out the CE loss in Eq. (3) for unused QR labels in training its initialized T5QR model, and we use the concatenated dialogue context as the BM25 input to obtain weak gold and hard negative passages for each training example, instead of using human rewrites (see details in Section 3.2). Figure 3 plots the curve of MRR on the overall QReCC test data using DE as the retriever versus the percentage of QR labels used for training. We see that CONQRR can already significantly outperform T5QR with even 0% or 1% of QR supervision.

The slight difference in performance for the 100% QR label case with respect to Table 3 is due to the different mechanism (using human rewrite vs. the dialogue context) for choosing the positive and hard negative passages for RL training. Performance of the RL and mixed loss are similar when there is little supervision, roughly tracking the trends of the T5QR model that it is initialized with. The finding that performance degrades for the mixed loss with 100% supervision may be due to a mismatch in the CE and RL losses as minimizing the CE loss does not directly optimize the retrieval performance. T5QR is more sensitive to QR supervision but also does not require many QR labels for training, as its curve becomes flattened after 1% supervision. We see similar trends with other metrics and BM25 (see Appendix A.4).

Effects of Topic Shift & Human Rewrites We hypothesize that a context involving a topic shift will present the greatest challenges for conversational passage retrieval. To explore this factor, we split the QReCC data into topic-concentrated and topic-shifted subsets as follows. A test example (with at least one previous turn) is considered *topic-concentrated* if the gold passage of the current question comes from a document that was used in *at least one* previous turn. In contrast, a test example (with at least one previous turn) is considered *topic-shifted* if the gold passage of the current question comes from a document that was *never* used in any previous turn. There are about 4.7k and 1.1k examples in the topic-concentrated and topic-shifted subsets, respectively. We compare the retrieval performance of different retriever inputs: dialogue context (which uses the concatenated dialogue history without QR), the predicted rewrite from T5QR and CONQRR with two loss alternatives, and the human rewrite. Table 4 shows that the dialogue context outperforms even the human rewrite on the topic-concentrated set by 22% and 17%, averaging

Input	IR	Topic-Concentrated MRR	R10	R100	Topic-Shifted MRR	R10	R100
Dial Context	BM25	**0.620**	**81.4**	**94.9**	0.154	39.1	68.6
T5QR	BM25	0.352	54.4	84.0	**0.252**	45.1	79.1
CONQRR (mix)	BM25	0.419	63.1	91.2	**0.252**	**45.9**	**82.1**
CONQRR (RL)	BM25	0.444	66.2	90.3	0.233	44.5	78.4
Human Rewrite	BM25	0.440	66.7	98.8	0.318	56.7	98.4
Dial Context	DE	**0.551**	**78.1**	**93.2**	0.179	35.7	61.4
T5QR	DE	0.353	55.7	75.4	0.329	50.8	69.2
CONQRR (mix)	DE	0.404	63.8	83.4	**0.334**	**53.2**	72.6
CONQRR (RL)	DE	0.445	69.3	87.8	0.303	50.4	**73.3**
Human Rewrite	DE	0.424	65.5	84.5	0.397	61.0	79.8

Table 4: Performance of using different retriever inputs for *Topic-Concentrated* or *Topic-Shifted* examples.

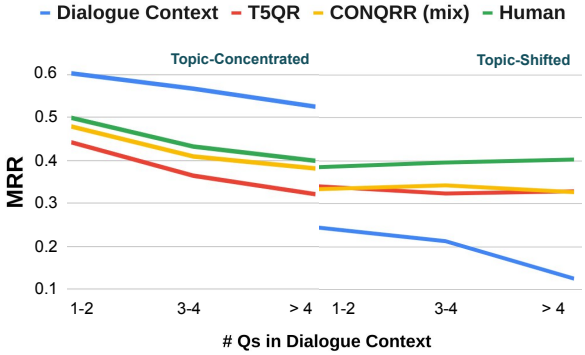

Figure 4: MRR versus the number of questions in the dialogue context, with DE as the retriever.

over three metrics, for BM25 and DE respectively, which shows the *limitation of human rewrites*. We also see that CONQRR (RL) surpass the human rewrite on the topic-concentrated set on MRR for BM25 and all three metrics for DE.

However, for the topic-shifted set, the human rewrite outperforms the dialogue context by 52% and 61%, averaging over three metrics, on BM25 and DE, respectively. The predicted rewrite by CONQRR (mix) outperforms the dialogue context by 30% and 44% on BM25 and DE, respectively. Therefore, compared with dialogue context, QR has great value in the aspect of *robustness to topic shifts*. When comparing with human rewrites, we also see room for improvement for QR models.

These observations are *largely unexplored* in previous work, and they motivate our work on the task of QR for conversational passage retrieval in general, and optimizing directly towards retrieval.

Effect of Dialogue Context Length Figure 4 shows the MRR score on topic-concentrated and topic-shifted subsets with DE as the retriever for various dialogue context lengths. Dialogue context lengths are grouped into 1-2, 3-4 and $\geq$ 4 pre-

Dialogue Context	*Q*: What were **John Stossel**'s most popular publications? *A*: Give Me a Break: How I Exposed Hucksters, Cheats, and Scam Artists and Became ... ... *Q*: What was the response?	*Q*: What were some notable live performances at the Buena Vista Social Club? *A*: **Ibrahim Ferrer and Rubén González** ... ... *Q*: What other live performances are important?
Gold Passage	**Stossel** has written three books. Give Me a Break: ... It was *a New York Times bestseller for 11 weeks* ...	The first performances ... **Ibrahim Ferrer and Rubén González** performed together ... *a 1999 Miami performance* ...
CONQRR (mix)	What was the response to **John Stossel**'s book, Give Me a Break? (Rank=2)	What other live performances at the Buena Vista Social Club are important besides **Ibrahim Ferrer and Rubén González**? (Rank=2)
T5QR	What was the response to the book Give Me a Break? (Rank >100)	What other live performances are important at the Buena Vista Social Club? (Rank=18)
Human	What was the response to Give Me a Break: How I Exposed Hucksters, Cheats, and Scam Artists and Became the Scourge of the Liberal Media? (Rank >100)	What other live performances of the Buena Vista Social Club are important? (Rank=17)

Table 5: Examples of predicted rewrites and the gold passage ranks by using them as the DE retriever input. *The gold answer is italicized in the gold passage.*

QR Model	QuAC-Conv L	QuAC-Conv % OL	NQ-Conv L	NQ-Conv % OL	TREC-Conv L	TREC-Conv % OL
T5QR	10.9	35.8	8.9	40.4	8.2	37.8
Ours (mix) w/ BM25	12.1	37.2	9.5	42.1	8.5	38.8
Ours (RL) w/ BM25	11.2	40.2	10.1	44.6	9.4	39.4
Ours (mix) w/ DE	12.1	37.2	9.6	41.7	8.7	39.1
Ours (RL) w/ DE	28.2	51.1	21.7	55.8	18.3	44.3
Human	12.1	37.2	9.3	43.0	8.4	41.7

Table 6: Average number of tokens (L) and the percentage of overlapping tokens (OL) with the gold passage(s) in output rewrites.

vious utterances (including the current question). For topic-concentrated conversations, all compared models have similar robustness to the dialogue context length and CONQRR (mix) is slightly more robust than T5QR. For topic-shifted conversations, both QR models and human rewrites show little drop or even an increase in performance as the context length gets longer. In contrast, the robustness of the dialogue context worsens with longer contexts, which confirms the importance of QR discussed above. We have similar observations for other metrics as well as for the BM25 retriever.

Quantitative Attributes of Rewrites Table 6 shows the average number of tokens per rewrite, and the percentage of overlapping tokens (excluding stopwords) between the rewrite and the gold passage(s). CONQRR generally generates longer rewrites with more overlapping tokens with gold passage(s), compared with T5QR. With DE as the retriever, CONQRR (RL) generates more than double the length of T5QR, CONQRR (mix) and even human rewrites. We show in Appendix A.4 that T5QR underperforms CONQRR even when we make it generate rewrites of similar lengths by applying a brevity penalty (Wu et al., 2016).

Rewrite Quality Analysis and Examples In order to understand why rewrites generated by CONQRR lead to better retrieval performance and even sometimes outperform human rewrites,[11] we sampled 50 examples where CONQRR (mix) leads to better ranking of gold passages than human rewrites (using DE retriever). We notice that 70% of CONQRR generated rewrites contain additional context and (correct) information when compared to human rewrites. The remaining 30% contain alternative or less context information than human rewrites. In such cases, potentially because the information in human rewrites is less relevant to gold passages, it led to a lower gold passage rank. Overall, these CONQRR rewrites are as fluent as human rewrites and contain no major misinterpretation of the dialogue context. Table 5 shows two examples of generated rewritten queries of T5QR and CONQRR (mix) trained with DE in the loop, as well as the human rewrites. In the left example, the CONQRR rewrite includes an entity "John Stossel" that is mentioned in the gold passage but not included by rewrites from T5QR or Human. Thus, even if the human rewrite is longer by containing the book's full name, CONQRR enables more efficient retrieval with a partial book name along with its author name. In the right example, CONQRR generates a longer rewrite containing richer contextual information. We have similar observations for BM25 and put more examples in Appendix A.4.

For error analysis, we sampled another 50 examples where CONQRR (mix) leads to worse ranking of gold passages than human rewrites with DE. All were deemed fluent. We found in most of these cases, CONQRR rewrites contain less context than human rewrites (56%) or additional information with a misinterpretation of the user request (34%).

[11] This is only for analysis purposes. Note that the goal of our predicted rewrites is to improve retrieval performance instead of directly being used by end users.

See examples in Appendix A.4 due to space limit.

5 Conclusion and Future Work

To summarize, we introduce CONQRR to address query rewriting for conversational passage retrieval with an off-the-shelf retriever. Motivated by our analysis showing both the limitations and utility of human rewrites, which are unexplored by prior work, we adopt RL with a novel reward to train CONQRR directly towards retrieval. As shown, CONQRR is the first QR model that can be trained adaptively to any off-the-shelf retriever, and achieves state-of-the-art retrieval performance on QReCC with conversations from 3 different sources. It shows better performance with zero QR supervision when compared with strong supervised baselines trained with full QR supervision.

A direction for future work includes leveraging QR to facilitate other tasks like question answering and response generation in a full CQA system, as well as sentence rewriting in a document (Choi et al., 2021). Future investigation is needed to explore conversations with other discourse relations like asking for clarifications besides alternating questions and answers in current CQA datasets.

Limitations

We show in Section 4.3 (Table 4) that compared to directly use dialogue context without QR, a QR model has great value in robustness to topic shifts when used with an off-the-shelf retriever. However, if most conversations of interest are topic-concentrated, we show that using the dialogue context itself may already work well. Although we focus on the *fixed retriever* setting in this work, we illustrate in Table 7 in Appendix A.4, that if the downstream retriever is *allowed to be fine-tuned*, our best QR model CONQRR (mix) underperforms compared to the dialogue context in both topic-concentrated and topic-shifted scenarios, and thus the benefits of QR as an intermediate step require further justification in that setting. Nevertheless, the table still shows that human rewrites have an advantage on topic-shifted conversations over dialogue contexts. Therefore, it would be interesting for follow-up studies to investigate the design of a QR model that reaches close performance with human rewrites on topic-shift scenarios with a fine-tunable retriever. Then, combining the dialogue context with the rewritten query for retrieval may help further improve the overall retrieval performance.

The training time of CONQRR is longer than fine-tuning a DE retriever of a similar model size (9 vs 2 hours) because for each training step of CONQRR, CONQRR needs to do autoregressive decoding to get greedily decoded and sampled q and q_s. However, re-indexing passages after fine-tuning the retriever can be very time-consuming (about 24 hours) and memory-consuming. In addition, unlike DE, CONQRR can also be used for any blackbox retriever such as search engines that are infeasible to fine-tune or be replaced.

Another downside of QR is that for out-of-domain and topic-shifted scenarios, QR may still require additional labels to achieve robust performance. Although we show that CONQRR (RL) initialized with T5 does not require QR labels and can work well on the overall QReCC test set, CONQRR (RL) does show worse robustness to out-of-domain and topic-shifted examples when compared with CONQRR (mix). Therefore, training a more robust CONQRR model may still require additional annotation efforts to collect human rewrites.

CONQRR has only been tested on the standard CQA dialogue format of alternating questions and answers. To facilitate more practical use cases with more diverse dialogue acts or discourse relations (e.g., the agent asks a clarification question to the user), further investigation is needed.

Ethical Considerations

Our work is primarily intended to leverage query rewriting (QR) models to facilitate the task of conversational passage retrieval in an open-domain CQA system. Retrieving the most relevant passage(s) to the current user query in a conversation would help to generate a more appropriate agent response. Predicted rewrites from our QR model are mainly intended to be used as *intermediate* results (e.g., the inputs to the downstream retrieval system). They may also be useful for interpretability purposes when a final response does not make sense to the user in a full CQA system, but that introduces a potential risk of offensive text generation. In addition, to prevent the retriever from retrieving passages from unreliable resources, filtering of such passages in the corpus should be performed before any practical use.