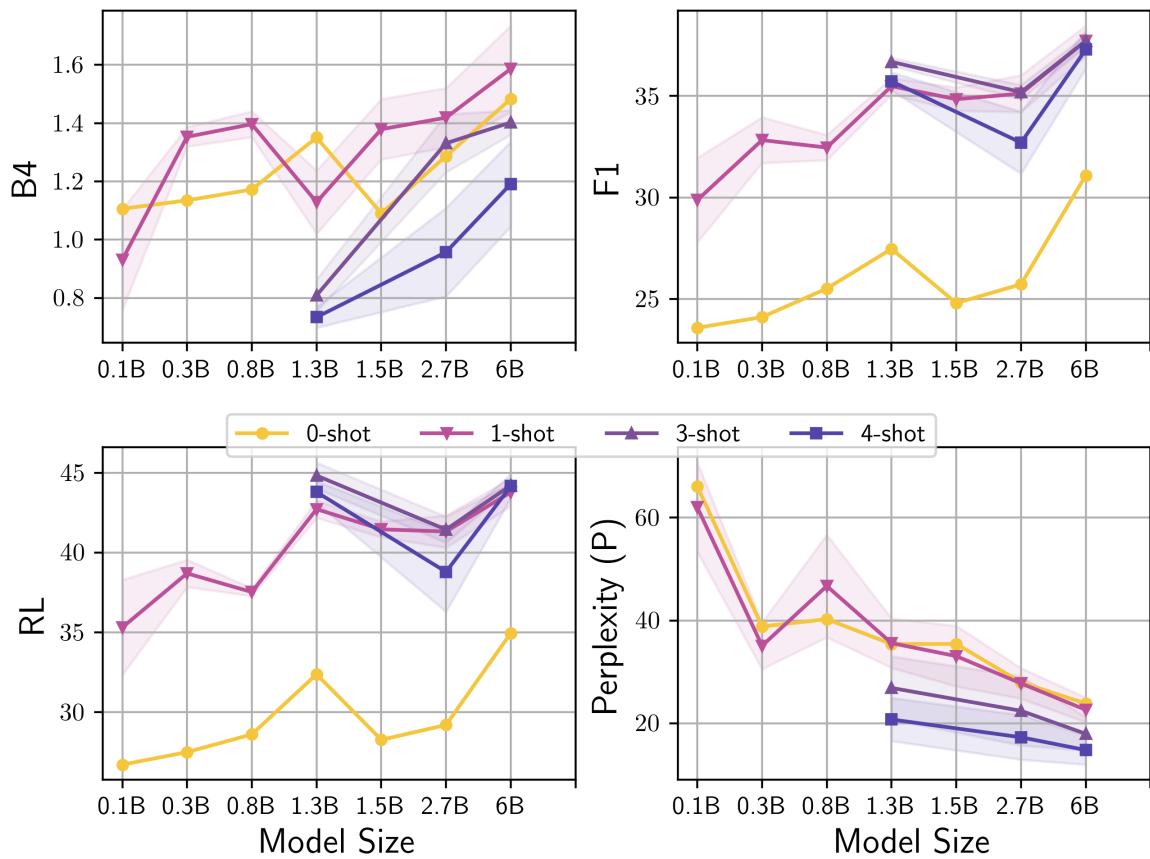


Figure 18: Model size vs metrics in the Wizard of Internet Parse (WiT-Parse) dataset.

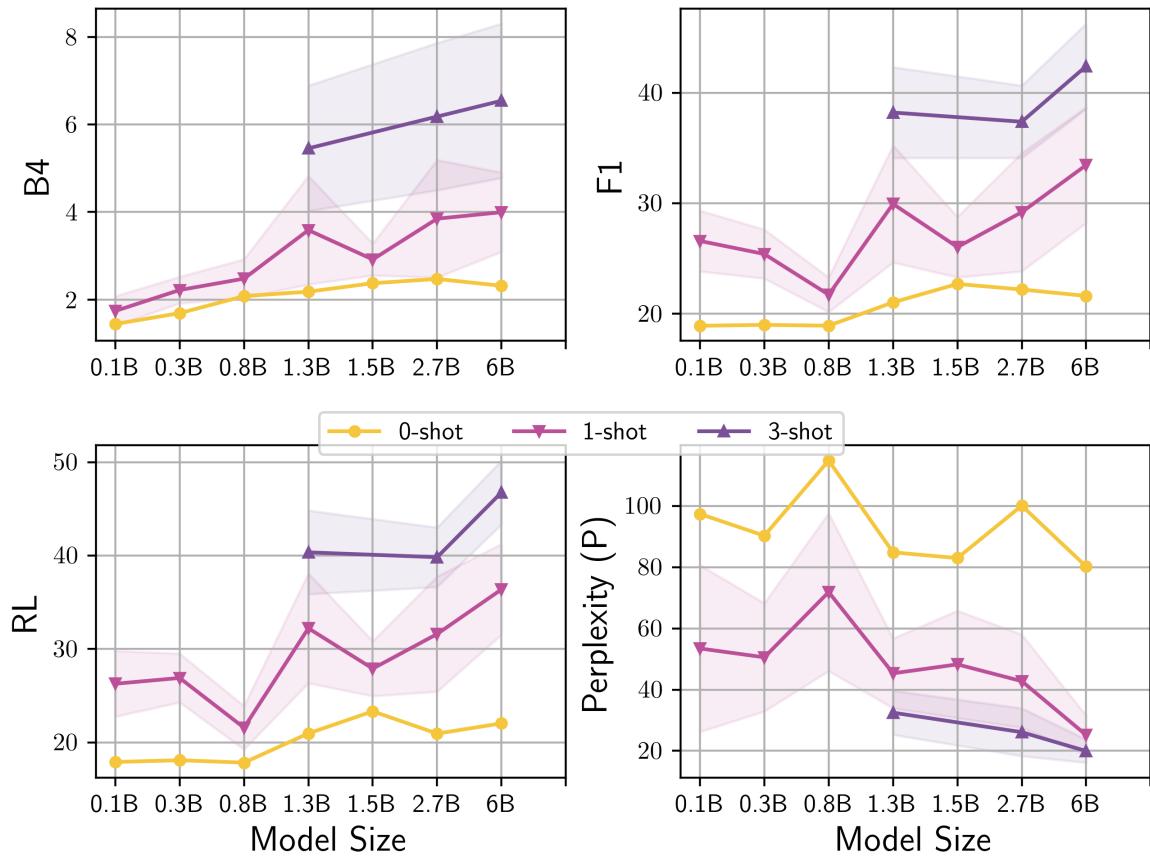


Figure 19: Model size vs metrics in the Multi-Session Chat Parse (MSC-Parse) dataset.

FEW-SHOT BOT

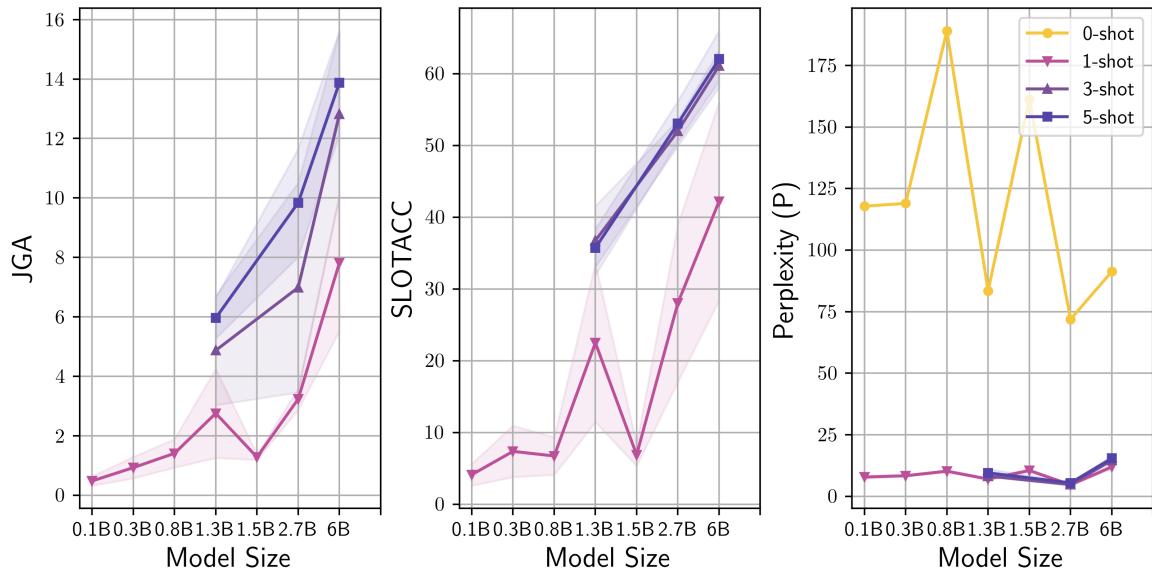


Figure 20: Model size vs metrics in the MWoZ-Single dataset.

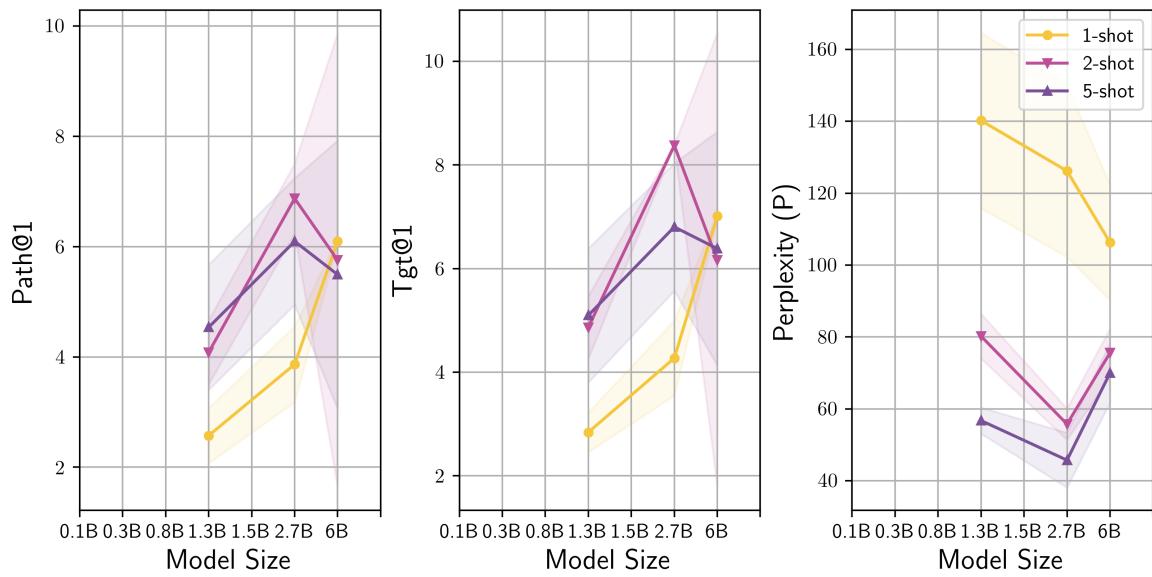


Figure 21: Model size vs metrics in the DialKG-Parse dataset.

```

1 # Initialize memory with user information if exist
2 memory = init()
3
4 # Dictionary with the prompt of each skill
5 prompts_response = get_prompts(["WoW", "WiT", "DialKG", "Persona", "IC",
6                                "DD", "ED", "MSC"])
7 prompts_parsing = get_prompts(["WoW-parse", "WiT-parse", "DialKG-parse",
8                                "MSC-parse"])
9
10 dialog_history = []
11 while True:
12     # Get the user input
13     usr_utt = input()
14
15     # Append utterance to the dialogue history
16     dialog_history.append(usr_utt)
17
18     # Select the skill using Equation (3)
19     skill = skill_selector(dialog_history, prompts)
20
21     # Retrieve knowledge from the Wiki/Internet/KG
22     if skill in ["WoW", "WiT", "DialKG"]:
23         query = parse(dialog_history, prompts_parsing[skill])
24         knowledge = retriever(query, skill)
25         update_memory(memory, skill, knowledge)
26
27     # Extract user knowledge
28     user_info = parse(dialog_history, prompts_parsing["MSC-parse"])
29     update_memory(memory, "MSC", user_info)
30
31     # Match memory content and dialogue history to FSB templates
32     dialogue_hist_prompt = convert_to_string(dialog_history, memory, skill)
33
34     # Generate system response using a LM or API (OpenAI or AI21)
35     sys_utt = generate(prompts_response[skill]+dialogue_hist_prompt)
36
37     # Append utterance to the dialogue history
38     dialog_history.append(sys_utt)
39
40     print(sys_utt)

```

Figure 22: Pseudo code for the FSB interaction algorithm.

```

1 # Dictionary with the prompt for each style in Image Chat (IC)
2 prompts_style = get_prompts(CG(["CG-IC"]))
3 all_styles = (prompts_response.keys())
4
5 dialog_history = []
6 while True:
7     # Get the user input
8     usr_utt = input()
9
10    # Get the style of the response
11    print(f"Style: {all_styles}")
12    style = input()
13
14    # Append utterance to the dialogue history
15    dialog_history.append(usr_utt)
16
17
18    # Match memory content and dialogue history to FSB templates
19    dialogue_hist_prompt = convert_to_string(dialog_history, {}, "CG-IC")
20
21    # Generate system response using a LM or API (OpenAI or AI21)
22    sys_utt = generate(prompts_style[style]+dialogue_hist_prompt)
23
24    # Append utterance to the dialogue history
25    dialog_history.append(sys_utt)
26
27    print(sys_utt)

```

Figure 23: Pseudo code for the controlled style generation.