Pan More Gold from the Sand: Refining Open-domain Dialogue Training with Noisy Self-Retrieval Generation

Yihe Wang[1]*, **Yitong Li**[2,3]†, **Yasheng Wang**[2], **Fei Mi**[2]
Pingyi Zhou[2], **Xin Wang**[1], **Jin Liu**[1]‡, **Xin Jiang**[2], **Qun Liu**[2]
[1]School of Computer Science, Wuhan University
[2]Noah's Ark Lab, Huawei [3]Huawei Technologies Ltd.
{yihewang, jinliu, xinwang0920}@whu.edu.cn
{liyitong3, wangyasheng, feimi2, zhoupingyi, Jiang.Xin, qun.liu}@huawei.com

Abstract

Real human conversation data are complicated, heterogeneous, and noisy, from which building open-domain dialogue systems remains a challenging task. In fact, such dialogue data still contains a wealth of information and knowledge, however, they are not fully explored. In this paper, we show existing open-domain dialogue generation methods that memorize *context-response* paired data with autoregressive or encode-decode language models underutilize the training data. Different from current approaches, using external knowledge, we explore a retrieval-generation training framework that can take advantage of the heterogeneous and noisy training data by considering them as "evidence". In particular, we use BERTScore for retrieval, which gives better qualities of the evidence and generation. Experiments over publicly available datasets demonstrate that our method can help models generate better responses, even such training data are usually impressed as low-quality data. Such performance gain is comparable with those improved by enlarging the training set, even better. We also found that the model performance has a positive correlation with the relevance of the retrieved evidence. Moreover, our method performed well on zero-shot experiments, which indicates that our method can be more robust to real-world data.

1 Introduction

Open-domain dialogue is a long-standing problem in natural language processing and has aroused the widespread interest of researchers. Many approaches have been studied, and recently, generation models trained on large-scale data have gained more attention (Adiwardana et al., 2020; Roller et al., 2020; Xu et al., 2021; Madotto et al., 2021; Bao et al., 2019, 2020; Zhang et al., 2019b; Wang et al., 2020). Open-domain dialogue systems are born to deal with diverse domains, and naturally their training data, usually crawled from online resources such as Reddit and Twitter, are heterogeneous and contain utterances with many various topics, more freedom of topic shifting, and vague responses (Kummerfeld et al., 2018). As a result, directly building generation models from such data will be inefficient and usually requires "knowledge" during the training.

One common solution is to introduce external knowledge, usually, in a form of unstructured knowledge passages from Wikipedia (Dinan et al., 2018) or Internet articles (Komeili et al., 2021), and then, to build retrieval-augmented generation (RAG) methods to improve the response quality (Lewis et al., 2020; Izacard and Grave, 2020). However, this assumes knowledge-intensive scenarios, which are not suitable for general open-domain or robust to noise. According to our preliminary study, in the Reddit dataset, 43% of the dialogues are merely chitchat and cannot match "knowledge". Moreover, building such a knowledge-augmented dataset is very expensive as it relies on large amounts of high-quality human annotations w.r.t. knowledge grounding. And thus, they are limited in size, making it hard for a knowledge-retrieval method to generalize on scale.

Motivated by the above, we would like to investigate *can we have better ways of utilizing open domain data without introducing external resources?* To tackle the aforementioned problem, we found that the context from the other relevant dialogue sessions can still be very useful for dialogue generation. To utilize such unstructured contexts, we take inspiration from retrieval-augmented methods (Lewis et al., 2020). Differently, we retrieve useful dialogue context as evidence, build context-evidence-response triples for each dialogue turn, and treat open-domain generation as an evidence-aware generation task. Such that our model can

*Work done during internship at Noah's Ark Lab, Huawei
†Equal Contribution
‡Corresponding Author

Proceedings of the 29th International Conference on Computational Linguistics, pages 636–647
October 12–17, 2022.

learn to respond with useful grounding evidences. To retrieve evidences, we adopt similarity-based BERTScore (Zhang et al., 2019a), which leverages pre-trained contextual embeddings from BERT and matches words in two sentences by cosine similarity. It has been shown to correlate with human judgment on sentence-level and system-level evaluation. Although it was proposed as an automatic evaluation metric for text generation, due to the high correlation with human judgment, we consider it as a better off-the-shelf method to pick high-relevant evidences, compared with lexicon-based BM25.

By this, we show that current training methods which learn merely using context-response pairs have not fully unleashed the potential of training data and that our methods, only retrieving from the training data, can consistently improve the generation performance. We also perform zero-shot experiments, demonstrating that our method can be robust and generalized to different domains. Moreover, we found that adding extra retrieval data only (without training them) can still help the model gain performance, and it can even outperform traditional methods directly trained on that part of retrieval data. This proves our method is compatible with current methods with external knowledge.

Our contributions are summarized as follows:

- we explore a retrieval-generation training framework that can increase the usage of training data by directly considering the heterogeneous and noisy training data as the "evidence".
- We show that adding extra retrieval data while not training them can still gain performance benefits, even better than traditional training with the retrieval data attached.
- The proposed method performs well on zero-shot experiments, which indicates that our method can generalize well in real-world applications.

2 Related Work

Open-domain Dialogue System Open-domain dialogue system aims to perform chit-chat with people without the task and domain restriction. Adiwardana et al. (2020) proposed Meena, a multi-turn open-domain chatbot trained end-to-end on data mined and filtered from public domain social media conversations. Blender (Roller et al., 2020; Xu et al., 2021) learn to provide engaging talking points and listen to their partners, as well as display knowledge, empathy and personality appropriately, while maintaining a consistent persona. Adapter-bot (Madotto et al., 2021) explored prompt-based few-shot learning in dialogue tasks. Plato (Bao et al., 2019, 2020) introduced discrete latent variables to tackle the inherent one-to-many mapping problem in response generation. Zhang et al. (2019b) proposed DialoGPT which was trained on 147M conversation-like exchanges extracted from Reddit comment chains. Wang et al. (2020) introduced CDial-GPT, a pre-training dialogue model which is trained on a large-scale cleaned Chinese conversation dataset. Mi et al. (2022) built PANGU-BOT with relatively fewer data and computation costs by inheriting valuable language capabilities and knowledge from pre-trained language model.

Retrieval Augmented Generation Retrieval is a long-considered intermediate step in dialogue systems, and recently, it has been an intensively studied topic for neural models(Song et al., 2018; Pandey et al., 2018; Weston et al., 2018; Wu et al., 2019; Cai et al., 2019). Lewis et al. explored a fine-tuning recipe for retrieval-augmented generation, which combined pre-trained parametric and non-parametric memory for language generation. Izacard and Grave proposed Fusion-in-Decoder which encoded each evidence independently with the context when generative model processing retrieved passages. Li et al. (2022) explored how to effectively utilize information with different channel settings of FiD in multi-turn topic driven Conversations. Most of these works retrieved external knowledge, usually unstructured knowledge passages, such as Wizard of Wikipedia (Dinan et al., 2018), persona-chat (Zhang et al., 2018), and Wizard of Internet (Komeili et al., 2021). Moreover, Li et al. (2020) proposed a zero-resource knowledge-grounded dialogue model which bridged a context and a response as knowledge and expressed it as a latent variable.

3 Self-retrieval Method

We start from an open-domain dialogue dataset $\mathcal{D} = \{(c_i, r_i)\}_{i=1}^{N}$, where c_i denotes multi-turn dialogue context, consisting of dialogue utterances, and r_i represents the response.

Generally, we aim to build open-domain dialogue systems that retrieve useful dialogue responses (as evidences) from other sessions to help response generation. To tackle this problem, we

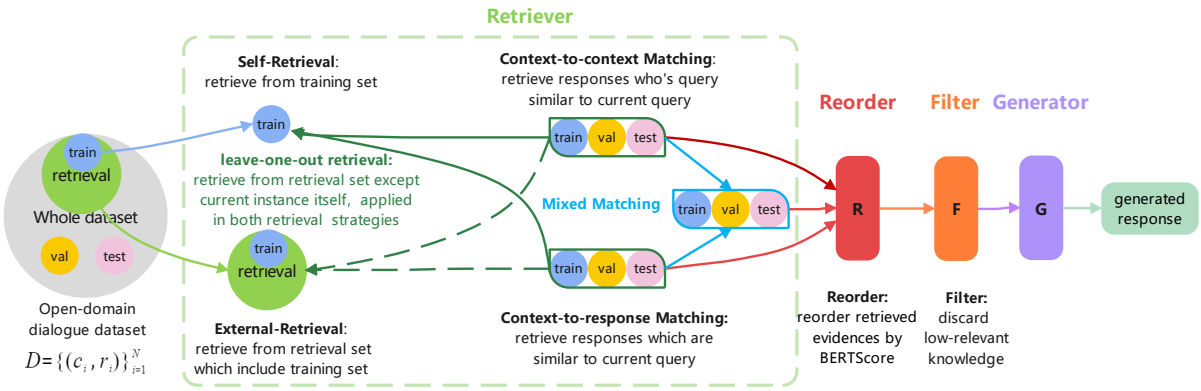


Figure 1: Overview of our self-retrieval approach as well as external-retrieval approach. In self-retrieval, our retriever first retrieves useful dialogue instances from the training dataset, which extends current data to context-evidence-response triples. And then, we adopt evidence-aware training models over the data with self-retrieval evidences.

proposed a two-step framework. The overview of our approach is shown in Figure 1.

1. Firstly, we extend an open-domain dialogue dataset with a *retriever*. Given the context of current dialogue turn c_i, the retriever $\mathscr{R}(e_{\{\cdot\}}|c_i)$ returns top-k relevant evidences as the *evidence set* $\mathcal{E}_i = \{e_{1:k}\}$ from a *retrieval set*. Note that different from existing knowledge-grounding methods, we do not introduce external data for our retriever, and we only consider retrieving evidence from the training data at hand. By that, we extend the dataset into context-evidence-response triples $\mathcal{D} = \{(c_i, \mathcal{E}_i, r_i)\}_{i=1}^{N}$.
2. Secondly, we adopt an evidence-aware generation model, which is a conditional language model to generate the response y given the context and the retrieved evidence $p(y|c, \mathcal{E})$. We investigate two widely used architectures, an auto-regressive GPT, and an encoder-decoder based language model T5.

Next, we introduce how to design an effective retriever in Section 3.1 and ways of implementing evidence-aware generation on the basis of state-of-the-art pre-trained generation models in Section 3.2.

3.1 Retrieve Dialogue Evidence

A variety of retrieval systems have been studied, including classic but effective bag-of-words system (Robertson et al., 1995) and up-to-date dense retriever, such as DPR (Karpukhin et al., 2020) and SparTerm (Bai et al., 2020). We utilized an off-the-shelf similarity based **BERTScore** to retrieve evidence (Zhang et al., 2019a).[1] BERTScore computes token similarity using pre-trained contextual embeddings rather than exact matches, which shows better coherent matching capability compared with human judgment. During the retrieval, for each context-response pair (c_i, r_i), we define the *retrieval set* by applying leave-one-out of the original training set $\mathcal{S} = \mathcal{D} - \{(c_i, r_i)\}$, to ensure the model cannot see the true response during generation.

We explore three retrieval strategies: context-to-context (C2C) retrieval, context-to-response (C2R) retrieval, and a MIX retrieval.

Context-to-context Matching C2C matches the context c_i of current dialogue and the context c_j from the retrieval set $\mathcal{S}$. And the evidence set of c_i is defined as:

$$\mathcal{E}_i^{\text{C2C}}(c_i, \mathcal{S}) = \underset{(c_j, r_j)\in\mathcal{S}}{\text{argmax}_{\text{K}}}\, \text{score}(c_i, c_j)\,,$$

where argmax_{K} means selecting top k corresponding responses $r_{1:k}$ as evidences $e_{1:k}$ with best matching score given by BERTSCORE.

Context-to-response Matching As the retrieval set contains the dialogue response, we also perform a Context-to-response (C2R) Matching. It is similar to C2C, while C2R directly matches the response in the retrieval set. In C2R, BERTSCORE computes the matching score based on the response r_j of the retrieval set.

$$\mathcal{E}_i^{\text{C2R}}(c_i, \mathcal{S}) = \underset{(c_j, r_j)\in\mathcal{S}}{\text{argmax}_{\text{K}}}\, \text{score}(c_i, r_j)\,.$$

Mixed Matching We observed that these two strategies, C2C and C2R, often obtain different results. Therefore, we complement the two retrieval

[1] We also did preliminary experiments over BM25 and it shows no significant differences for our findings.

sets of C2C and C2R with each other and combine them into a MIX retrieval set by re-ranking them using BERTScore. Finally, we take their responses as evidences:

$$\mathcal{E}_i^{\text{MIX}}(u_i, \mathcal{S}) = \text{argmax}_{\text{K}}\{\mathcal{E}_i^{\text{C2C}}, \mathcal{E}_i^{\text{C2R}}\}\,.$$

Filter During preliminary studies, we found that some retrieved evidences are not relevant to the current context. It is arguable that very few relevant evidences can be retrieved for some dialogue instances, and to study this we perform analysis in Section 4.5, where we study different sizes of the retrieval set to ensure more relevant evidences can be found. Undoubtedly, these low-relevant evidences are harmful to response generation. Therefore, we approach a simple filter to discard evidences with very low matching scores.

3.2 Evidence-aware Dialogue Generation

For generating more appropriate responses, our generator is a language model but also conditional on the retrieved evidence set.[2]

$$p(y|c_i, \mathcal{E}_i) = \prod_t p(y_t|c_i, \mathcal{E}_i, y_{<t})\,.$$

Generally speaking, it can be modeled by any auto-regressive or encoder-decoder generation architecture for open-domain dialogue. To demonstrate, we adopt both widely used architectures, i.e. a GPT-2 (Brown et al., 2020) and a Fusion-in-Decoder (FID; Izacard and Grave, 2020).[3]

GPT-2 GPT (Radford et al., 2019; Brown et al., 2020) is auto-regressive language model based on multi-head self-attention transformers (Vaswani et al., 2017). For our task, the model takes the dialogue context and the support evidences as the input, and then it generates the response. More precisely, for any instance $(c_i, \mathcal{E}_i, r_i)$, all retrieved evidences are concatenated before the dialogue context c_i, and the model directly generates the response y after c_i. We add special token `[p]` before each retrieved evidence passage, and following Wang et al. (2020), we add `[speaker1]`, `[speaker2]` to each utterance to indicate different speakers of muti-turn dialogue.

[2]Note that responses from the retrieval set are not directly trained by the language model, but used as the evidences at the input side only.

[3]We also experiment with T5 architectures via concatenating the context and evidences and decoding the response. Yet the performance does not significantly vary from GPT thus we do not report T5 in our main results.

Fusion-in-Decoder In our setups, we have multiple evidences for one instance, thus we adopt a slightly different model than the standard encoder-decoder T5 (Raffel et al., 2020). We use FID (Izacard and Grave, 2020), which was originally proposed for open-domain question answering. It considers encoding each evidence independently with context, so that these evidences will not affect each other on the encoder side, which is a better solution to encode multiple evidences. In detail, FID encodes a concatenation of the context c_i with each retrieved evidence e_j. It concatenates all the encoded hidden representations and then passed to the decoder for response generation. Slightly different from the original architecture, we add an additional passage that only encodes the dialogue context, in case one dialogue does not use any retrieved evidence (discussed in Section 4.5). Similarly, we add special tokens as we did for GPT-2.

4 Experiments

4.1 Datasets

To evaluate the performance of the proposed model, we conduct experiments on two publicly available dialogue datasets.

Reddit Dataset The `Reddit` dataset is extracted from comment chains scraped from Reddit spanning. Reddit discussions can be naturally expanded as tree-structured reply chains, since a thread replying to one thread forms the root node of subsequent threads. We derived the dataset from DialoGPT (Zhang et al., 2019b), and use their script to obtain and process the full dataset or demo dataset.[4] We report results on the demo dataset which comprises 770k multi-turn dialogue instances and is sufficient for our experiments.

Movie Dialog Dataset `Movie dialog` dataset collects movie discussions from real conversation taken directly under the *movie* subreddit (Dodge et al., 2015).[5] We discard instances with long turns or long sentences. In total, the movie dialog dataset has $940k$ dialogue sessions after preprocessing.

For both datasets, we randomly sample a training set of 100k samples, a validation set of 10k samples, and a test set of 10k samples. Data outside the above sets can be considered as retrieval resources. Noted that in our main experiments, the retrieval

[4]https://github.com/microsoft/DialoGPT.

[5]https://research.fb.com/downloads/babi/.

		Automatic Metrics					Human Evaluation			
Reddit		**PPL↓**	**F1↑**	**BLEU↑**	**Dist-1↑**	**Dist-2↑**	**Flue↑**	**Info↑**	**Relv↑**	**SSA↑**
GPT-2	BASELINE	31.3	5.3	3.4	65.4	96.7	3.0	2.9	2.8	46%
w. SR	BM25 MIX	28.1	6.6	4.2	73.5	**98.2**	3.4	3.3	3.4	51%
	BERTScore C2C	27.7	7.2	4.8	75.2	96.1	3.4	3.4	3.4	54%
	BERTScore C2R	27.9	7.0	4.7	75.0	96.4	3.3	3.4	3.3	53%
	BERTScore MIX	**27.1**	**7.8**	**5.4**	**76.1**	96.8	**3.5**	**3.4**	**3.5**	**55%**
T5	BASELINE	25.5	5.2	3.7	**95.7**	96.3	3.1	3.0	3.1	48%
FID w. SR	BM25 MIX	23.8	9.5	6.9	95.3	**97.2**	3.5	3.4	3.5	52%
	BERTScore C2C	23.3	9.9	7.3	94.3	94.7	3.5	3.5	3.4	54%
	BERTScore C2R	23.4	9.8	7.2	94.0	94.4	3.5	3.4	3.4	54%
	BERTScore MIX	**22.7**	**10.4**	**7.8**	95.6	96.5	**3.6**	**3.5**	**3.5**	**56%**
Movie		**PPL**	**F1**	**BLEU**	**Dist-1**	**Dist-2**	**Flue**	**Info**	**Relv**	**SSA**
GPT-2	BASELINE	25.6	5.4	3.3	64.3	96.0	3.0	2.9	2.8	47%
w. SR	BM25 MIX	22.7	6.7	4.2	71.4	96.1	3.4	3.3	3.3	52%
	BERTScore C2C	22.1	7.1	4.7	71.7	94.9	3.4	3.4	3.3	53%
	BERTScore C2R	22.3	7.0	4.7	72.0	94.3	3.3	3.4	3.3	53%
	BERTScore MIX	**21.6**	**7.6**	**5.2**	**73.4**	**96.2**	**3.5**	**3.4**	**3.4**	**55%**
T5	BASELINE	20.5	5.2	3.7	95.2	95.8	3.1	2.9	2.9	48%
FID w. SR	BM25 MIX	18.9	9.2	6.6	94.9	96.8	3.6	3.5	3.6	53%
	BERTScore C2C	18.4	9.5	7.0	94.4	95.6	3.6	3.6	3.5	55%
	BERTScore C2R	18.5	9.4	6.8	93.8	94.9	3.5	3.5	3.6	54%
	BERTScore MIX	**17.9**	**10.1**	**7.5**	**95.3**	**96.9**	**3.7**	**3.7**	**3.6**	**57%**

Table 1: Automatic and human evaluation of the in-domain setups over Reddit and Movie Dialog, using 8 evidences passages. GPT-2 and T5 are baselines. "w. SR" (with self-retrieval) indicate our methods. The best results are in **bold**.

set (for train/dev/test) is exactly the training set, where we only retrieve from the training set. And experimental results using a larger retrieval set are investigated and reported in Section 4.5, which involves more evidence than the training set.

4.2 Metrics

To evaluate response quality, we adopt both automatic metrics and human evaluations.

Automatic Metrics We deploy four commonly used automatic metrics for the dialogue generation, the perplexity (**PPL**), unigram overlap (**F1**), **BLEU**, and distinct 1,2 (**Dist-1,2**). **F1** and **BLEU** are commonly used to measure how similar the machine-generated responses is to referenced golden response (Miller et al., 2017; Papineni et al., 2002). **Dist-1,2** measure the diversity of the generated responses (Li et al., 2016).

Human Evaluations We perform human evaluation over the generated response. Following Song et al. (2021), we consider three conventional criteria: fluency (**Flue.**), informativeness (**Info.**), and relevance (**Relv.**). We recruit a team on Amazon Mechanical Turk consisting of several professional annotators, who are proficient in language tasks but know nothing about the models.[6] We sample 200 instances for each model's evaluation and each sample was evaluated by three people. Each criterion is rated on five scales, where 1, 3, and 5 indicate unacceptable, moderate, and perfect performance, respectively. We report the average Fleiss's kappa score (Fleiss and Cohen, 1973) on Reddit and Movie Dialogue, 0.49 and 0.45 respectively, indicating annotators have reached moderate agreement. We also consider Sensibleness and Specificity Average (**SSA**), which evaluates two aspects of responses: making sense and being specific (Adiwardana et al., 2020).

4.3 Implementation and Setup

As the context has a different number of turns, we use the latest utterance of dialogue context as the BERTScore query in practice, which can yield more consistent matching scores. Specifically, we compute $F1_{BERT}$ of context c_i of current dialogue and the corresponding context of every evidence. We use DEBERTA-XLARGE-MNLI (He et al., 2020) following the suggestion of authors.[7] The filter is used in all retrieval setups except the baselines.

[6] https://www.mturk.com/

[7] https://github.com/Tiiiger/bert_score

Speaker1: Why do you get to decide who has something to offer ?
Speaker2: He doesn't , he is entitled to his opinion , this is the internet and a forum discussion thread . People post their opinions not the truth .
Baseline Generation: Why have you already voted to make sure you for yourself to support yourself ?
Key Evidence 1: Everyone is entitled to an opinion , but those with experience in the area of discussion usually have more pertinent and accurate opinions than others .
Key Evidence 2: No you're entitled to your opinion . I'd just prefer an opinion that didn't contain a logical fallacy .
Our Generation: I agree with you. People are entitled to their opinion . I just posted my own opinion .
Ground Truth: I know , I was taking a round about way of trying to get him to questions his opinion .

Table 2: Examples of responses generated by baseline and our approach based on FID.

		Movie Dialogue → Reddit					Reddit → Movie Dialogue				
		PPL	**F1**	**BLEU**	**Dist-1**	**Dist-2**	**PPL**	**F1**	**BLEU**	**Dist-1**	**Dist-2**
T5	BASELINE	29.2	5.3	3.9	95.6	96.2	33.0	5.1	3.6	94.5	95.9
FID w. SR	C2C	26.1	9.2	6.8	95.9	97.2	27.3	8.8	6.7	95.8	96.7
	C2R	26.2	9.1	6.6	95.2	96.6	27.5	8.6	6.6	95.2	96.1
	MIX	**25.6**	**9.8**	**7.3**	**96.4**	**98.1**	**26.8**	**9.5**	**7.1**	**95.5**	**97.8**

Table 3: Automatic evaluation results of zero-shot experiments over Reddit and Movie Dialog with 8 retrieved evidence passages. BERTScore is used to retrieve. The best results are in **bold**.

We perform an in-domain evaluation over the two datasets. For each dataset, we adopt the proposed three self-retrieval (SR) method, C2C, C2R, and MIX, comparing against the GPT-2 and FID baselines. We experiments with different numbers of retrieval evidence passages (see Section 4.5). Note that FID degenerates to a standard T5 model without any evidence. We retrain our model based on the pretrained checkpoint of GPT-2,[8] and T5 checkpoint for FID.[9] We do model selection based on PPL over the validation set.

We additionally perform a zero-shot cross-domain evaluation for both datasets using FID.[10] In this setup, we only train our best in-domain FID model on one dataset and then directly test on the other, while the retrieval set for inference is the training set of the target domain. All other setups follow the in-domain experiments.

4.4 Results

In-domain Table 1 reports the overall in-domain experimental results. Overall, our self-retrieval methods achieve better performance consistently across almost all automatic and human evaluation metrics in terms of generating quality. For generation diversities (**Dist-1 and Dist-2**), our SR can still have comparable performance with the strong baselines. For both GPT-2 and FID, all three used matching strategies can improve the overall performance, and MIX consistently outperforms the other two. Comparing with GPT-2 and FID, two baselines achieve similar performance, while when adding our retrieved evidences, we observed FID based methods performance better, demonstrating the effectiveness of evidence-aware training of FID in modeling multiple evidence passages. We also illustrate the example generated by our approach and baselines in Table 2. Above all, these results demonstrate that our approach could utilize more of the dialogue data without introducing more data compared with the baselines.

Zero-shot Cross-domain Table 3 reports the results of zero-shot experiments using FID. Again, we find that our methods with evidence achieve better performance compared to the baselines without knowledge and MIX performs the best. This result indicates that our approach has good generalization and is robust to different datasets.

Overall, both in-domain and zero-shot results demonstrate our self-retrieval method can improve the performance of open-domain dialogue generation, and worth noting that our self-retrieval do not use any additional resources. This indicates our methods can unleash more potential of the dialogue data compared with the vanilla training methods.

4.5 Analysis

Retrieval Methods Table 1 shows the experimental results of different retrieval methods. We find that both methods achieve better results compared to baseline, which shows the generality of our self-

[8] https://huggingface.co/gpt2/tree/main
[9] https://huggingface.co/t5-small/tree/main
[10] We ensure there is no overlap between the two datasets.

		Reddit					Movie Dialog				
		PPL	**F1**	**BLEU**	**Dist-1**	**Dist-2**	**PPL**	**F1**	**BLEU**	**Dist-1**	**Dist-2**
GPT-2	BASELINE	31.3	5.3	3.4	65.4	96.7	25.6	5.4	3.3	64.3	96.0
	p1	28.3	6.9	4.7	74.5	95.8	22.8	6.8	4.6	71.3	94.2
	p2	27.9	7.1	4.9	74.2	95.6	22.5	7.1	4.8	71.6	94.8
SR	p4	27.5	7.4	5.1	75.1	96.3	22.1	7.3	5.0	72.8	95.3
	p8	27.1	7.8	5.4	76.1	96.8	21.6	7.6	5.2	73.4	96.2
	p16	**26.8**	**7.9**	**5.4**	**76.5**	**97.0**	**21.3**	**7.8**	**5.3**	**73.8**	**96.5**
T5	BASELINE	25.5	5.2	3.7	95.7	96.3	20.5	5.2	3.7	95.2	95.8
	p1	23.8	9.5	6.9	93.7	94.8	19.1	9.0	6.3	94.6	95.7
	p2	23.5	9.8	7.2	94.1	95.3	18.7	9.4	6.7	94.4	95.5
FID w. SR	p4	23.1	10.1	7.6	94.6	96.2	18.2	9.8	7.2	94.9	96.3
	p8	22.7	10.4	7.8	95.6	96.5	17.9	10.1	7.5	95.3	96.9
	p16	**22.4**	**10.6**	**7.9**	**95.9**	**98.2**	**17.7**	**10.3**	**7.6**	**95.5**	**97.0**

Table 4: Experimental results of different numbers of evidences used for generation using Reddit and Movie Dialog. p-k indicates the number of evidence passages used for generation. The best results are in **bold**.

Reddit		**PPL**	**F1**	**BLEU**
GPT-2	BASELINE	31.3	5.3	3.4
	RANDOM	31.4	5.4	3.4
w. SR	w/o FILTER	27.6	7.2	4.8
	w. FILTER	**27.1**	**7.8**	**5.4**
FID (T5)	BASELINE	25.5	5.2	3.7
	RANDOM	25.7	5.2	3.6
w. SR	w/o FILTER	23.3	9.8	7.2
	w. FILTER	**22.7**	**10.4**	**7.8**

Table 5: Effectiveness of the Filter.

retrieval method. We can also find that BERTScore performs better than BM25,[11] which indicates that BERTScore could be used to get better retrieval evidences.

Retrieval Strategies Table 1 also shows the experimental results of different retrieval strategies. We find that MIX perform better than context-to-context retrieval (C2C) and context-to-response retrieval (C2R), and the latter two methods show no significant difference. We thought that both C2C and C2R can retrieve useful evidences while from different aspects. And thus mixing them can yield more useful informative and relevant evidences and better performance as well.

Effectiveness of the Filter Table 5 shows the ablation study without using the filter during the retrieval step on Reddit. Here the finding is that experiment with the filter (w. FILTER), has better performance than experiments without it (w/o FILTER), as well as a setup using random evidences (RANDOM). These show that noisy evidences give no assistance, or even harm, to the model and that the necessity of discarding low-relevant evidence in our method.

Number of Retrieved Evidences We also carried out experiments with a different number of retrieved evidences. Table 4 reports the experimental results of using k evidences (p-k) for generation. We observe that experiment using more retrieved evidences (p16) performs better than experiments with fewer retrieved evidences (i.e. p1, p2, p4, p8). While the performance gap is getting smaller when increasing the evidence numbers. Considering the trade-off between efficiency and performance, we report results using 8 evidence as our main results, which is considered to be good enough. These results indicate that we can use more retrieved evidences to obtain better experimental results. In addition, supporting more information is significant for the generative model.

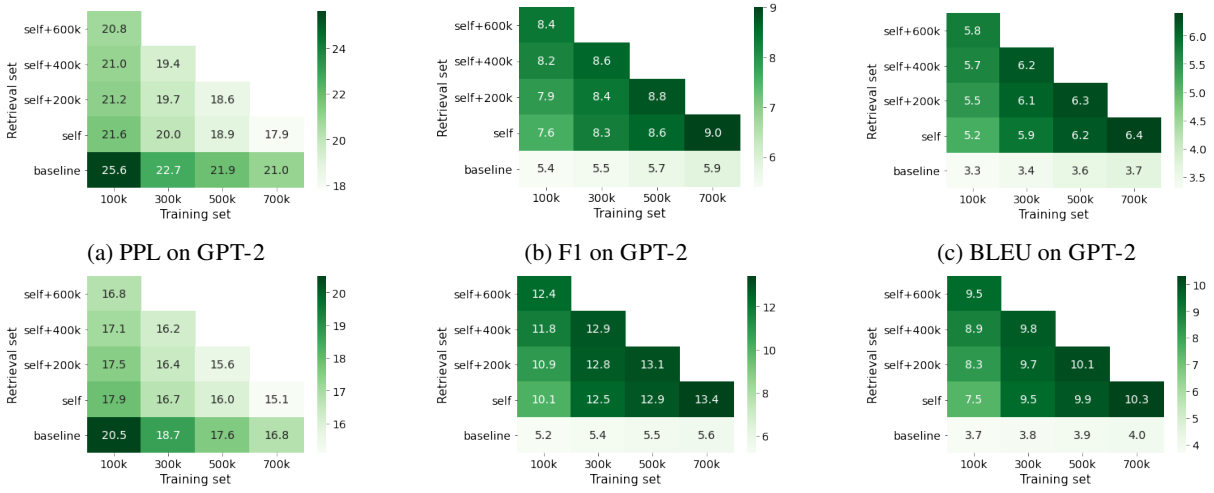
Self-retrieval vs. Extra Evidences We made the retrieval set exactly the same as the training set, denoted as the "self-retrieval (SR)" setup. One natural question is *can we use extra data for retrieval set?* To further understand this question and to validate the usefulness of our method, we carried out experiments with different sizes of the training set and retrieval set. Specifically, we experiment with additional setups by enlarging the retrieval sets, i.e. +200k, +400k, +600k, where "+" means extra data for retrieval sets, and we also adopt baselines with different training sizes of 100k, 300k, 500k, 700k (denoted before "+").[12]

Figure 2 shows the experimental results.[13] We

[11] We only report the mix results for BM25. Refer to the appendix for full results.

[12] Due to data size limitation, we did not occupy all setups.

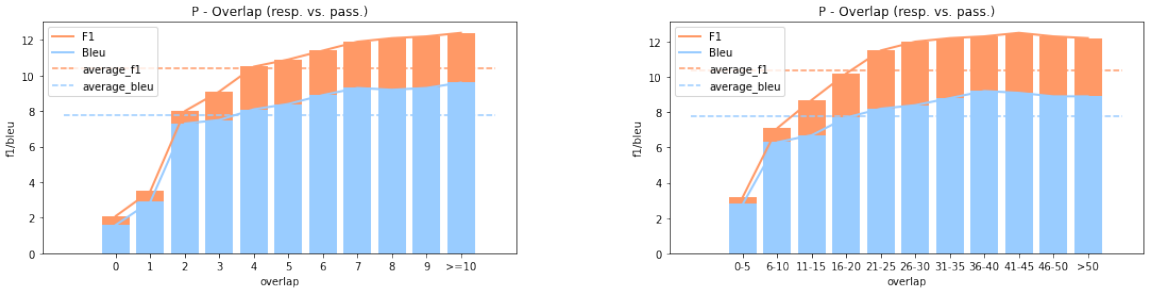
[13] We also report a detailed results using (100+600k) setup in Appendix A.1.



(a) PPL on GPT-2 (b) F1 on GPT-2 (c) BLEU on GPT-2

(d) PPL on FID (e) F1 on FID (f) BLEU on FID

Figure 2: Results of different sizes of training set and retrieval set on the `Movie Dialog` with 8 retrieved evidences. "Self" indicates the training set used for self-retrieval and "+" means adding extra data for retrieval.



(a) max setup over overlaps with bins = $\{0, \cdots, 9, \geq 10\}$ (b) sum setup over overlaps using bin size = 5

Figure 3: Performance by different overlaps between evidences and ground-truth responses over `Reddit`.

observe that experiments with larger retrieval sets achieve better results than those with small retrieval sets across different training sizes. We believe larger retrieval sets can introduce more relevant evidences, which brings performance gain for the model. Another interesting finding is that adding extra data for retrieval (100+600k, 300+400k, 500+200k) can outperform the baselines (700k) with extra data added via direct training. Also, under the same amount of total data (700k), leveraging more data for retrieval (100+600k, 300+400k, 500+200k) has approaching performance with the self-retrieval with full data (self, 700k). It indicates that our methods can increase the usage of the training data only in a retrieval way without directly training these responses, and our method has good generalization over the retrieval evidences.

Relevance of Evidence and Ground-truth To further study how our methods make sense, we study *how the relevance of the retrieved evidences and ground-truth response can influence the generation performance*. For each instance (c_i, r_i) which used n retrieval evidences $\mathcal{E}_i^{\text{MIX}} = \{e_1, e_2, ..., e_n\}$, we compute the number of overlapped words between the ground-truth r_i and each retrieved evidence. We study two setups by computing the overall $\text{overlap}(\mathcal{E}, r_i)$ using $\max$ and sum over the individual overlaps.

Figure 3 shows the results of these two setups. We observed that higher overlap leads to better performance. It indicates that high relevant retrieval evidences can help to generate better responses and low relevant knowledge are harmful, which is consistent with the findings in Section 4.5. Also, there are low-relevant evidences left in the retrieval step, which indicates that open-domain dialogue generation is still a difficult task, and better retrieval methods are required to further improve our generation performance.

5 Conclusion

In this paper, we propose a self-retrieval training framework for open-domain dialogue generation. Different from other knowledge-intensive tasks, our framework only retrieves relevant dialogue instances from the training data (which can be extended to a retrieval set) without the need to train them in the generation model. It is significant that we demonstrate that traditional training baselines underutilize the training data and our method can utilize more potential of data. We show that our method improves the robustness and generality of generative models as well as generate proper response for complicated human conversation. We also find that BERTScore can be used for better evidence retrieval. In future works, we would like to study better ways of evidence retrieval and evidence-aware training and we believe our approach can benefit to other NLP tasks, such as classification task.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive and insightful comments. This work was supported by the National Natural Science Foundation of China under Grants No. 61972290.

References

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

Yang Bai, Xiaoguang Li, Gang Wang, Chaoliang Zhang, Lifeng Shang, Jun Xu, Zhaowei Wang, Fangshan Wang, and Qun Liu. 2020. Sparterm: Learning term-based sparse representation for fast text retrieval. *arXiv preprint arXiv:2010.00768*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2019. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019. Skeleton-to-response: Dialogue generation guided by retrieval memory. 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 1219–1228.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2015. Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

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

Jonathan K Kummerfeld, Sai R Gouravajhala, Joseph Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros Polymenakos, and Walter S Lasecki. 2018. A large-scale corpus for conversation disentanglement. *arXiv preprint arXiv:1810.11118*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation

for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Jiatong Li, Bin He, and Fei Mi. 2022. Exploring effective information utilization in multi-turn topic-driven conversations. *arXiv preprint arXiv:2209.00250*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *arXiv preprint arXiv:2008.12918*.

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*.

Fei Mi, Yitong Li, Yulong Zeng, Jingyan Zhou, Yasheng Wang, Chuanfei Xu, Lifeng Shang, Xin Jiang, Shiqi Zhao, and Qun Liu. 2022. Pangubot: Efficient generative dialogue pre-training from pre-trained language model. *arXiv preprint arXiv:2203.17090*.

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

Gaurav Pandey, Danish Contractor, Vineet Kumar, and Sachindra Joshi. 2018. Exemplar encoder-decoder for neural conversation generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1329–1338.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. Bob: Bert over bert for training persona-based dialogue models from limited personalized data. *arXiv preprint arXiv:2106.06169*.

Yiping Song, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, Dongyan Zhao, and Rui Yan. 2018. An ensemble of retrieval-based and generation-based human-computer conversation systems.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *NLPCC*.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92.

Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7281–7288.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

A Appendix

A.1 Full results of Retrieving Extra data

We present a full results of enlarging the retrieval set to (100+600k) for both `Reddit` and `Movie Dialogue`, shown in Table 6. The training set is 100k as the same as the self-retrieval setup in main results. BM25 is used to retrieve.

`Reddit`		**PPL↓**	**F1↑**	**BLEU↑**
GPT-2	BASELINE	31.3	5.3	3.4
	C2C	28.0	6.2	3.8
GPT-2 w. DR	C2R	28.2	6.0	3.7
	MIX	**26.9**	**6.8**	**4.3**
T5	BASELINE	25.5	5.2	3.7
	C2C	23.6	9.6	7.2
FID w. DR	C2R	23.8	9.4	7.1
	MIX	**21.9**	**12.0**	**9.0**
`Movie`		**PPL↓**	**F1↑**	**BLEU↑**
GPT2	BASELINE	25.6	5.4	3.3
	C2C	22.5	6.0	3.7
GPT-2 w. DR	C2R	22.6	5.9	3.5
	MIX	**21.7**	**7.3**	**4.7**
T5	BASELINE	20.5	5.2	3.7
	C2C	19.2	9.1	6.9
FID w. DR	C2R	19.4	9.0	6.7
	MIX	**17.7**	**11.5**	**8.5**

Table 6: Automatic evaluations of the in-domain setups on the `Reddit` and `Movie Dialog` datasets with 8 evidences for retrieval. The best results are in **bold**.

A.2 Full results of Self-Retrieval

		Automatic Metrics					Human Evaluation			
Reddit		**PPL↓**	**F1↑**	**BLEU↑**	**Dist-1↑**	**Dist-2↑**	**Flue↑**	**Info↑**	**Relv↑**	**SSA↑**
GPT-2	BASELINE	31.3	5.3	3.4	65.4	96.7	3.0	2.9	2.8	46%
w. SR	BM25 C2C	29.4	6.1	3.8	69.3	95.6	3.2	3.0	3.1	49%
	BM25 C2R	29.7	6.0	3.6	68.4	95.3	3.2	3.1	3.1	50%
	BM25 MIX	28.1	6.6	4.2	73.5	**98.2**	3.4	3.3	3.4	51%
	BERTScore C2C	27.7	7.2	4.8	75.2	96.1	3.4	3.4	3.4	54%
	BERTScore C2R	27.9	7.0	4.7	75.0	96.4	3.3	3.4	3.3	53%
	BERTScore MIX	**27.1**	**7.8**	**5.4**	**76.1**	96.8	**3.5**	**3.4**	**3.5**	**55%**
T5	BASELINE	25.5	5.2	3.7	**95.7**	96.3	3.1	3.0	3.1	48%
FID w. SR	BM25 C2C	25.0	8.0	5.9	91.2	93.8	3.3	3.2	3.3	51%
	BM25 C2R	25.2	7.9	5.7	90.4	92.3	3.3	3.2	3.2	50%
	BM25 MIX	23.8	9.5	6.9	95.3	**97.2**	3.5	3.4	3.5	52%
	BERTScore C2C	23.3	9.9	7.3	94.3	94.7	3.5	3.5	3.4	54%
	BERTScore C2R	23.4	9.8	7.2	94.0	94.4	3.5	3.4	3.4	54%
	BERTScore MIX	**22.7**	**10.4**	**7.8**	95.6	96.5	**3.6**	**3.5**	**3.5**	**56%**
Movie		**PPL**	**F1**	**BLEU**	**Dist-1**	**Dist-2**	**Flue**	**Info**	**Relv**	**SSA**
GPT-2	BASELINE	25.6	5.4	3.3	64.3	96.0	3.0	2.9	2.8	47%
w. SR	BM25 C2C	23.5	6.1	3.8	66.9	93.9	3.2	3.1	3.1	51%
	BM25 C2R	23.5	6.0	3.7	67.8	92.7	3.2	3.0	3.1	50%
	BM25 MIX	22.7	6.7	4.2	71.4	96.1	3.4	3.3	3.3	52%
	BERTScore C2C	22.1	7.1	4.7	71.7	94.9	3.4	3.4	3.3	53%
	BERTScore C2R	22.3	7.0	4.7	72.0	94.3	3.3	3.4	3.3	53%
	BERTScore MIX	**21.6**	**7.6**	**5.2**	**73.4**	**96.2**	**3.5**	**3.4**	**3.4**	**55%**
T5	BASELINE	20.5	5.2	3.7	95.2	95.8	3.1	2.9	2.9	48%
FID w. SR	BM25 C2C	20.1	7.7	5.5	92.3	94.1	3.3	3.2	3.2	52%
	BM25 C2R	20.2	7.7	5.4	91.7	93.6	3.3	3.1	3.2	51%
	BM25 MIX	18.9	9.2	6.6	94.9	96.8	3.6	3.5	3.6	53%
	BERTScore C2C	18.4	9.5	7.0	94.4	95.6	3.6	3.6	3.5	55%
	BERTScore C2R	18.5	9.4	6.8	93.8	94.9	3.5	3.5	3.6	54%
	BERTScore MIX	**17.9**	**10.1**	**7.5**	**95.3**	**96.9**	**3.7**	**3.7**	**3.6**	**57%**

Table 7: Automatic and human evaluation of the in-domain setups over Reddit and Movie Dialog, using 8 evidences passages. GPT-2 and T5 are baselines. "w. SR" (with self-retrieval) indicate our methods. The best results are in **bold**.