only. During the RL training of CONQRR, due to the complexity of applying Pyserini to calculate rewards on-the-fly, we instead use a Pyserini approximate called BM25-light. The only differences between them are that BM25-light (1) uses T5's subword tokenization instead of whole word tokenization and (2) does not use special operations (e.g., stemming) as applied in Pyserini. After training, we still run inference and report retrieval performance on BM25. Pyserini simply encodes the whole query input and each passage without truncating. We set maximum query and passage length as 128 and 2000 for BM25-light, but only less than 0.1% cases require truncation with these thresholds.

For the dual encoder, the maximum query or passage length is 384. The average passage length is 378, but we observe performance drop by further increasing the maximum length for the dual encoder.

A.2 Additional Data Details

QReCC reuses questions in QuAC and TREC conversations and re-annotates answers. For each NQ-based conversation, they only use one randomly chosen question from NQ to be the starting question and then annotate the remaining conversation. In total, there are 63k, 16k and 748 question and answer pairs in the three subsets QuAC-Conv, NQ-Conv, TREC-Conv respectively, where TREC-Conv only appears in the test set. The original data is only divided into train and test sets. We randomly choose 5% examples from the train set to be our validation set.

In some conversations from QuAC-Conv, the first user query is ambiguous as it depends on some topical information from the original QuAC dataset. Therefore, in order to fix this issue, we follow Anantha et al. (2021) to replace all first user queries in QReCC conversations with the their corresponding human rewrites.

QReCC is a publicly available dataset that was released under the Apache License 2.0 and we use the same task set-up proposed by the original QReCC authors.

A.3 Additional Evaluation Details

Some agent turns in QReCC do not have valid gold passage labels,[13] and the (provided) original evaluation script assigns a score of 0 to all such examples. Their updated evaluation script calculates the scores by removing those examples from the evaluation set (roughly 50%), which results in 6396, 1442 and 371 test instances for QuAC-Conv, NQ-Conv and TREC-Conv, respectively. This leads to a total of 8209 test instances in QReCC. We use the *updated* evaluation script for most of our experiments, except that we also use the *original* version for calculating scores in Table 2 to compare with their reported QReCC baseline results. We note that these two evaluation scripts only differ by a scaling factor so they should lead to the same conclusions regarding model comparisons.

A.4 Additional Analysis

Lower Recall@100 with DE Previous work (Karpukhin et al., 2020) shows that DE retrievers generally lead to better recall scores than BM25. However, in Table 3, we observe that across all subsets, the best MRR and Recall@10 results are consistently from DE, whereas BM25 has better Recall@100 scores. One reason to explain the observation difference is that we use an *off-the-shelf* retriever for our retrieval task while most previous work that compares BM25 and DE focuses on fine-tuning the DE model. Without being fine-tuned, a DE model may be more vulnerable to domain shift than BM25. On the other hand, prior work (Luan et al., 2021) proves that a DE model's performance would drop as the passage length increases. In the QReCC dataset, the average passage length is 378, which is relatively long (Luan et al., 2021).

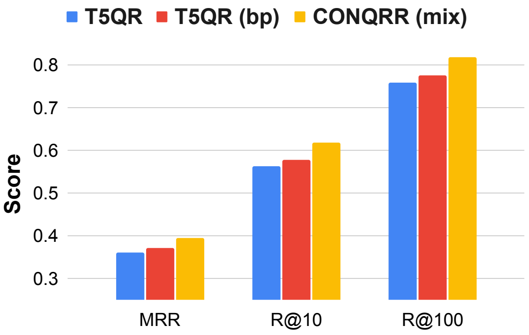

Figure 5: Evaluation scores on QReCC for T5QR w/ or w/o brevity penalty and CONQRR (mix), with DE as the retriever. Recall scores (R@k) are divided by 100.

Analysis of Longer Rewrites We hypothesize that simply generating a longer rewritten query is not the only factor that contributes to better retrieval performance. We investigate this by applying a

[13] Missing gold labels for certain examples in the dataset has no effect on the training of CONQRR as we induce weak labels without using the provided labels.

Input	Topic-Concentrated			Topic-Shifted		
	MRR	R10	R100	MRR	R10	R100
Dial Context	**0.643**	**87.7**	**96.9**	**0.312**	**56.2**	**81.9**
ConQRR (mix)	0.588	84.0	96.9	0.259	48.3	77.2
Human Rewrite	0.510	79.9	95.2	0.380	61.3	86.0

Table 7: Results of using the dialogue context, predicted rewrite or human rewrite as the retriever input with the *finetuned* DE as the retriever.

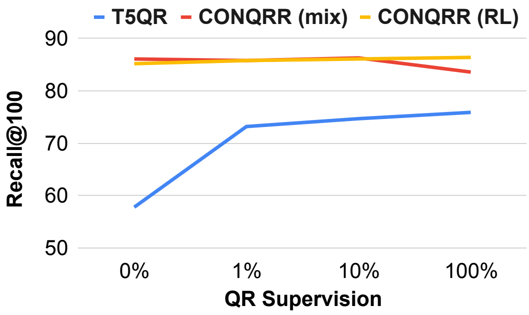

Figure 6: Recall@100 on QReCC versus the percentage of QR supervision used for training, with DE as the retriever.

brevity penalty (Wu et al., 2016) during decoding for T5QR such that its average query length matches that of ConQRR (mix). Figure 5 shows that ConQRR (mix) still outperforms T5QR with the brevity penalty for all three evaluation metrics on QReCC.

Fine-tuned Retriever Although our work focuses on the off-the-shelf retriever setting, we also conduct an experiment of fine-tuning the DE retriever with the concatenated dialogue context, the predicted rewrite from ConQRR (mix) or the human rewrite as the query input, with results in Table 7. The numbers are comparable to those in Table 4. Fine-tuning the DE retriever improves results for all scenarios, but the dialogue context benefits substantially, to the extent that it outperforms ConQRR in topic-shifted cases. However, there is still improvement room as we see benefits of human query-rewrites for topic shifts.

Additional Data Efficiency Figure Figure 6 shows the curve of Recall@100 on the overall QReCC test data using DE as the retriever versus the percentage of QR labels used for training. We also observe similar trends with Recall@10 and using BM25 as the retriever.

Additional Rewrite Examples In addition to Table 5, we put more examples in Table 8 for using DE as the retriever. We also put predicted rewrites from ConQRR (mix) that is trained towards BM25 instead of the DE retriever in Table 9. Gold passage ranks are shown in the table, using the predicted rewrites as the BM25 retriever input.

Table 10 and 11 contain examples where CONQRR (mix) rewrites have worse ranking of the gold passage than human rewrites, from our error analysis. In the two examples, the ConQRR rewrite contains less context than human rewrites or a misinterpretation of the user request.

Dialogue Context	*Q*: How did Michael Anthony's career start? *A*: While attending Pasadena City College, Anthony met Eddie Van Halen . . . Bassist Mark Stone left Mammoth. *Q*: How was that band formed?	*Q*: What kind of instrumentation did Pink Floyd use on the album The Dark Side of the Moon? . . . *Q*: Were there any particular songs they used this technique on? *A*: Speak to Me and Money. *Q*: What other different techniques did they use?
Gold Passage	**Anthony** met . . . *Van Halens decided to audition Anthony as a replacement. Anthony was impressed by their skill during subsequent jam sessions even though he had seen the brothers play before . . .*	The album features metronomic sound effects . . . *The sound effects on "Money" were created by splicing together Waters' recordings of clinking coins, tearing paper, a ringing cash register, and a clicking adding machine . . .* Pink Floyd . . .
CONQRR (mix)	How was the band Mammoth formed by **Michael Anthony**? (Rank=0)	What other different techniques did Pink Floyd use besides metronomic sound effects and tape loops? (Rank=4)
T5QR	How was the band formed? (Rank >100)	What other different techniques did Pink Floyd use on the album The Dark Side of the Moon besides metronomic sound effects and tape loops? (Rank=55)
Human	How was Mammoth formed after Mark Stone left Mammoth? (Rank=31)	What other different techniques did Pink Floyd use on the album The Dark Side of the Moon besides metronomic sound effects and tape loops? (Rank=55)

Table 8: Additional Examples of predicted rewrites and the gold passage ranks by using them as the **DE retriever** input. In these examples, CONQRR predicts alternative or less context information than human rewrites, but leads to a lower gold passage rank. *The gold answer is italicized in the gold passage.*

Dialogue Context	*Q*: What is Get 'Em Girls? *A*: Jessica Mauboy's second studio album, **Get 'Em Girls** (**2010**). . . . *Q*: Did she receive any awards or honors during these years?	*Q*: What is one actress who was a Bond girl? *A*: Ursula Andress in **Dr. No** is widely regarded as the first Bond girl. . . . . . . *Q*: Who was another Bond girl?
Gold Passage	. . . Mauboy performed "**Get 'Em Girls**" at the **2010** . . . received *her first nomination for Young Australian of the Year . . .*	. . . Ursula Andress (as Honey Ryder) in **Dr. No** (1962) is widely regarded as the first Bond girl, although she was preceded by both *Eunice Gayson* as Sylvia Trench and . . .
CONQRR (mix)	Did Jessica Mauboy receive any awards or honors during the years she released **Get 'Em Girls**? (Rank=7)	Who was another Bond girl besides Ursula Andress in **Dr. No**? (Rank=7)
T5QR	Did Jessica Mauboy receive any awards or honors during these years? (Rank >100)	Who was another Bond girl? (Rank=68)
Human	Did Jessica Mauboy receive any awards or honors during the **2010**s? (Rank=24)	Who was another Bond girl, besides Ursula Andress? (Rank=12)

Table 9: Examples of predicted rewrites and the gold passage ranks by using them as the **BM25 retriever** input. *The gold answer is italicized in the gold passage.*

Dialogue Context	*Q*: What did Jan Howard do in the early 60s? *A*: In 1960, Jan Howard went to Nashville, Tennessee, where they appeared on The Prince Albert Show, the Grand Ole Opry segment carried nationally by NBC Radio. *Q*: Did she get a record deal?
CONQRR (mix)	Did Jan Howard get a record deal? (Rank=69)
Human	Did Jan Howard get a record deal in 1960 after her appearance on The Prince Albert Show? (Rank=6)

Table 10: **Error analysis example 1**: CONQRR (mix) rewrite contains less context than the human rewrite, which leads to worse ranking of the gold passage.

Dialogue Context	*Q*: What is the keto diet? . . . *A*: The Paleolithic diet, Paleo diet, caveman diet, or stone-age diet is a modern fad diet requiring the sole or predominant eating of foods presumed to have been available to humans during the Paleolithic era. *Q*: What do they have in common?
CONQRR (mix)	What do the Paleolithic diet and the stone-age diet have in common? (Rank=78)
Human	What do paleo diet and keto diet have in common? (Rank=1)

Table 11: **Error analysis example 2**: CONQRR (mix) rewrite contains a misinterpretation of the user request, which leads to worse ranking of the gold passage than the human rewrite.