Multi-Stage Prompting for Knowledgeable Dialogue Generation

Anonymous ACL submission

Abstract

Existing knowledge-grounded dialogue systems typically use finetuned versions of a pretrained language model (LM) and large-scale knowledge bases. These models typically fail to generalize on topics outside of the knowledge base, and require maintaining separate potentially large checkpoints each time finetuning is needed. In this paper, we aim to address these limitations by leveraging the inherent knowledge stored in the pretrained LM as well as its powerful generation ability. We propose a multi-stage prompting approach to generate knowledgeable responses from a single pretrained LM. We first prompt the LM to generate knowledge based on the dialogue context. Then, we further prompt it to generate responses based on the dialogue context and the previously generated knowledge. Results show that our knowledge generator outperforms the state-of-the-art retrieval-based model by 5.8% when combining knowledge relevance and correctness. In addition, our multi-stage prompting outperforms the finetuning-based dialogue model in terms of response knowledgeability and engagement by up to 10% and 5%, respectively. Furthermore, we scale our model up to 530 billion parameters and demonstrate that larger LMs improve the generation correctness score by up to 10%, and response relevance, knowledgeability and engagement by up to 10%.[1]

1 Introduction

Dialogue systems face the problem of producing bland and generic outputs that are devoid of content (Wolf et al., 2019; Holtzman et al., 2019; Ma et al., 2020). Recent efforts have been made to address these concerns by grounding dialogue responses on a source of knowledge (Dinan et al., 2018; Zhou et al., 2018; Zhao et al., 2019; Santhanam et al., 2020; Prabhumoye et al., 2021).

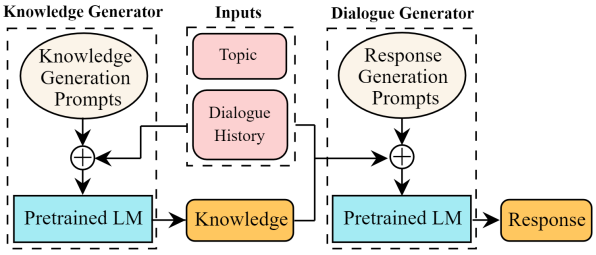

Figure 1: Our proposed framework (MSDP) for the knowledgeable dialogue generation.

Therefore, building a knowledgeable dialogue system has become one of the key milestone tasks in conversational research.

Current knowledge-grounded dialogue systems highly rely on a massive external knowledge corpus for a retrieval model to obtain relevant knowledge (Dinan et al., 2018; Kim et al., 2019; Zhao et al., 2020), which inevitably brings several limitations. First, retrieval systems are constrained to the size and domains of the database, and they cannot generalize to out-of-domain topics that are not covered by the database. Second, retrieval from a massive corpus takes substantial resources. Reimers and Gurevych (2021) show that it is more difficult for the state-of-the-art retrieval model (Karpukhin et al., 2020) to retrieve relevant knowledge when the size of the database increases. The larger database increases the chance that an irrelevant document is closer to the query embedding than the relevant document.

We aim to address these limitations by using a relatively small database and a pretrained language model (LM) (Shoeybi et al., 2019; Brown et al., 2020) as an additional source of knowledge to ground a dialogue system. Since the LM inherently stores a variety of knowledge (Petroni et al., 2019), it can help dialogue systems generalize to out-of-domain topics that are not explicitly present in the database. We propose a prompt-based approach to directly generate the context-relevant knowledge

[1] Our code is available at annoymous.com.

from the LM. Specifically, we select a few dialogue contexts and their associated knowledge from the database to be given as prompts to the LM for the knowledge generation. These samples are chosen such that the dialogue contexts are semantically closer to the current dialogue context.

Moreover, finetuning LMs, which current dialogue systems rely on, could lead to overfitting when the finetuning dataset is relatively small. Also, gigantic LMs like GPT-3 (Brown et al., 2020) and the recent 530B (Patwary et al., 2021), may only be available through APIs. Hence, finetuning them on the dialogue task might not be a feasible solution. To bypass the finetuning process, we propose to further prompt the LM to generate the response based on the dialogue context and previously generated knowledge. We select a few dialogue contexts and corresponding knowledge and responses to be given as prompts to the LM for the response generation. The samples are chosen such that their responses are knowledgeable and highly conditioned on the corresponding knowledge.

In summary, we present a novel **M**ulti-**S**tage **D**ialogue **P**rompting (MSDP) framework, which consists of a first-stage prompting for the knowledge generation and a second-stage prompting for the response generation. Our framework does not need any finetuning or updates to the pretrained weights of the LM, can generate relevant and factually correct knowledge, and is effective at producing knowledgeable and engaging responses.

Our contributions are summarized as follows:

- We propose a novel multi-stage prompting framework for knowledgeable dialogue generation that only uses a single LM and does not require any finetuning.
- We show that for in-domain dialogue topics, our knowledge generator can outperform the state-of-the-art retrieval model by 5.8% when combining relevance and correctness, and it can also better generalize to out-of-domain topics by a 6.4 F1-score improvement.
- We show that MSDP can outperform the finetuning-based dialogue model for response knowledgeability and engagement by up to 10% and 5%, respectively.
- We scale our technique up to a 530-billion-parameter LM and demonstrate that larger LMs improve the generation correctness score by up to 10%, and response relevance, knowledgeability and engagement by up to 10%.

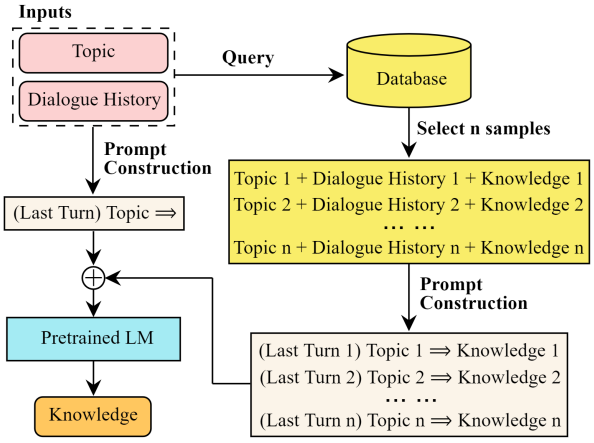

Figure 2: Prompting for the knowledge generation.

2 Framework

Our proposed multi-stage dialogue prompting (MSDP) framework is illustrated in Figure 1. It consists of a knowledge generator and a dialogue generator, both using the same pretrained LM. The knowledge generator produces relevant knowledge to the input topic and dialogue history, while the dialogue generator generates engaging and knowledgeable responses based on the dialogue context and the generated knowledge.

We denote the input topic as t, the input dialogue history as h, the last dialogue turn as h^*, and a database of samples as D. Each data sample in D is denoted by d_i, and consists of a topic t_i, a dialogue history h_i with the last turn as h_i^*, corresponding knowledge k_i, and a response r_i.

2.1 Knowledge Generator

To bypass the dependence on a large-scale knowledge base, we propose a prompt-based knowledge generation approach, which uses a relatively small database (about 70K samples) and a pretrained LM to generate context-relevant knowledge. As shown in Figure 2, the knowledge generation consists of sample selection and knowledge generation.

Sample Selection We hypothesize that selecting appropriate samples as prompts is the key to generating high-quality knowledge sentences. Intuitively, leveraging the knowledge from similar topics or dialogue context can help the LM to generate contextually relevant and factually correct knowledge sentences. Hence, we propose a query-based sample selection method, which aims to search similar

samples from D based on the input query (q). To ensure that the selected examples are relevant to the query, we utilize a pretrained sentence encoder (SE) (Devlin et al., 2019; Karpukhin et al., 2020) to obtain the representations for the query and each data sample (d_i) in D. Then, we calculate the similarity between the query and each sample using the dot product of their representations:

$$Sim(q, d_i) = SE(t + h)^\intercal \cdot SE(t_i + h_i),$$

where the input of the SE is a concatenation of the topic and dialogue history. Finally, we select n samples that have the highest similarity scores to q. This selection process can be done efficiently since the database is relatively small.

Knowledge Generation Inspired by the few shot approach in Brown et al. (2020), feeding the pretrained LM with suitable and intuitive prompts can allow it to generate relevant content. The template of the constructed prompts is illustrated in Figure 2. Concretely, the prompt for the i^{th} sample ($prompt_i$, $i \in [1, n]$) is "(h_i^*) $t_i \Rightarrow k_i$"[2], and the prompt for the current dialogue context ($prompt_{curr}$) is "(h^*) $t \Rightarrow$", where we use the symbol "$\Rightarrow$" to guide the LM for knowledge generation. We only use the last dialogue turn to construct prompts because the previous turns are mostly not relevant to the knowledge, and adding redundant information could lead to negative effects for knowledge generation. Given that k_i usually has a closer connection with t_i than h_i^*, we put k_i closer to t_i than h_i^* in the prompts. Finally, we concatenate the constructed prompts using "\n" and feed them into the LM to generate the knowledge:

$$k' = \mathcal{LM}(prompt_1 \texttt{\textbackslash n} \ldots prompt_n \texttt{\textbackslash n}\ prompt_{curr})$$

where k' denotes the generated knowledge for the input. Since "\n" is used to separate the prompts, the model will start generating "\n" followed by another random example after finishing the knowledge generation. Hence, we consider the generated sentence before "\n" as k'.

2.2 Dialogue Generator

The architecture of our proposed dialogue generator is illustrated in Figure 3. Finetuning a LM could lead to overfitting when the finetuning dataset is relatively small. In addition, since usually one can only access to the gigantic LMs, like GPT-3 (Brown et al., 2020) and the recent 530B (Patwary et al., 2021) using only APIs, finetuning them might not be a feasible solution. Therefore, we propose to circumvent the finetuning by prompting the pretrained LM for the response generation, which requires only a few dialogue examples. To generate knowledgeable and engaging responses, we focus on how to select samples and how to effectively prompt the LM for the response generation.

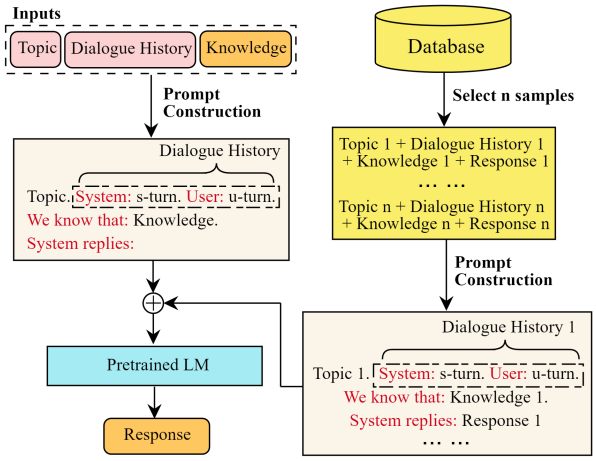

Figure 3: Prompting for the dialogue response generation. We use comprehensive words (denoted in red color) to connect the dialogue history, knowledge and response for the prompt construction.

Sample Selection One of the essential skills for the knowledgeable dialogue model is to effectively leverage the knowledge produced in the first stage, in order to make the generated responses knowledgeable. Considering that we can provide the LM with only a few dialogue samples, it could be difficult for it to learn how to generate a response based on the knowledge unless there are strong connections between the dialogue response and knowledge in the samples that we provide. Hence, we focus on selecting the samples in which the responses are knowledgeable and highly conditioned on the corresponding ground truth knowledge. Concretely, for each example in the database, we calculate how much ground truth knowledge accounts for the dialogue response by using the word overlap ratio. Then, we filter out the examples where the ratio is lower than 0.6 (this number is decided based on a hyper-parameter search among {0.4, 0.5, 0.6, 0.7, 0.8}). Also having responses be too knowledgeable could make it less engaging. Therefore, we also filter out the examples where the ratio is higher than 0.9 since we expect the response to contain

[2]For example, (I love pizza) Pizza $\Rightarrow$ Pizza is a traditional Italian dish typically topped with tomato sauce and cheese.

information other than the knowledge. After the filtering, to ensure that our approach does not depend on the dialogue context, we randomly select n samples from the rest of the dialogue examples. These selected n samples will be later constructed as prompts and used for the response generation.

Response Generation Aside from the ability to leverage the generated knowledge, another essential skill for the dialogue model is to have the ability to chat based on the dialogue context. To equip our model with this skill, we focus on constructing intuitive prompts for the selected examples and feed them into the LM. The constructed prompts for the selected examples and inputs are illustrated in Figure 3. For prompts from the selected examples, we use "System:" and "User:" to connect different turns in the dialogue history, and "We know that:" and "System replies:" are used to introduce the knowledge and response, respectively. For prompts from the current conversation (i.e., inputs), we follow the same template except that we keep the response empty for the pretrained LM to generate.

After the prompt construction, we concatenate the prompts for selected samples and the inputs using "`\n`", and then feed them into the pretrained LM to generate the response. Similar to what we have described in Section 2.1, we consider the generated sentence before "`\n`" as the response.

3 Experimental Setup

3.1 Datasets

We evaluate our model using two knowledge-grounded dialogue datasets: Wizard of Wikipedia (WoW) (Dinan et al., 2018) and Wizard of Internet (WoI) (Komeili et al., 2021).

WoW uses Wikipedia as the knowledge base and covers a wide range of topics (1365 in total). Its test dataset is split into two subsets: test seen and test unseen. Each data sample has a chosen topic, a dialogue history, a ground truth knowledge sentence, and a corresponding dialogue response. The dialogue topics in the test seen subset appear in the training dataset, while the topics in the test unseen subset do not. Different from WoW, the collection of WoI is grounded on the whole Internet, which covers a wider range of topics than Wikipedia.

In the experiments, we only use the training set of WoW (as the database) for the sample selection of our prompting framework. All the models (our model and baselines) *do not use any training sample from WoI*, and we directly evaluate them on the test set of WoI. The motivation for doing this is to test how well our model can generalize to the unseen scenario where topics do not exist in the database. The topics in the WoW test unseen set do not exist in the database, and only 5.76% of topics in the WoI test set exist in the database. We calculate the 13-gram overlap (Brown et al., 2020) between the knowledge used in WoI test set and the database, and find the overlap is as little as 0.39%.

3.2 Baselines for Knowledge Generation

DPR DPR, Dense Passage Retriever (Karpukhin et al., 2020), is the state-of-the-art retrieval model. To make DPR fit into the dialogue scenario, we finetune it on the training dataset of WoW. Concretely, it is finetuned to map the dialogue context (topic and dialogue history pair) and corresponding ground truth knowledge into similar vector space.[3]

FKG FKG denotes the finetuning-based knowledge generation. We use the training dataset of WoW to finetune the LM. Concretely, the input is a concatenation of a topic and dialogue history, and the LM is finetuned to generate relevant knowledge. We use FKG as a baseline to compare the performance of the knowledge generation between the prompt-based and finetuning-based methods.

3.3 Baselines for Response Generation

PPLM PPLM denotes the plug and play language model (Dathathri et al., 2019). We choose it as a baseline because our MSDP can be considered as using topics to control the LM to generate responses, and PPLM, which does not need finetuning either, can be also used to control LMs for topic-relevant generation. We follow Madotto et al. (2020) and use dialoGPT (Zhang et al., 2020) for PPLM to enable the response generation. We use ConceptNet (Speer et al., 2017) to produce topic-relevant bag-of-words for the response generation.

FCM w/ DPR FCM denotes the finetuning-based conversational model. We use the training dataset of WoW to finetune the LM. This baseline has the same pipeline as that of our MSDP. Instead of doing prompting, it uses DPR for producing the knowledge and FCM to generate a response.

FCM w/ FKG This baseline follows the same setting as FCM w/ DPR, except that we use FKG instead of DPR to produce knowledge.

[3]The details of this finetuning is placed in Appendix F.

Models	Wizard of Wikipedia (Seen)				Wizard of Wikipedia (Unseen)				Wizard of Internet			
	B	M	R-L	F1	B	M	R-L	F1	B	M	R-L	F1
DPR (seen)	18.32	12.82	21.91	24.86	8.09	6.80	12.04	13.71	2.37	3.90	5.73	7.03
DPR (wiki)	9.95	9.27	15.11	18.42	10.06	9.80	15.46	18.24	3.49	5.36	7.35	9.16
FKG	21.08	14.61	25.57	27.83	9.01	8.26	15.61	16.07	3.45	4.69	6.55	7.14
MSDP-KG†	**23.68**	**15.93**	**27.88**	**31.55**	**11.54**	**10.53**	**19.05**	**20.15**	**5.20**	**7.38**	**10.47**	**11.12**

Table 1: Results of automatic metrics for the knowledge generation/retrieval models across three datasets. B, M, and R-L denote the averaged BLEU, METEOR, and ROUGE-L, respectively. DPR (seen) can only access the knowledge in the training dataset of WoW, while DPR (wiki) can access all the knowledge in Wikipedia. †We use "-KG" to denote the knowledge generation part of MSDP (same for the following tables). Both FKG and MSDP-KG use a 126m LM to match the size of DPR, which is based on a 110m LM.

Models	Wizard of Wikipedia (Seen)			Wizard of Wikipedia (Unseen)			Wizard of Internet		
	Relevance	Correctness	Combination	Relevance	Correctness	Combination	Relevance	Correctness	Combination
DPR (110m)	3.39	**4.00**	3.39	3.38	**4.00**	3.38	2.79	**4.00**	2.79
MSDP-KG (126m)	3.76*	3.71	3.59*	3.80*	3.19	3.12	3.60*	2.93	2.83
MSDP-KG (357m)	3.79*	3.80	3.69*	3.84*	3.56*	3.47	3.74*	3.29*	3.21*
MSDP-KG (1.3b)	3.81*	3.90*	3.72*	3.89*	3.72*	3.62*	3.77*	3.51*	3.38*
MSDP-KG (530b)	**3.88***	3.96*	**3.84***	**3.92***	3.94*	**3.87***	**3.81***	3.84*	**3.70***

Table 2: Human evaluations for the knowledge generation/retrieval models. We compare MSDP-KG with DPR (seen) on the WoW (seen) dataset, and DPR (wiki) on the WoW (unseen) and WoI datasets. We directly use a score of 4 to rate the correctness of the knowledge retrieved by DPR since all knowledge in the database is correct. For relevance and combination, we conduct a t-test between MSDP-KG and DPR. For the correctness, we conduct a t-test between MSDP-KG (357m-530b) and MSDP-KG (126m). * denotes the result is significant at $p < 0.05$.

3.4 Automatic Evaluation

For evaluating both knowledge generation and response generation, we follow previous works (Rashkin et al., 2019; Dinan et al., 2018; Prabhumoye et al., 2021) to evaluate the generated sentences against the reference sentences on averaged BLEU (an average of BLEU-1,2,3,4) (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Denkowski and Lavie, 2011), and unigram F1. Additionally, we follow Komeili et al. (2021) to use knowledge F1 (KF1) to evaluate the knowledgeability of the response generation.

3.5 Human Evaluation

Knowledge Generation For evaluating the quality of the knowledge generation, we use **relevance**, **correctness**, and a **combination** of the two metrics. To evaluate the relevance, we provide annotators with the topic and dialogue, as well as the model-produced knowledge, and ask them to rate how relevant the knowledge is to the topic and dialogue on a scale from 1 to 4, where 1 means not relevant at all, 2 is only a little relevant, 3 is somewhat relevant, and 4 is highly relevant. To evaluate the correctness, we provide the annotators with the topic and the model-generated knowledge, and ask them to rate how correct the knowledge is on a scale from 1 to 4, where 1 is not correct at all, 2 is less than half is correct, 3 is half and more than half is correct, and 4 is all correct.

In addition, given that the overall quality of the knowledge depends on both relevance and correctness, we calculate a combination score based on the minimum between the relevance and correctness for each evaluated sample:

$$\text{combination} = min(\text{relevance}, \text{correctness}).$$

We use minimum instead of average or maximum because both relevance and correctness are indispensable for the quality of the knowledge.

Response Generation For evaluating the quality of the response generation, we use **relevance**, **engagement**, and **knowledgeability**. To evaluate the relevance, we provide the annotators with a topic and dialogue history, as well as a pair of generated responses from two models and ask them to choose which is more relevant to both topic and dialogue history. For engagement and knowledgeability, we provide the annotators with the same samples as for relevance, and ask them to choose which is more engaging and knowledgeable, respectively. For all these metrics, we let annotators choose a tie when the quality is comparable.[4]

[4]We put the human evaluation setup in the Appendix E, and the generation samples are in Appendix D.

Models	Wizard of Wikipedia (Seen)					Wizard of Wikipedia (Unseen)					Wizard of Internet				
	B	M	R-L	F1	KF1	B	M	R-L	F1	KF1	B	M	R-L	F1	KF1
PPLM	2.08	4.89	6.32	11.40	6.63	2.15	4.86	6.30	11.38	6.77	1.78	4.58	5.70	9.83	4.48
FCM w/ DPR (seen)	8.72	8.40	14.91	17.40	17.13	6.51	6.88	12.12	13.71	11.54	4.06	6.27	9.17	12.90	7.38
FCM w/ DPR (wiki)	7.36	7.63	13.65	16.00	13.80	6.98	7.43	13.33	15.46	13.38	4.47	6.65	9.65	13.52	7.78
FCM w/ FKG	8.97	8.67	15.36	18.31	18.85	6.73	7.19	12.97	14.68	12.59	4.75	6.56	9.72	13.71	7.89
FCM w/ MSDP-KG	**10.17**	9.34	16.00	**19.45**	21.02	7.12	7.70	13.93	**16.75**	13.96	**4.80**	6.82	**10.21**	**14.39**	8.77
MSDP	9.97	**9.95**	**18.62**	17.57	**22.95**	**8.30**	**8.65**	**17.40**	16.00	**16.57**	4.66	**8.00**	9.80	14.09	**9.67**

Table 3: Results of automatic metrics for the knowledgeable conversational model. Both FKG and MSDP-KG (associated with FCM) use a 126m LM to match the size of DPR, which is based on a 110m LM. MSDP uses a 357m LM to match the size of FCM, which is also based on a 357m LM.

3.6 Training Details

The LMs used for our MSDP model, and baselines FKG and FCM are GPT-style (Brown et al., 2020) models and are pretrained using the toolkit in Shoeybi et al. (2019). PPLM uses dialoGPT-medium, which has 355 million parameters (355m). The LM in FCM has 357m parameters, and DPR consists of two encoders (question encoder and passage encoder) with a size of 110m parameters each. To test how different model sizes affect the performance, we evaluate our methods with 126m, 357m, 1.3 billion (1.3b), and 530 billion (530b) parameters LMs. For the sample selections, we choose 10 samples for the prompting in the knowledge generation, and 20 samples for the prompting in the response generation. To ensure a fair comparison, we select the top-1 knowledge from the DPR model, and we use deterministic greedy search for the generation of LM. We use the question encoder of DPR as the sentence encoder in the sample selection of the knowledge generation. Note that this sentence encoder can be replaced with any pretrained model (e.g., BERT (Devlin et al., 2019)), and as shown in Section 4.3, there is only a marginal difference between using BERT and DPR's question encoder (about 0.5 F1 for the dialogue response generation).

4 Results

In this section, we compare our framework with baselines for the knowledge and response generation. Then, we conduct ablation studies to further analyze the effectiveness of our framework.

4.1 Knowledge Generation

We first analyze how DPR performs when different sizes of databases are available. From Table 1, we can see that in the WoW (seen) scenario, DPR (seen) can retrieve generally better knowledge compared to DPR (wiki) since the corpus size for DPR (wiki) is much larger. This further confirms that larger database makes retrieval of relevant information more difficult DPR as shown in Reimers and Gurevych (2021). However, DPR (seen) cannot work in the unseen scenarios (WoW (unseen) and WoI) due to the absence of a topic-relevant knowledge base. Compared to DPR, FKG achieves better results when the topics are covered in the training dataset (WoW (seen)), while its generalization ability to unseen topics is relatively limited since we can see that DPR (wiki) has better performance than FKG in WoW (unseen) and WoI. Our approach, MSDP-KG, demonstrates a powerful generalization ability to unseen topics, which leads to better results across the three datasets compared to all the baselines.

To evaluate the generation quality, we compare MSDP-KG with DPR using human evaluation, and the results are shown in Table 2. We find that MSDP-KG (126m) can generate much more relevant knowledge compared to DPR (with more than 10% improvement in the relevance score). In addition, MSDP-KG (126m) can produce generally correct knowledge in WoW (seen) since it can refer to the knowledge in similar topics, which leads to a better combination score than DPR (a 5.8% improvement). Meanwhile, its generation correctness is somewhat limited in WoW (unseen) and WoI, which can be attributed to the relatively small model size and the pretraining corpus. We notice that MSDP-KG (126m) also achieves a better combination score in WoI due to a very significant improvement in the relevance score. This is because the knowledge base for DPR is limited in the Wikipedia domain, which lowers its generalization ability to a wider range of topics on the Internet.

Furthermore, we observe that larger LMs bring improvements on all metrics. MSDP-KG (357m) can outperform DPR in all datasets for the combination score. We find that larger LMs can also bring

Model A	Rele.	Enga.	Know.	Model B
Wizard of Wikipedia (Seen)				
M (357m)	41.5 - 40.0	**43.7** - 38.5	**50.4** - 37.8	F (357m)
M (1.3b)	**48.9** - 40.0	**47.8** - 37.8	**47.8** - 35.6	M (357m)
M (530b)	**54.4** - 41.1	**53.3** - 41.1	**51.1** - 42.2	M (1.3b)
Wizard of Wikipedia (Unseen)				
M (357m)	39.3 - 40.0	**46.7** - 43.0	**48.9** - 37.8	F (357m)
M (1.3b)	**50.0** - 38.9	**51.1** - 41.1	**46.7** - 41.1	M (357m)
M (530b)	**52.2** - 42.2	**51.1** - 40.0	**50.0** - 38.9	M (1.3b)
Wizard of Internet				
M (357m)	42.2 - 43.7	41.5 - 40.7	**44.4** - 39.3	F (357m)
M (1.3b)	**51.1** - 42.2	**50.0** - 38.9	**44.4** - 41.1	M (357m)
M (530b)	**54.4** - 38.9	**52.2** - 42.2	**56.7** - 38.9	M (1.3b)

Table 4: Human evaluation results on the dialogue models in terms of relevance (Rele.), engagement (Enga.), and knowledgeability (Know.). M denotes the MSDP and F denotes the FCM w/ DPR (DPR (seen) for WoW (seen), and DPR (wiki) for WoW (unseen) and WoI). For each number pair, the left number denotes the win rate for model A and the right one for model B. Note that the numbers in each pair might not sum to 100 since the annotators can choose "tie".

Models	WoW (Seen)				WoW (Unseen)			
	B	M	R-L	F1	B	M	R-L	F1
MSDP-KG	**24.5**	**16.4**	**28.7**	**33.2**	**12.4**	**11.1**	**19.6**	**22.0**
w/ BERT	23.1	15.5	27.3	31.1	12.1	10.5	19.0	21.2
w/ random	12.9	9.72	17.6	18.8	9.85	10.1	17.5	19.8
w/o topic	21.5	14.2	25.3	27.2	7.37	6.86	13.3	14.2

Table 5: Ablation studies for the knowledge generation, in terms of the sentence encoder (w/ BERT), sample selection method (w/ random), and the importance of the input topic (w/o topic). The size of the LM is 357m.

Models	Wizard of Wikipedia (Unseen)				
	B	M	R-L	F1	KF1
MSDP	**8.30**	**8.65**	**17.40**	**16.00**	**16.57**
w/ BERT	8.13	8.38	17.16	15.51	16.13
w/ random	5.56	6.50	16.48	14.32	13.13
w/o topic	6.32	7.17	15.70	13.06	11.77

Table 6: Ablation studies for the response generation, in terms of the sentence encoder in the knowledge generation, sample selection method, and the importance of an input topic. The size of the LM is 357m.

significant improvement on the correctness score (e.g., 357m improves over 126m by 11.5% in WoW (unseen)). Moreover, MSDP-KG (530b) achieves a 3.94 correctness score for WoW (unseen), which means about 95% of the generations are all correct.

4.2 Response Generation

The automatic metrics for conversational models are shown in Table 3. We notice that PPLM does not perform as well as the other models for this task since it does not explicitly use the relevant knowledge for the response generation. For the FCM-based models, we find that a better knowledge generation leads to a performance improvement as does a better retrieval model. "FCM w/ MSDP-KG" outperforms baseline models. Interestingly, our MSDP also generally outperforms the FCM-based baselines on different automatic metrics, especially the KF1 score. For example, compared to "FCM w/ DPR (wiki)", MSDP has a 3.19 higher KF1 score in WoW (unseen) and a 1.89 higher KF1 score in WoI. This can be attributed to the sample selection for the response generation, which selects knowledgeable responses that are highly based on the knowledge sentence. We also observe that MSDP achieves comparable results to the "FCM w/ MSDP-KG", which further illustrates the effectiveness of our proposed framework.

The human evaluations from Table 4 further confirms the effectiveness of MSDP. Compared to "FCM w/ DPR", MSDP can generate relevant responses, and more engaging and knowledgeable responses. For WoW (seen) and WoW (unseen), MSDP has more a than 10% higher win rate in terms of knowledgeability, and about 3% to 5% higher win rate in terms of the engagement.

Furthermore, we observe that larger LMs generally improve on response relevance, engagement, and knowledgeability by about 10% win rate.

4.3 Ablation Studies

Sentence Encoder In the sample selection of the knowledge generation, we obtain the similarity based on the DPR's question encoder, and we investigate how effective the generation will be if we replace the question encoder with a simpler model, like BERT (Devlin et al., 2019). From Table 5, using BERT as the sentence encoder achieves comparable performance to using DPR's question encoder. Also, from Table 6, we can see that using BERT in MSDP-KG only slightly lowers the performance in the response generation. These results confirms the effectiveness of our proposed method.

Sample Selection We study the effectiveness of our sample selection methods in both knowledge generation and response generation by using the random selection as a comparison. From Table 5, we can see that using randomly selected samples consistently decreases the performance in all metrics. Since the random selection does not leverage

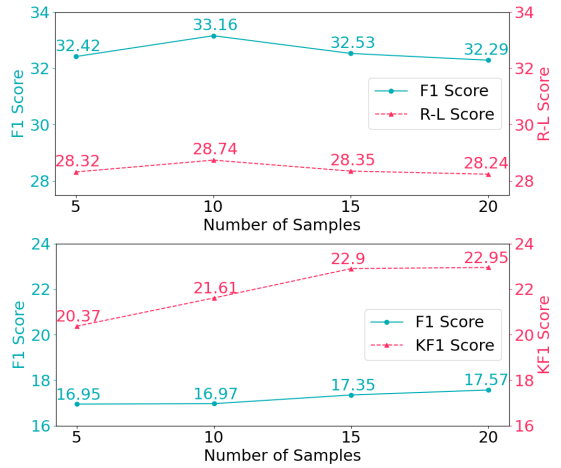

Figure 4: Effectiveness for different numbers of samples for the knowledge generation (top) and response generation (bottom). The size of the LM is 357m, and the results are from WoW (unseen).

the information from the database, the performance drop is especially significant in WoW (seen). In addition, from Table 6, "MSDP" significantly outperforms "MSDP w/ random" in all metrics, which confirms the effectiveness our proposed sample selection for the response generation.

Importance of Input Topic In our framework, a topic is a part of the input. To investigate the effectiveness of using a topic, we remove the input topic from the knowledge generation and response generation. As shown in both Table 5 and Table 6, we can see that providing a topic in the input is important, especially for the unseen scenario, where we observe a 7 F1-score decrease for "MSDP-KG w/o topic" in WoW (unseen).

Number of Samples for Prompting We further study how sample size affects the prompting performance. From Figure 4 (top), the number of samples will not significantly affect the knowledge generation. Interestingly, the performance of knowledge generation starts to slightly drop when sample size increases from 10. We conjecture that selecting too many samples might induce less similar samples to the input dialogue context, which could impact the performance negatively. As shown in Figure 4 (bottom), having more samples can slightly bring better responses. This is because, with more samples as references, the LM can better understand how to generate responses based on the given knowledge, which leads to a higher F1 and KF1 scores.[5]

[5]More ablation studies and results of automatic metrics for the model scaling are in the Appendix A, B, and C.

5 Related Work

Knowledge & Prompting Pretrained LMs are shown to possess commonsense knowledge (Davison et al., 2019; Bosselut et al., 2019; Rajani et al., 2019; Zhou et al., 2020), and can be prompted to do cloze questions (Petroni et al., 2019; Jiang et al., 2020; Brown et al., 2020; Shin et al., 2020; Schick and Schütze, 2021; Qin and Eisner, 2021), question answering (Brown et al., 2020; Patwary et al., 2021), etc. Despite the extensive research having explored the knowledge in LMs, little research has focused on directly generating context-relevant knowledge from LMs.

Recently, Zheng and Huang (2021) and Madotto et al. (2021), in concurrent works to ours, proposed to prompt LMs for the dialogue generation. Different from them, we focus on the knowledge-grounded scenario and propose a multi-stage prompting framework to leverage the inherent knowledge in LMs.

Knowledge-grounded Dialogues Grounding dialogue responses based on a knowledge base is emerging as an important step in research of human-machine conversation (Dinan et al., 2018; Zhou et al., 2018; Kim et al., 2019; Moon et al., 2019; Zhao et al., 2019; Chen et al., 2020; Li et al., 2020; Zhan et al., 2021; Prabhumoye et al., 2021; Komeili et al., 2021). Our proposed framework circumvents the need of LM finetuning and a massive knowledge base, which current models typically rely on.

6 Conclusion

We propose a novel multi-stage dialogue prompting framework which consists of a first-stage prompting for the knowledge generation and a second-stage prompting for the response generation. Both automatic metrics and human evaluations show that compared to the state-of-the-art retrieval-based model, our knowledge generator can generate better context-relevant knowledge for both in-domain and out-of-domain dialogue topics. Moreover, our framework is able to produce more knowledgeable and engaging responses compared to the finetuning-based dialogue model. Additionally, we conduct comprehensive ablation studies to show the effectiveness of our proposed methods. Furthermore, we scale the LM up to 530 billion parameters and demonstrate that larger LMs consistently improve the generation correctness, and response relevance, knowledgeability, and engagement.

References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Joe Davison, Joshua Feldman, and Alexander M Rush. 2019. Commonsense knowledge mining from pretrained models. 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 1173–1178.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91.

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

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

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

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. 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.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2019. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.

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

Linxiao Li, Can Xu, Wei Wu, YUFAN ZHAO, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. In *Advances in Neural Information Processing Systems*, volume 33, pages 8475–8485. Curran Associates, Inc.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020. Plug-and-play conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2422–2433.

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

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

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

Mostofa Patwary, Mohammad Shoeybi, Patrick LeGresley, Shrimai Prabhumoye, Jared Casper, Vijay Korthikanti, Vartika Singh, Julie Bernauer, Michael Houston, Bryan Catanzaro, Shaden Smith,

Brandon Norick, Samyam Rajbhandari, Zhun Liu, George Zerveas, Elton Zhang, Yazdani Aminabadi, Xia Song, Yuxiong He, Jeffrey Zhu, Jennifer Cruzan, Umesh Madan, Luis Vargas, and Tiwary Saurabh. 2021. Using deepspeed and megatron to train megatronturing nlg 530b, the world's largest and most powerful generative language model.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. 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.

Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. 2021. Focused attention improves document-grounded generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4274–4287.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212.

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

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Nils Reimers and Iryna Gurevych. 2021. The curse of dense low-dimensional information retrieval for large index sizes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 605–611, Online. Association for Computational Linguistics.

Sashank Santhanam, Wei Ping, Raul Puri, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Local knowledge powered conversational agents. *arXiv preprint arXiv:2010.10150*.

Timo Schick and Hinrich Schütze. 2021. 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.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. 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.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021. Augmenting knowledge-grounded conversations with sequential knowledge transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5621–5630.

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

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2019. Low-resource knowledge-grounded dialogue generation. In *International Conference on Learning Representations*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390.

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

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

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9733–9740.

A Perplexity-based Sample Selection

We investigated another sample selection method (i.e., perplexity-based selection) for the knowledge generation. The knowledge generation using perplexity-based selection is depicted in Figure 5. The details of this sample selection is described as follows. Note that we denote the sample selection method for the knowledge generation in the main paper (Section 2.1) as the query-based sample selection.

Instead of selecting samples based on the current conversation (i.e., query), perplexity-based method will complete the sample selection before the inference, and the selected examples can be used for all inputs (i.e, topic and dialogue history pairs). Intuitively, using easy to understand prompts (instead of incomprehensible ones) enables the pretrained language models quickly comprehend the task and push it to generate the knowledge that is more topic-relevant and factually correct. To find comprehensible prompts, we first perform the prompt construction[6] for each data example in the database. We then calculate the perplexity for each prompt using a GPT-2 model (Radford et al., 2019) and select top-n prompts that have the lowest perplexities.[7]

Compared to query-based selection, the prompts selected based on perplexities are less relevant to the test example, which could generally lead to a worse generation quality. However, its advantage is that we do not need to select samples from the database for every input. Technically, it needs only a few easy to understand samples (i.e., 10 samples) for prompting.

B Ablation Studies Results

In the ablation study, we compare the query-based sample selection method (used in MSDP) and the perplexity-based sample selection method. We also provide the automatic metrics for different model sizes. We denote the sample selection method for the knowledge generation in the main paper (Section 2.1) as the query-based selection. In the tables, we use "ppl." to denote that the model is using the perplexity-based sample selection for the knowledge generation, and "que." to denote that the model is using the query-based sample selection for the knowledge generation.

[6]The prompt construction is the same as the query-based sample selection proposed in the main paper.

[7]To ensure a fair comparison with the query-based sample selection in the main paper (Section 2.1), we choose top-10 samples for the perplexity-based sample selection.

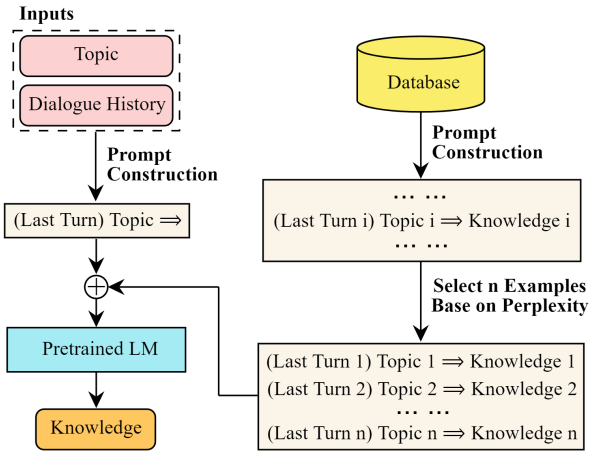

Figure 5: Prompting for the knowledge generation using the perplexity-based sample selection.

Models	B	M	R-L	F1
Wizard of Wikipedia (Seen)				
FKG	21.08	14.61	25.57	27.83
MSDP-KG (ran.)	8.73	8.56	15.35	16.37
MSDP-KG (ppl.)	9.61	9.48	16.95	17.83
MSDP-KG (que.)	**23.68**	**15.93**	**27.88**	**31.55**
Wizard of Wikipedia (Unseen)				
FKG	9.01	8.26	15.61	16.07
MSDP-KG (ran.)	8.89	9.11	16.19	16.42
MSDP-KG (ppl.)	9.94	10.08	17.91	18.44
MSDP-KG (que.)	**11.54**	**10.53**	**19.05**	**20.15**

Table 7: Ablation study for knowledge generation models. "ran." denotes the prompts are randomly selected, "ppl." denotes the prompts are selected based on the lowest perplexity, and "que." denotes the prompts are selected based on the query.

The ablation studies between perplexity-based sample selection and query-based sample selection are shown in Table 7 and Table 8. We also add finetuning-based knowledge generation (FKG), and sample selection by random into the comparison to better analyze the perplexity-based sample selection method.

Knowledge Generation From Table 7, we can see that perplexity-based selection generally achieves better results across all automatic metrics compared to the sample selection by random, which confirms the effectiveness of using easy to understand samples for prompting. We find that MSDP-KG (ppl.) performs much worse than FKG in WoW (seen). It is because FKG fully utilize the

Models	B	M	R-L	F1	KF1
Wizard of Wikipedia (Seen)					
FCM w/ FKG	8.97	8.67	15.36	18.31	18.85
FCM w/ MSDP-KG (ppl.)	6.93	7.67	14.01	16.89	13.59
FCM w/ MSDP-KG (que.)	**10.17**	9.34	16.60	**19.45**	21.02
MSDP (ppl.)	8.18	8.43	17.46	15.92	14.73
MSDP (que.)	9.97	**9.95**	**18.62**	17.57	**22.95**
Wizard of Wikipedia (Unseen)					
FCM w/ FKG	6.73	7.19	12.97	14.68	12.59
FCM w/ MSDP-KG (ppl.)	7.03	7.58	13.81	16.54	13.23
FCM w/ MSDP-KG (que.)	7.12	7.70	13.93	**16.75**	13.96
MSDP (ppl.)	7.95	8.46	17.14	15.56	15.49
MSDP (que.)	**8.30**	**8.65**	**17.40**	16.00	**16.57**

Table 8: Ablation study for knowledgeable conversational models. "MSDP (ppl.)" and "MSDP (que.)" uses "MSDP-KG (ppl.)" and "MSDP-KG (que.)", respectively, as the knowledge generator.

Models	B	M	R-L	F1
Wizard of Wikipedia (Seen)				
MSDP-KG (126m)	23.68	15.93	27.88	31.55
MSDP-KG (357m)	24.48	16.37	28.74	33.16
MSDP-KG (1.3b)	25.62	17.18	29.66	34.52
MSDP-KG (530b)	**27.45**	**19.34**	**33.09**	**35.73**
Wizard of Wikipedia (Unseen)				
MSDP-KG (126m)	11.54	10.53	19.05	20.15
MSDP-KG (357m)	12.38	11.10	19.64	21.98
MSDP-KG (1.3b)	13.49	11.94	20.68	23.65
MSDP-KG (530b)	**18.50**	**15.15**	**25.87**	**29.40**

Table 9: Ablation study for MSDP-KG (que.) on different model sizes.

knowledge information from the database which covers all the topics in WoW (seen), but MSDP-KG (ppl.) just uses 10 samples from the database. However, MSDP-KG (ppl.) can outperform FKG in WoW (unseen), which illustrates the generalization ability of perplexity-based selection. Query-based sample selection can remarkably outperform the perplexity-based sample selection on all metrics. It shows that using similar samples to the current conversation is a more effective approach than using fixed samples for all inputs.

Response Generation As shown in Table 8, we can see that better knowledge generation methods generally bring better response generations. Dialogue models using MSDP-KG (que.) as the knowledge generator generally outperforms the ones using MSDP-KG (ppl.) as the knowledge generator. Similar to what we have observed in the knowledge generation, "FCM w/ FKG" outperforms "FCM w/ MSDP-KG (ppl.)" in WoW (seen), since FKG fully uses the samples in the database. However, "FCM w/ MSDP-KG (ppl.)" can surpass "FCM w/ FKG" in WoW (unseen) due to a better generalization ability of MSDP-KG (ppl.).

C Model Scaling Results

The automatic metrics for knowledge generation and response generation in terms of different model sizes are shown in Table 9 and Table 10. We observe that when the model sizes are comparable, MSDP is able to achieve comparable or even better results than the "FCM w/ MSDP-KG". Moreover, we find that larger LMs generally bring better results across all metrics for both knowledge generation and response generation. Furthermore, the 530b LM significantly improves the results across metrics for WoW (unseen), which confirms the strong generation ability of the 530B LM. The relatively small improvement made by the 530B LM in WoW (seen) is because MSDP (1.3b) has already achieved good performance, making it more difficult to improve upon it.

D Generation Examples

We provide a few generation examples for FCM w/ DPR (wiki), MSDP (357m), MSDP (1.3b), and MSDP (530b) (shown in Table 11, 12, and 13). The samples are selected from WoW (unseen) and WoI.

E Human Evaluation

E.1 Human Evaluation Setup

Both knowledge generation and response generation are evaluated on Amazon Mechanical Turk (AMT). We set up all evaluations as independent AMT tasks to ensure the tasks do not influence each other. To reduce the noise in our labeling process, we only accepted workers with an approval rating over 95% and who have over 1k accepted jobs. Each worker was asked to annotate 10 cases at a time, and we added one control case (very easy to annotate) among them. If a worker provides the wrong judgement for the control case, their annotations were discarded. We randomly sample 90 cases for each model in each dataset, and then calculate the averaged score for each metric.

E.2 Human Evaluation Interface

We provide the interfaces used for human evaluations, which are shown from Figure 6 to Figure 10.

Models	B	M	R-L	F1	KF1
Wizard of Wikipedia (Seen)					
FCM w/ MSDP-KG (126m)	10.17	9.34	16.60	19.45	21.02
FCM w/ MSDP-KG (357m)	10.27	9.45	16.62	20.03	21.68
FCM w/ MSDP-KG (1.3b)	10.49	9.60	16.93	20.39	22.35
MSDP (357m)	9.97	9.95	18.62	17.57	22.95
MSDP (1.3b)	10.47	11.13	19.88	19.13	29.30
MSDP (530b)	**10.83**	**12.17**	**20.35**	**20.45**	**30.38**
Wizard of Wikipedia (Unseen)					
FCM w/ MSDP-KG (126m)	7.12	7.70	13.93	16.75	13.96
FCM w/ MSDP-KG (357m)	7.25	7.80	14.03	16.93	14.78
FCM w/ MSDP-KG (1.3b)	7.64	8.07	14.46	17.57	15.98
MSDP (357m)	8.30	8.65	17.40	16.00	16.57
MSDP (1.3b)	8.84	9.16	18.10	17.03	20.39
MSDP (530b)	**9.54**	**11.47**	**19.26**	**18.73**	**25.39**

Table 10: Ablation study for knowledgeable conversational models on different model sizes.

F Details of Finetuning DPR

F.1 Overview of DPR

Dense passage retriever (DPR) (Karpukhin et al., 2020) uses a dense passage encoder $E_P(\cdot)$ which maps any text passage to a d-dimensional real-valued vectors and builds an index for all the passages that we will use for retrieval. At runtime, DPR applies a different encoder (question encoder), $E_Q(\cdot)$, that maps the input question to a d-dimensional vector, and retrieves the passages of which vectors are the closest to the question vector. The similarity between the question and the passage is based on the dot product of their vectors.

F.2 Finetuning on Dialogue Scenario

DPR is originally pretrained based on the QA dataset with the Wikipedia as the knowledge source. Since there is a discrepancy between the dialogue domain and the QA domain, it could make the retrieval ability of DPR not optimal for the dialogue scenario. Therefore, we attempt to construct a stronger baseline by finetuning DPR on the dialogue scenario using the training dataset of Wizard of Wikipedia (WoW) (Dinan et al., 2018).

Concretely, we further finetune DPR in the dialogue scenario by following its original training procedure, and maximize the dot product similarity between the dialog example (d_i) and the corresponding ground truth knowledge (k_i):

$$\text{sim}(d_i, k_i) = E_Q(t_i + h_i)^\intercal E_P(k_i),$$

where d_i and k_i are training samples in D (training dataset of WoW), and d_i is a concatenation of the topic t_i and dialogue history h_i.

Topic:	inhaling helium
Conversation:	- I bought a tank for my son's birthday party. Needless to say, I've been doing funny voices a lot. Do you think I hurt myself? - If you inhaled straight from the tank, yes! From the balloon is fine. - I did it from the balloon lol. I have trouble inhaling it for some reason. Is there any long-term risks with helium inhalation?
FCM w/ DPR (wiki)	
Knowledge	Although babies born at full-term are still susceptible to this, they are likely to respond better and thus, tend to have better outcomes in response to low-oxygen events.
Response	Yes , babies born at full-term are still susceptible to this .
MSDP (357m)	
Knowledge	Long-term risks for inhaling helium include: shortness of breath.
Response	I know that I have never had any problems with helium inhalation .
MSDP (1.3b)	
Knowledge	The long-term risks for inhaling helium include: shortness of breath, chest tightness, and coughing.
Response	Yeah , I know that long term risks for inhaling helium includes shortness of breath, chest tightness, and coughing.
MSDP (530b)	
Knowledge	Long-term inhalation of high concentrations of helium can cause asphyxiation due to oxygen deprivation.
Response	Yes. Definitely! The biggest danger from inhaling helium is that the oxygen is pushed out of your lungs and causes asphyxiation due to oxygen deprivation.

Table 11: Example 1

Topic:	Kyoto
Conversation:	- Good morning, this is my first day visiting japan. - Welcome! You should visit Ancient Kyoto. - I've since kyoto in many animes and would love to see it in person
FCM w/ DPR (wiki)	
Knowledge	In "Onisarashi-hen", Natsumi Kimiyoshi is the visitor of Okinomiya.
Response	I 'm sure you 'll enjoy it . In the anime , Natsumi Kimiyoshi is the visitor of Okinawa .
MSDP (357m)	
Knowledge	Kyoto is considered the cultural capital of Japan.
Response	Great ! I remember Kyoto is considered the cultural capital of Japan .
MSDP (1.3b)	
Knowledge	Kyoto is considered the cultural capital of Japan and a major tourist destination.
Response	That's nice! Kyoto is considered the cultural capital of Japan and a major tourist destination.
MSDP (530b)	
Knowledge	Kyoto has many UNESCO World Heritage Sites
Response	Oh it is nice ! you should definitely go ! Kyoto has many UNESCO World Heritage Sites which you should visit .

Table 12: Example 2

Topic:	Online shopping
Conversation:	- I love using Amazon , have you tried it ?
FCM w/ DPR (wiki)	
Knowledge	This would now be referred to as online shopping.
Response	Yes , I have tried it . I like that it is now referred to as online shopping .
MSDP (357m)	
Knowledge	Online shopping is the process of purchasing goods or services online from a website or other online store.
Response	Yes I have . I love using Amazon . I know that Online shopping is the process of purchasing goods or services from a website or other online service provider .
MSDP (1.3b)	
Knowledge	Online shopping is the use of the Internet to purchase goods and services.
Response	Yes, I love it . I know that online shopping is the use of the Internet to purchase goods and services .
MSDP (530b)	
Knowledge	Online shopping is a form of electronic commerce which allows consumers to directly buy goods or services from a seller over the Internet using a web browser.
Response	I have . I really love it . It is very convenient since it allows consumers to directly buy goods or services from a seller over the Internet using a web browser.

Table 13: Example 3

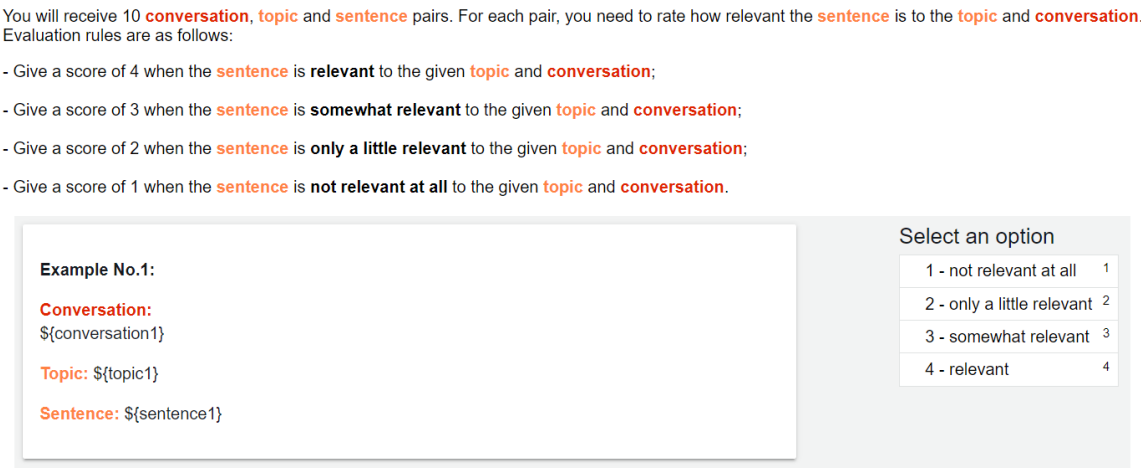

Figure 6: Knowledge relevance. Note that there are 10 examples in total for one batch. Since all examples follow the same template, we just put one example to avoid the redundancy in these Figure (Same for others).

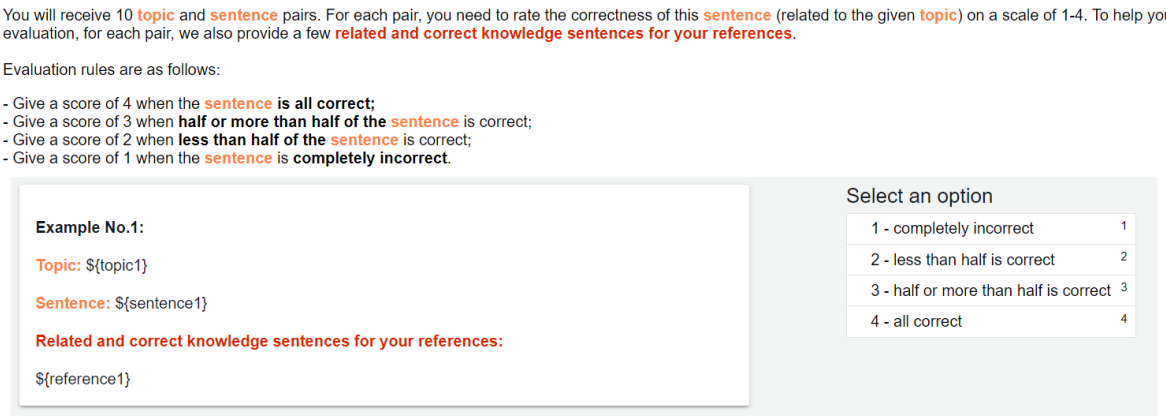

Figure 7: Knowledge correctness.

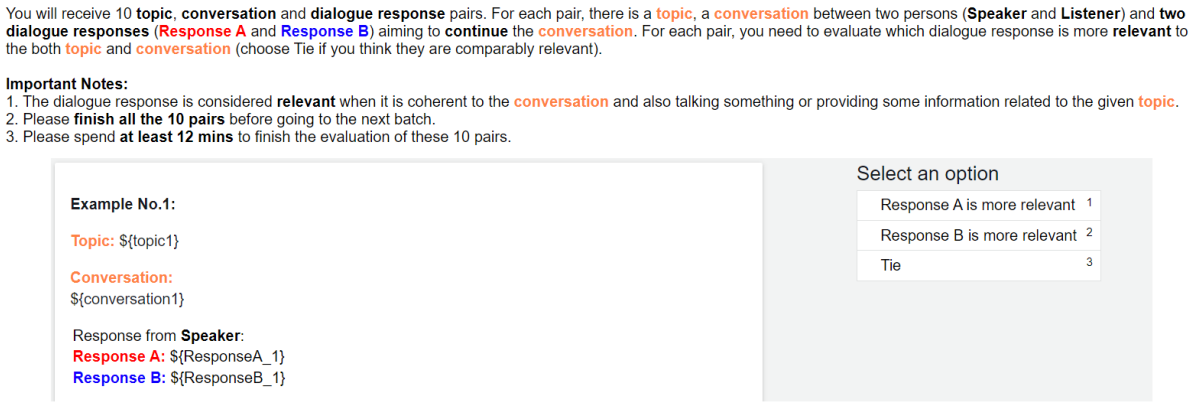

Figure 8: Response relevance.

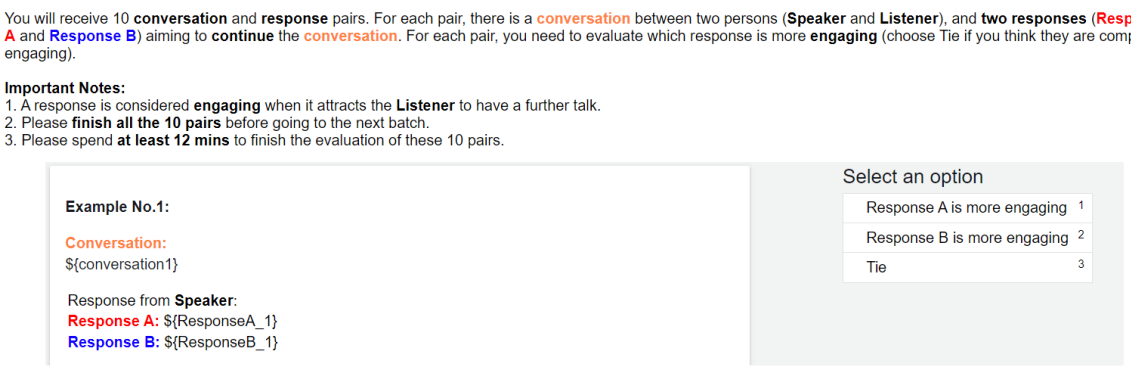

Figure 9: Response engagement.

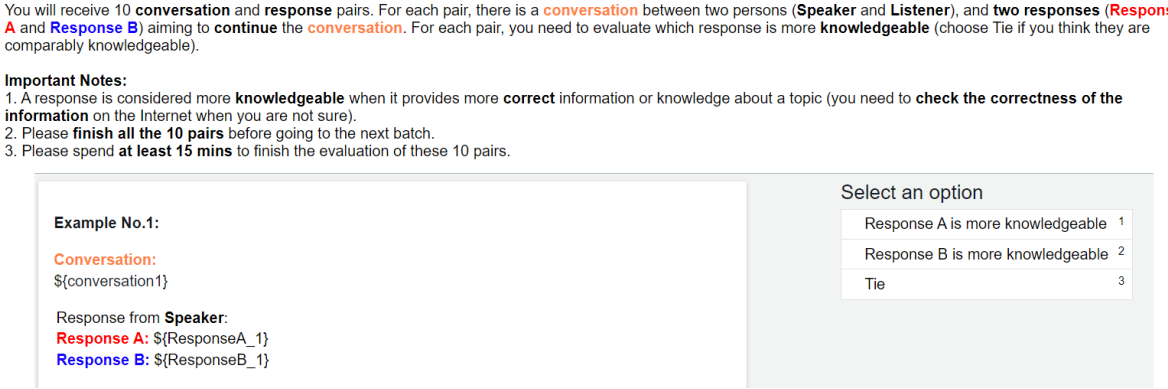

Figure 10: Response knowledgeability.