CONVERSER: Few-Shot Conversational Dense Retrieval with Synthetic Data Generation

Chao-Wei Huang   Chen-Yu Hsu   Tsu-Yuan Hsu   Chen-An Li   Yun-Nung Chen

National Taiwan University, Taipei, Taiwan

f07922069@csie.ntu.edu.tw   y.v.chen@ieee.org

Abstract

Conversational search provides a natural interface for information retrieval (IR). Recent approaches have demonstrated promising results in applying dense retrieval to conversational IR. However, training dense retrievers requires large amounts of in-domain paired data. This hinders the development of conversational dense retrievers, as abundant in-domain conversations are expensive to collect. In this paper, we propose CONVERSER, a framework for training conversational dense retrievers with at most 6 examples of in-domain dialogues. Specifically, we utilize the in-context learning capability of large language models to generate conversational queries given a passage in the retrieval corpus. Experimental results on conversational retrieval benchmarks OR-QuAC and TREC CAsT 19 show that the proposed CONVERSER achieves comparable performance to fully-supervised models, demonstrating the effectiveness of our proposed framework in few-shot conversational dense retrieval.[1]

1 Introduction

Conversational information retrieval (CIR) has been an important area of research in recent years, aiming to retrieve relevant information from a large corpus of text in a conversational format. It has gained considerable interest due to its potential to deliver information in a natural format in response to a user's queries. Unlike traditional IR, CIR poses distinctive challenges, including its multi-turn and context-dependent nature, which require more nuanced approaches (Yu et al., 2021; Fang et al., 2022).

Dense retrieval methods have demonstrated their ability to understand the semantics of complex user queries and shown promising performance on open-domain retrieval (Karpukhin et al., 2020). One

of the major obstacles to conversational dense retrieval is the scarcity of training data, given the high cost and extensive time to collect high-quality information-seeking conversations (Adlakha et al., 2022). Previous work has explored various approaches to address this issue (Dai et al., 2022; Kim et al., 2022). However, most methods still rely on the assumption that a large amount of in-domain data is present and build data augmentation models upon it.

In this paper, we aim to develop a few-shot conversational dense retrieval model that can effectively retrieve relevant passages based on a small number of in-domain dialogues. To achieve this, we leverage the in-context learning capability of large language models (LLMs) to generate synthetic passage-dialogue pairs with few-shot demonstrations. Specifically, in-domain passages are sampled from the retrieval corpus, and dialogues are synthesized by asking LLMs to generate a series of queries based on a few examples. We also employ a self-consistency filtering mechanism to automatically discard inconsistent generated queries, ensuring the accuracy and reliability of the generations.

We conduct experiments on two benchmark datasets, including OR-QuAC (Qu et al., 2020) and TREC CAsT 19 (Dalton et al., 2019). The experimental results demonstrate that our proposed framework, CONVERSER, performs comparably to fully-supervised models that are trained on *thousands* of annotated dialogues while using only 6 examples at most. Furthermore, analyses show that CONVERSER rivals other data augmentation methods that utilize full in-domain datasets, demonstrating its effectiveness.

2 Related Work

Conversational Dense Retrieval Conversational dense retrieval poses a unique challenge in that the questions are context-dependent. Prior works have explored various modeling techniques for conver-

---

sational history to address this challenge (Huang et al., 2018; Choi et al., 2018; Yeh and Chen, 2019; Chiang et al., 2020). However, these works only examined the modeling ability for conversational question answering (CQA), where the relevant passages are provided.

More recently, Qu et al. (2020) proposed OR-ConvQA, which extends CQA to the open-domain setting where a retrieval module is required. ConvDR (Yu et al., 2021) utilizes an ad-hoc dense retriever and manually rewritten context-independent queries for training few-shot retrievers and rerankers, while our method does not require an ad-hoc model and additional annotation. Others have explored various methods for encoding conversational queries (Li et al., 2021; Fang et al., 2022; Wu et al., 2022; Liang et al., 2022), which are orthogonal to our work.

2.1 Synthetic Data Generation for Dense Retrieval

Due to the data-hungry nature of dense retrievers, synthetic data generation for dense retrieval has drawn considerable interest.

Previous works have worked on generating information-seeking conversations via transforming documents (Dai et al., 2022; Kim et al., 2022) or web search sessions (Mao et al., 2022). However, these methods all require training query generators with conversational data, which does not mitigate the data scarcity issue. Our method requires only 6 in-domain dialogues with their relevant passages and demonstrates comparable performance to models trained on thousands of manually annotated dialogues.

InPars (Bonifacio et al., 2022) and Promptagator (Dai et al., 2023) are the most closely related works to our method. They both proposed to generate synthetic queries with LLMs from few-shot examples, which achieved comparable performance to supervised methods in dense retrieval. Inspired by these works, our method further extends few-shot query generation to the conversational setting. We propose novel techniques for generating conversational queries and show that they are crucial to handle the unique challenges of conversational dense retrieval.

3 Proposed Method: CONVERSER

We propose few-shot conversational dense retrieval with synthetic data generation, CONVERSER,

which aims to generate synthetic conversational queries given few examples. More formally, given a conversational retrieval task T, its retrieval corpus $\mathcal{P}_T$, and k examples, we aim to generate synthetic conversational query-passage pairs $\{\hat{C}_1, \cdots, \hat{C}_n\}$ for training dense retrievers.

3.1 Few-Shot Conversational Query Generation

The core of our method is *few-shot query generation*. We leverage the in-context learning ability of LLMs (Brown et al., 2020) to generate conversational queries. Specifically, we start with k examples $\{C_1, C_2, \cdots, C_k\}$, where each C_i is a conversation represented as a series of query-passage pairs, $(q_i^1, p_i^1), \cdots, (q_i^{n_i}, p_i^{n_i})$, with n_i denoting the length of C_i. Using these examples, we construct the following template $\mathcal{T}$ as a few-shot demonstration for LLMs:

$$\left[(p_1^{n_1}, q_1^1, \cdots, q_1^{n_1}), \cdots, (p_k^{n_k}, q_k^1, \cdots, q_k^{n_k}) \right]$$

Note that we always choose the relevant passage that corresponds to the last query in the examplar, indicating that the last query $q_i^{n_i}$ is generated given $p_i^{n_i}$ and previous queries $q_i^1, \cdots, q_i^{n_i-1}$.

The generation process for a synthetic conversation starts with randomly sampling a passage $\hat{p}$ from the retrieval corpus, i.e., $\hat{p} \sim \mathcal{P}_T$. We concatenate the template and the sampled passage to form an input text sequence $[\mathcal{T}, \hat{p}]$. An LLM is employed for generating synthetic queries. It is expected to generate the first query $\hat{q}_1$ that is relevant to $\hat{p}$ based on the provided examples. We then append $\hat{q}_1$ to the input sequence, forming the input sequence for generating the next query $\hat{q}_2$, and so forth. We sequentially perform the generations for a conversation until a predefined number of turns is reached.

3.2 Two-Stage Generation

One unique characteristic of conversational queries is that the queries are *context-dependent* (Choi et al., 2018) except for the first query, which should be a self-contained query without any ambiguity. To address this difference, we propose to split the generations into two-stage: first query generation and follow-up query generation. When generating the first query for each conversation, we use an alternative template $\mathcal{T}_1 = [p_1^1, q_1^1, \cdots, p_k^1, q_k^1]$, which contains only the first queries and their relevant passages of the examples. We then replace
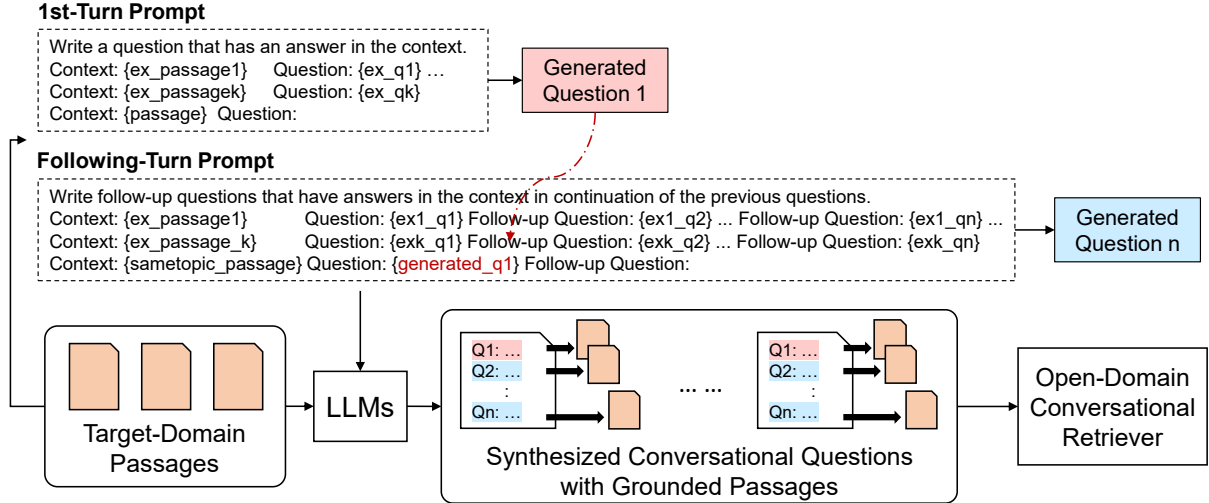
Figure 1: Illustration of our proposed framework.

$\mathcal{T}_1$ with $\mathcal{T}$ for generating all the follow-up queries. In practice, we found that this two-stage approach reduces the number of generated first queries that are not self-contained and thus ambiguous.

3.3 Passage Switching

In a conversation, relevant passages may vary for different queries. To this end, we incorporate passage switching into the generation process. We randomly replace the current passage $\hat{p}$ with a related passage $\hat{p}'$ in each turn with a probability p_{ps}. The LLM is expected to generate queries based on the new passage.

3.4 Consistency Filtering

The generation process sometimes generates queries that are nonsensical, degenerated, ambiguous, or not grounded by the given passage. We adopt a filtering mechanism via ensuring *round-trip consistency* (Alberti et al., 2019). We follow the procedure in Dai et al. (2023), where an initial retriever is trained on all synthetic query-passage pairs. For each synthetic pair $(\hat{q}, \hat{p})$, we use the initial retriever to retrieve the most relevant passages for $\hat{q}$ from $\mathcal{P}_T$. We keep the pair $(\hat{q}, \hat{p})$ only if $\hat{p}$ is in the top-k retrieved passages.

4 Experiments

To evaluate if our generated conversational questions can help train a conversational retriever, we conduct experiments on a conversational question answering dataset, OR-QuAC (Qu et al., 2020), and a conversational search benchmark, TREC CAsT-19 (Dalton et al., 2019).

4.1 Experimental Setup

We describe our experimental setup in the section. Additional details can be found in Appendix A.

Few-Shot Examples We manually select 6 examples for OR-QuAC and 5 examples for CAsT-19 and use the same set of examples in all experiments. Due to resource constraints, we use the remaining 15 conversations for evaluating on CAsT-19 without performing 5-fold cross-validation.

Generation We employ LLaMA-13B (Touvron et al., 2023) as our pretrained LLM, which is not instruction-tuned and is open to the research community. We use nucleus sampling (Holtzman et al., 2020) for decoding and set $top\_p = 0.95$, temperature $= 0.75$. We generate 427k turns (61k conversations) for OR-QuAC and 230k turns (32k conversations) for An example of generation results can be found in Section 5.

Retrieval Corpus We generate synthetic conversations based on the retrieval corpus for each task respectively. For OR-QuAC, we use the provided 11M passages from English Wikipedia. For TREC CAsT-19, we use the official passage collection, which consists of 8M webpage passages from MS-MARCO (Bajaj et al., 2016) and 30M Wikipedia passages from TREC-CAR (Dietz et al., 2017).

Model Details We follow the procedures from DPR (Karpukhin et al., 2020) to train our retrievers and use BERT-base as the pretrained model. We concatenate all previous queries and the current query as the input to the retriever. Additional details can be found in Appendix A.

Method	OR-QuAC			CAsT-19	
	MRR@5	R@5	MAP@10	MRR	NDCG@3
Supervised OR-ConvQA (Qu et al., 2020)	22.5	31.4	-	-	-
Supervised DPR	50.5	64.7	49.7	29.4	19.1
Few-Shot CONVERSER (Ours)	49.6	63.4	48.7	35.8	21.4

Table 1: Evaluation results (%). We report the result of OR-ConvQA from the original paper.

Method	OR-QuAC	
	MRR@5	R@5
OR-QuAC	50.5	64.7
WikiDialog (31k)	44.6	58.2
CONVERSER (31k)	46.8	61.5
- Two-Stage	45.1	59.9
- Consistency Filtering	45.2	59.8
- Passage Switching	45.6	58.1
- Only 1-Shot	42.1	55.2

Table 2: Results of ablation study. We use the identical training procedure and training data size for each experiment to make them comparable.

Baseline Systems

- **OR-ConvQA**: A supervised dense retriever trained on OR-QuAC (Qu et al., 2020).

- **DPR**: We train a DPR model (Karpukhin et al., 2020) on the training set of OR-QuAC for a fair comparison.

4.2 Main Results

Table 1 shows the experimental results. Note that both ConvDR and WikiDialog utilized multiple additional datasets and techniques, which are complementary to our method. On the OR-QuAC dataset, our proposed CONVERSER outperforms the supervised baseline OR-ConvQA by a large margin and performs comparably to the supervised DPR trained on OR-QuAC. This result demonstrates the effectiveness of our few-shot generation strategy, as our model trained on a synthetic dataset based on only 6 annotated examples can rival the performance of supervised DPR, which is trained on 4000 annotated dialogues.

On CAsT-19, CONVERSER outperforms supervised DPR, which is trained on OR-QuAC. This shows that our task-specific generation strategy can effectively synthesize conversational queries on a new task given a few examples of the new task. Our

proposed method provides better adaptability without requiring another supervised dataset as done in conventional transfer learning.

4.3 Ablation and Comparative Study

We conduct an ablation study on different settings of our proposed method, where we remove one component at a time to validate its effectiveness. We also compare our method with two datasets: OR-QuAC and WikiDialog (Dai et al., 2022). To ensure the results are comparable, we limit the size of every dataset to 31k turns, which is the same as the training set of OR-QuAC. The training process and hyperparameters are also identical for all datasets. For WikiDialog, we subsample the original WikiDialog dataset and use it to fine-tune a retriever, without further fine-tuning on OR-QuAC. The results are shown in Table 2.

Given the same number of synthesized turns, our CONVERSER outperforms WikiDialog, which requires supervised conversational datasets for training a query generator. This result validates the effectiveness of our proposed few-shot generation method. The ablation study demonstrates that all of our proposed components contribute to the improvement.

4.4 Effect of Generated Data Size

We explore the impact of the generated data size on the performance, where we conduct a series of experiments, systematically varying the number of generated turns used for training presented in Figure 2. It clearly illustrates that as the number of turns increases, the system's performance improves significantly. This finding highlights the crucial role of conversational data in enhancing the effectiveness of our model.

5 Qualitative Study

We present a generated example in Table 3 to perform qualitative analysis. WikiDialog is capable of generating follow-up questions. However, it often

Generated Question Example	
Title	Palazzo Primoli
Passage	Palazzo Primoli Palazzo Primoli is a Palazzo in Rome, Italy. It is owned by the city of Rome and houses several museums and collections. The palazzo was built in the seventeenth century. In 1901 Count Giuseppe Primoli (1851–1927) became its sole owner. He extended and partly modernised the palazzo with a new facade and entrance between 1901 and 1911. The Count's maternal grandparents were Charles Lucien Bonaparte and Zénaïde Bonaparte, and the Count brought together a collection of objects (now the Museo Napoleonico), documenting the relationship between Rome and the Bonaparte family. He also was an avid photographer. In 1927 Giuseppe Primoli donated the palazzo and his collections to the municipality of Rome. The Museo Napoleonico is located on the palazzo's ground floor, and the third floor is occupied by the Museo Mario Praz, the former residence of Mario Praz. Also located in the palazzo are the Count's library and photographic archive.
WikiDialog	Q1: What is known about Palazzo Primoli?
	Q2: Are there any other interesting aspects about this article?
	Q3: When was it built?
	Q4: Who were the previous owners of the palazzo?
	Q5: What was his role in the palazzo?
	Q6: What is the most interesting aspect of this article?
CONVERSER	Q1: What is the name of the building that houses the Museo Napoleonico?
	Q2: What is the Palazzo Primoli best known for?
	Q3: What year was it donated?
	Q4: What famous collection was donated?
	Q5: What other collections are housed in this museum?
	Q6: Who was the Count's maternal grandparents?
CONVERSER only 1 example	Q1: What is Palazzo Primoli in regards to Rome?
	Q2: What is the Palazzo Primoli's collection of objects?
	Q3: What is the Museo Napoleonico?
	Q4: What was the Count's library?
	Q5: What was the Count's photographic archive?

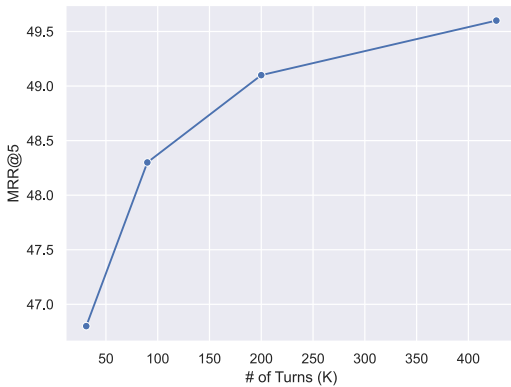Table 3: A qualitative example from WikiDialog, CONVERSER, and CONVERSER with only 1 example.

Figure 2: MRR@5 with regard to different number of generated turns on OR-QuAC.

generates generic queries, such as *Are there any other interesting aspects about this article*. On the other hand, CONVERSER with only 1 example suffers from a lack of diversity. Due to limited demonstrations, it generates queries that are very similar to the only example it is given. Our proposed CONVERSER can generate a context-independent first question and follow-up questions, demonstrating its effectiveness.

6 Conclusion

This paper introduces CONVERSER, a synthetic data generation method for training few-shot conversational dense retrievers. We leverage the in-context learning capability of LLMs and propose techniques that are designed for generating conversational queries. Experimental results demonstrate that our proposed CONVERSER achieves comparable performance to fully-supervised models while only requiring 6 annotated examples. Further analyses demonstrate that our method outperforms a fully-supervised data augmentation method. Future work could explore instruction-following LLMs, better filtering mechanisms, and synthesizing specialized data for conversational dense retrieval, such as query rewrites.

Acknowledgements

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2387–2392, New York, NY, USA. Association for Computing Machinery.

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

Ting-Rui Chiang, Hao-Tong Ye, and Yun-Nung Chen. 2020. An empirical study of content understanding in conversational question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7578–7585.

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

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4558–4586. PMLR.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. Cast 2019: The conversational assistance track overview. In *TREC*.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. Trec complex answer retrieval overview. In *TREC*.

Hung-Chieh Fang, Kuo-Han Hung, Chen-Wei Huang, and Yun-Nung Chen. 2022. Open-domain conversational question answering with historical answers. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 319–326, Online only. Association for Computational Linguistics.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2018. Flowqa: Grasping flow in history for conversational machine comprehension. In *International Conference on Learning Representations*.

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, Online. Association for Computational Linguistics.

Gangwoo Kim, Sungdong Kim, Kang Min Yoo, and Jaewoo Kang. 2022. Generating information-seeking conversations from unlabeled documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2362–2378, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yongqi Li, Wenjie Li, and Liqiang Nie. 2021. A graph-guided multi-round retrieval method for conversational open-domain question answering. *arXiv preprint arXiv:2104.08443*.

Tingting Liang, Yixuan Jiang, Congying Xia, Ziqiang Zhao, Yuyu Yin, and Philip S Yu. 2022. Multifaceted improvements for conversational open-domain question answering. *arXiv preprint arXiv:2204.00266*.

Kelong Mao, Zhicheng Dou, Hongjin Qian, Fengran Mo, Xiaohua Cheng, and Zhao Cao. 2022. ConvTrans: Transforming web search sessions for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2935–2946, Abu

Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.

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

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf, and Gaurav Singh Tomar. 2022. CONQRR: Conversational query rewriting for retrieval with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10000–10014, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yi-Ting Yeh and Yun-Nung Chen. 2019. FlowDelta: Modeling flow information gain in reasoning for conversational machine comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 86–90, Hong Kong, China. Association for Computational Linguistics.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 829–838.

A Implementation Details

Generation Text generation with language models often results in degeneration, i.e., repeating the same text sequence. Hence, we heuristically filter out degenerated generations. Initially, we examined the generation quality of LLaMA-7B. However, it showed an increased amount of degeneration and queries of lower quality. We have also tried several open-source instruction-tuned LLMs. To our surprise, these models failed to generate conversational queries given instructions, with or without few-shot examples. Using instruction-tuned LLMs for conversational query generation could be a direction for future exploration. Generations are conducted on 2 NVIDIA V100 GPUs. Generating one conversation takes roughly 10 seconds on a single GPU.

Training Details All retrievers are trained with a batch size of 64 queries. We use in-batch negatives as it is found to be important (Karpukhin et al., 2020). We train all retrievers for 10 epochs with a learning rate of 2e-5. To reduce GPU memory consumption, we use the DPR implementation with gradient cache (Gao et al., 2021), enabling larger batch size. The training process is done on 4 NVIDIA 2080Ti GPUs.

Evaluation Details We evaluate the models on the test sets of the evaluation datasets. There are 20 conversations for evaluation in CAsT-19. Previous work has conducted 5-fold cross-validation to address the lack of training in CAsT-19. However, due to resource constraints, we could not run generations for 5 different sets of examples. Hence, we manually select 5 conversations for building the few-shot examples and use the remaining 15 conversations for evaluation.

We report the most commonly-used evaluation metrics on each dataset: **MRR@5**, **R@5**, and **MAP@10** for OR-QuAC, and **MRR** and **NDCG@3** for CAsT-19.