RALI@TREC iKAT 2024: Achieving Personalization via Retrieval Fusion in Conversational Search

Yuchen Hui
RALI, Université de Montréal
Montréal, Québec, Canada
yuchen.hui@umontreal.ca

Fengran Mo
RALI, Université de Montréal
Montréal, Québec, Canada
fengran.mo@umontreal.ca

Milan Mao
RALI, Université de Montréal
Montréal, Québec, Canada
milan.mao@umontreal.ca

Jian-Yun Nie
RALI, Université de Montréal
Montréal, Québec, Canada
nie@iro.umontreal.ca

Abstract

The **R**echerche **A**ppliquée en **L**inguistique **I**nformatique (**RALI**) team participated in the 2024 TREC Interactive Knowledge Assistance (iKAT) Track. In personalized conversational search, effectively capturing a user's complex search intent requires incorporating both contextual information and key elements from the user profile into query reformulation. The user profile often contains many relevant pieces, and each could potentially complement the user's information needs. It is difficult to disregard any of them, whereas introducing an excessive number of these pieces risks drifting from the original query and hinders search performance. This is a challenge we denote as over-personalization. In this paper, we tackle the problem via employing different strategies based on fusing ranking lists generated from the queries with different levels of personalization.

CCS Concepts

• **Information systems → Information retrieval**; **Users and interactive retrieval**; **Personalization**; **Query reformulation**.

Keywords

Conversational Search, Personalized Query Reformulation, Retrieval Fusion

1 Introduction

Personalized conversational information retrieval (CIR) systems aim to deliver tailored search results for users' specific information needs through multi-turn dialog, requiring the model to consider the previous user-system interactions and user profiles. The key challenge of personalized CIR lies in simultaneously extracting relevant information from the contextual history and the user profile,

and then combining these elements to conduct effective retrieval via understanding the user's complex search intent.

In CIR, the user's contextual search intent can be distilled into a stand-alone query through conversational query reformulation techniques [8]. In personalized CIR [1], it is intuitive to reformulate the query to be both context and persona independent, allowing better encapsulation of the user's nuanced search intent. A recent study [28] has demonstrated that a de-contextualized and personalized query can be built by leveraging language models (LLMs) as a search intent interpreter. However, while LLMs have shown remarkable success in conversational query reformulation (CQR) [20], the stand-alone queries produced by selecting and incorporating relevant pieces in the user profile are not always good search queries. This is because user profile terms are inherently noisier than those extracted from conversation history. Terms added by LLMs in CQR typically address missing pieces caused by coreference and ellipsis—key components that contribute to the majority of the query's semantic meaning. They are crucial to make the user's primary search intent understandable. On the other hand, terms from user profiles are usually only weakly semantically related to the query. Although these profile terms may hint at the user's preferences and potentially complement the search intent, explicitly expanding them into the reformulated queries might lead to the query drift issue, i.e., retrieving irrelevant documents that focus solely on these terms. This creates a dilemma: excluding profile terms risks omitting valuable personalized context, whereas introducing an excessive number of these pieces risks drifting from the original query and hinders search performance. This is a challenge we denote as over-personalization, which could be illustrated by the following de-contextualized and personalized query example: *What Turkish souvenir do you suggest for my mom, considering she has a collection of antique crystals and porcelains?* where "*collection of antique crystals and porcelains*" originated from the user's profile. Although the reformulated query condition on user profiles indeed enriches the user's search intent by implying a preference for fine art pieces as souvenirs, it also introduces potential noise. Highly relevant documents addressing both "*Turkish souvenir*" and "*crystals and porcelains*" should be ranked at the top. However, many irrelevant documents that focus solely on "*crystals*" and "*porcelains*" without mentioning "*Turkish souvenir*" may also rank high. To address this issue, one may consider distinguishing highly relevant documents from irrelevant ones among the top-ranked results. To achieve this,

note that these highly relevant documents would also score well when retrieved using a de-contextualized but non-personalized query. This suggests that combining document scores from both personalized and non-personalized queries can help identify truly valuable candidates: all those performing well in both ranking lists.

This simple yet practical principle is named "***The Chorus Effect***" [35]. Introduced in retrieval fusion approaches, it states that when several retrieval methods suggest that an item is relevant to a query, this tends to be stronger evidence for relevance. Dating back to the 20th century, retrieval fusion methods have already exploited this principle by merging search results from different retrieval systems to improve performance.

In this year's iKAT track, we investigate the effectiveness of retrieval fusion applied to personalized and non-personalized context-independent queries rewritten by LLMs.

2 Related Works

Conversational Search. Conversational search is an information-seeking process through interactions with a conversation system [23, 46]. Existing methods can be roughly categorized into two groups: conversational query rewriting (CQR) and conversational dense retrieval (CDR). Conversational dense retrieval [18, 21, 26, 27, 32] directly encodes the whole conversational search session to perform end-to-end dense retrieval. On the other hand, CQR methods aim to produce a stand-alone de-contextualized query that can be then submitted to any ad-hoc search models for retrieval purpose. Existing studies try to select useful tokens from the conversation context [9, 14, 25, 38] or train a generative rewriter model with conversational sessions to mimic the human rewrites [16, 34, 45]. To optimize query rewriting for the search task, some studies adopt reinforcement learning [5, 42] or apply the ranking signals with the rewriting model training [19, 24], while others jointly learn query rewriting and context modeling [31]. Remarkably, some recent methods are proposed to directly prompt the LLMs to generate rewrites [20, 22, 28, 43]. In our submissions, contextual and personalized elements are incorporated into current turn user queries by LLM-based CQR methods.

Retrieval fusion. Retrieval fusion utilizes a group of retrieval strategies (e.g., different information retrieval systems or query variations) to search in the same document collection, then merges the resulting ranking lists to improve retrieval effectiveness. Retrieval fusion methods are typically categorized into rank-based methods and score-based methods. Rank-based methods, such as RRF [6], Round-Robin fusion [36], Borda count [2], and Condorcet fusion [29], rely solely on the order of documents in multiple ranking lists. Score-based methods, by contrast, require relevance scores from candidate lists. Among these, the linear combination method (LC) is highly flexible, allowing different weights for each retrieval strategy. In LC, the relevance score of a document in the fused list is calculated as a weighted average of its scores in all candidate lists, where it is crucial to determine an appropriate weight assignment. Methods like CombSum and CombMNZ [10] assign equal weights to all candidate lists, while more sophisticated weighting assignment usually relies on relevance judgment. Vogt and Cottrell [35] and Wu et al. [41] assign weights according to the performance of candidate retrieval strategies on a group of training query. i.e., the

weight of a strategy is based on its average performance on training queries using specific metrics like Mean Average Precision. Wu [40] further employs multiple linear regression, using binary relevance judgments as the dependent variable and candidate list scores as input variables, with regression coefficients serving as weights. A comprehensive review of fusion in ad-hoc search is available in [15]. Beyond ad-hoc search, retrieval fusion has other application scenarios in the broad context of IR. For example, it is employed by [11] for combining multi-modal information in multimedia retrieval. [3, 39] explore fusion strategies to combine semantic match by dense retrievers [13] with lexical match by sparse retrievers. A recent attempt at retrieval fusion in personalization is that Huang et al. [12] trains a personalized Reinforcement Learning policy for weight assignment in the scenario of linearly fusing candidate item lists in the multi-channel recall phase of a recommender system. In personalized conversational search, it is worth exploring other reasonable weight assignment approaches since no public training data is available [28].

3 Methodology
3.1 Manual runs

We submitted two manual runs, both using the two-stage retrieval-reranking paradigm. In both stages, only the provided human rewrites are used as the search query for each conversational turn. BM25 serves as the retriever for both manual runs. In the reranking stage, we compare the performance of the MonoT5[1] [30] and RankLlama[2] [17] rerankers in the out-of-domain evaluation setting of the iKAT track. Top 50 documents recalled by BM25 are reranked. We did not participate in the response generation task for these manual runs.

3.2 Automatic runs

Our automatic runs aim to evaluate the effectiveness of a personalized ranking list fusion approach. We submitted four automatic runs in total, though our analysis here focuses on two key runs, as the others do not incorporate PTKB. Details are as follows:

*3.2.1 **RALI_gpt4o_fusion_rerank**.* This run involves four main steps:

(1) **Query Reformulation**. For each turn, three types of query reformulations are generated via prompting the GPT-4o[3] model:
  - A de-contexualized but non-personalized query rewrite;
  - The above query concatenated with a GPT-4o-generated response as a form of expansion;
  - A de-contextualized and personalized query rewrite.
  To obtain the first query and the response used in the second query, we apply the *Rewriting-And-Response* prompt from the

---

[1]Checkpoint available at castorini/monot5-base-msmarco-10k
[2]Checkpoint available at castorini/rankllama-v1-7b-lora-passage
[3]The version is gpt-4o-2024-08-06

LLM4CS framework [20]. The third query is generated using a Chain-of-Thought prompt guiding GPT-4o to incorporate relevant PTKB elements. This prompt includes human-annotated reasoning processes that analyze how PTKB elements are integrated into human rewrites in iKAT-23 test conversations.

(2) **Ranking List Fusion.** This component is central to our approach. For each conversational turn, we first get three ranking lists by querying BM25 with the reformulated queries from the previous step. Each ranking list contains the top 1000 documents retrieved for that turn. A score-based linear combination method is then used to fuse these lists into a single ranking list. Specifically, for a document D, its score in the final ranking list is computed as:

$$S_{final}(D) = \alpha_1 S_1(D) + \alpha_2 S_2(D) + \alpha_3 S_3(D)$$

Where $S_n(D)$ is the score for document D in ranking list n and α_n is the fixed weight assigned to each type of query reformulation. Note that the weights α are set consistently across all conversational turns for each reformulation type. If a document D does not present within the top 1000 documents in ranking list n, $S_n(D)$ defaults to the score of the 1000th document in that list. The weights are optimized based on performance tuning on the iKAT-23 test collection.

(3) **Reranking.** The top 50 documents from the fused ranking list are reranked using the MonoT5[4] [30] reranker. Here, the de-contextualized and personalized query generated in the first step is regarded as the search query for reranking.

(4) **Response Generation.** The top three documents from the reranked list, along with the conversational context and the user's PTKB, are provided to GPT-4o to generate a response.

3.2.2 **RALI_gpt4o_fusion_norerank.** This run uses the same query reformulation and BM25 ranking list fusion methods as described in the previous run. However, it does not rerank the fused ranking list or generate responses.

4 Evaluation Results

The evaluation results for passage ranking task of the afore-mentioned four submissions are presented in Table 1. Among the four methods, the run combining BM25 ranking list fusion and monoT5 reranking achieves the best performance, proving the effectiveness of retrieval-fusion. The evaluation results for response generation task has not been made available yet.

It is worth noting that we also evaluated all our submission methods on the iKAT-23 test collection, and the results show that the automatic fusion-then-reranking approach could not outperform runs based on manual human rewrites. Another similar interesting finding is that RankLlama outperforms MonoT5 this year, whereas the opposite is true on last year's test collection. A possible explanation for this discrepancy between the two years is **assessment bias**. Specifically, iKAT test collections are built using pooling, which tends to advantage participant methods since more documents retrieved by them are assessed. Last year, neither RankLlama-based reranking nor our proposed data fusion approach

---

[4]Checkpoint available at castorini/monot5-base-msmarco-10k

participated, which may explain why they can not achieve comparable performance to this year. See more detailed discussions in Section 5.

5 Assessment Bias

During our exploration for better personalization strategies, various types of query reformulation are tested for BM25 lexical search. Surprisingly, according to our manual observation, although some of these reformulated queries express the search intent better than human rewrites, they do not perform as well as the latter. This counter-intuitive finding drives us to investigate further via case studies. Specifically, we expanded human rewrites manually by adding clearly relevant terms and then compared the performance of these expanded versions with the original rewrites. Intuitively, the improved queries is expected to outperform the originals; However, the original ones always produce superior search results.

Table 2: Original and expanded human rewrite for iKAT 2023 test turn 9-1-3 utterance

Variation	Query Reformulation
Original	What about the DASH diet? I heard it is a healthy diet.
Expanded	What about the DASH diet? I heard it is a healthy diet. + **Dietary Approaches to Stop Hypertension**

An example is shown in Table 2. The original human rewrite for utterance 9-1-3 from iKAT 2023 asks about a healthy diet named "*DASH*", which stands for "*Dietary Approaches to Stop Hypertension*". Generally, concatenating a query with the full name of an acronym in it would help BM25 to retrieve more relevant documents. However, evaluation results in Table 3 indicate that these conceptually reasonable expansions drastically harm retrieval metrics calculated based on the relevance judgment.

Table 3: Evaluation and statistics on the original and expanded human rewrite

Variation	NDCG@10	R@20	# Assessed docs in Top20
Original	52.9	0.12	14
Expanded	29.4	0.04	5

We attribute the unexpectedly poor performance of improved query reformulation to biases in the iKAT track's assessment process. The iKAT track has two main submission classes, *automatic* and *manual*. All manual runs utilize the human rewrite as an input query, hence more documents retrieved by human rewrites are assessed in the TREC pooling process. This fact places all other query reformulations at a disadvantage as the assessment is less complete for them. For instance, in the turn 9-1-3 case, only 5 of the top 20 documents recalled by our more reasonable expanded query are assessed, compared to 14 for the original human rewrites. This bias might explain the inferior performance of the expanded 9-1-3 human rewrite.

A similar observation is reported in [30], where the authors find that due to the use of early year TREC test collections built before the advent of BERT-based retrieval models, documents retrieved

Table 1: TREC iKAT 2024 evaluation results

Submissions	MRR	NDCG@5	NDCG@10	Recall@10	Recall@100	MAP@1K
automatic runs						
BM25 fusion + no rerank	75.2	37.9	35.6	7.37	**28.4**	18.1
BM25 fusion + monoT5	**84.1**	**51.1**	**47.7**	**9.72**	**28.4**	**21.1**
manual runs						
BM25 + monoT5	79.8	40.4	35.4	7.7	20.2	13.6
BM25 + rankllama	80.0	42.3	36.7	7.7	20.2	13.6

by their T5-based models have lower proportions of judgment rate compared to those searched by BM25.

These raise questions about the reusability of iKAT test collections: can the pooling-built iKAT test collections reliably evaluate non-participant retrieval strategies that did not contribute to the relevance judgments? Investigations conducted by [37] confirmed that a test collection is reusable for new retrieval methods if ***deep pools*** are used over highly effective runs during assessment. This conclusion is supported by experiments on the TREC-8 ad hoc collection, which are created with a large pooling depth of 100, contrasting with the smaller pooling depth of 10 for iKAT 2023, as shown in Table 4. In the scenario of shallow pooling however, a separate study [44] on TREC 2019 Deep Learning track [7] shows that, with depth of 10, the reliability of evaluation results using such test collections can not be guaranteed, aligning with our observations for iKAT 23.

Table 4: Statistics of test collections

Collection	# Query	Pool Depth	# Assessed docs/psgs	Collection size
TREC-8	50	100	86,830	525,000
TREC DL-19	43	10	9,260	8,841,823
iKAT-23	174	10	26,159	116,838,987
iKAT-24	116	up to 30	20,575	116,838,987

Further more, [4] found that a biased relevance set may occur as the size of the document collection increases. Notably, as shown in Table 4, the iKAT test collections has a tremendous document set while remaining a low pooling depth — precisely the case identified in that study as susceptible to assessment bias.

Deeper pooling depth may help alleviate the assessment bias problem, but it is a costly remedy. Fortunately, regardless of a biased test collection, if a study claims that a new method outperforms TREC participant results, this conclusion remains valid despite the bias. Any unfairness in the assessment would only reduce the metrics for the new method rather than inflate them. Nevertheless, it remains valuable to explore innovative pooling and evaluation practices that could lead to more reusable test collections [33].

6 Conclusion

In this paper, we present our solution for the 2024 TREC iKAT passage ranking task, with a focus on achieving personalization through retrieval fusion in conversational search. The overall evaluation results demonstrate the effectiveness of retrieval fusion. Additionally, we discuss the assessment bias and reusability of the iKAT test collections. Future work will involve exploring more effective fusion strategies.

References

[1] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffery Dalton, and Leif Azzopardi. 2024. Trec ikat 2023: The interactive knowledge assistance track overview. *arXiv preprint arXiv:2401.01330* (2024).

[2] Javed A. Aslam and Mark Montague. 2001. Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, Louisiana, USA) *(SIGIR '01)*. Association for Computing Machinery, New York, NY, USA, 276–284. https://doi.org/10.1145/383952.384007

[3] Sebastian Bruch, Siyu Gai, and Amir Ingber. 2023. An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems* 42, 1 (2023), 1–35.

[4] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. 2007. Bias and the limits of pooling for large collections. *Information retrieval* 10 (2007), 491–508.

[5] Zhiyu Chen, Jie Zhao, Anjie Fang, Besnik Fetahu, Oleg Rokhlenko, and Shervin Malmasi. 2022. Reinforced Question Rewriting for Conversational Question Answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Yunyao Li and Angeliki Lazaridou (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 357–370. https://doi.org/10.18653/v1/2022.emnlp-industry.36

[6] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 758–759. https://doi.org/10.1145/1571941.1572114

[7] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the TREC 2019 deep learning track. *arXiv preprint arXiv:2003.07820* (2020).

[8] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 5918–5924. https://doi.org/10.18653/v1/D19-1605

[9] 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*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 319–326. https://aclanthology.org/2022.findings-aacl.30

[10] Edward Fox and Joseph Shaw. 1994. Combination of multiple searches. *NIST special publication SP* (1994), 243–243.

[11] Ilias Gialampoukidis, Anastasia Moumtzidou, Dimitris Liparas, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2016. A hybrid graph-based and non-linear late fusion approach for multimedia retrieval. In *2016 14th International workshop on content-based multimedia indexing (CBMI)*. IEEE, 1–6.

[12] Junjie Huang, Jiarui Qin, Jianghao Lin, Ziming Feng, Yong Yu, and Weinan Zhang. 2024. Unleashing the Potential of Multi-Channel Fusion in Retrieval for Personalized Recommendations. *arXiv preprint arXiv:2410.16080* (2024).

[13] 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)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[14] Vaibhav Kumar and Jamie Callan. 2020. Making information seeking easier: An improved pipeline for conversational search. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 3971–3980.

[15] Oren Kurland and J Shane Culpepper. 2018. Fusion in information retrieval: Sigir 2018 half-day tutorial. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1383–1386.

[16] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. *arXiv preprint arXiv:2004.01909* (2020).

[17] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2421–2425.

[18] Kelong Mao, Chenlong Deng, Haonan Chen, Fengran Mo, Zheng Liu, Tetsuya Sakai, and Zhicheng Dou. 2024. ChatRetriever: Adapting Large Language Models for Generalized and Robust Conversational Dense Retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1227–1240. https://aclanthology.org/2024.emnlp-main.71

[19] Kelong Mao, Zhicheng Dou, Bang Liu, Hongjin Qian, Fengran Mo, Xiangli Wu, Xiaohua Cheng, and Zhao Cao. 2023. Search-oriented conversational query editing. In *Findings of the Association for Computational Linguistics: ACL 2023*. 4160–4172.

[20] Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023. Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1211–1225. https://doi.org/10.18653/v1/2023.findings-emnlp.86

[21] Kelong Mao, Hongjin Qian, Fengran Mo, Zhicheng Dou, Bang Liu, Xiaohua Cheng, and Zhao Cao. 2023. Learning denoised and interpretable session representation for conversational search. In *Proceedings of the ACM Web Conference 2023*. 3193–3202.

[22] Fengran Mo, Abbas Ghaddar, Kelong Mao, Mehdi Rezagholizadeh, Boxing Chen, Qun Liu, and Jian-Yun Nie. 2024. CHIQ: Contextual History Enhancement for Improving Query Rewriting in Conversational Search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 2253–2268. https://aclanthology.org/2024.emnlp-main.135

[23] Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2024. A survey of conversational search. *arXiv preprint arXiv:2410.15576* (2024).

[24] Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. ConvGQR: Generative Query Reformulation for Conversational Search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4998–5012. https://doi.org/10.18653/v1/2023.acl-long.274

[25] Fengran Mo, Jian-Yun Nie, Kaiyu Huang, Kelong Mao, Yutao Zhu, Peng Li, and Yang Liu. 2023. Learning to relate to previous turns in conversational search. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1722–1732.

[26] Fengran Mo, Chen Qu, Kelong Mao, Yihong Wu, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. 2024. Aligning query representation with rewritten query and relevance judgments in conversational search. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1700–1710.

[27] Fengran Mo, Bole Yi, Kelong Mao, Chen Qu, Kaiyu Huang, and Jian-Yun Nie. 2024. Convsdg: Session data generation for conversational search. In *Companion Proceedings of the ACM on Web Conference 2024*. 1634–1642.

[28] Fengran Mo, Longxiang Zhao, Kaiyu Huang, Yue Dong, Degen Huang, and Jian-Yun Nie. 2024. How to Leverage Personal Textual Knowledge for Personalized Conversational Information Retrieval. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 3954–3958.

[29] Mark Montague and Javed A Aslam. 2002. Condorcet fusion for improved retrieval. In *Proceedings of the eleventh international conference on Information and knowledge management*. 538–548.

[30] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 708–718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

[31] Hongjin Qian and Zhicheng Dou. 2022. Explicit query rewriting for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 4725–4737.

[32] 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*. 539–548.

[33] Ian Soboroff. 2024. Don't Use LLMs to Make Relevance Judgments. *arXiv preprint arXiv:2409.15133* (2024).

[34] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM international conference on web search and data mining*. 355–363.

[35] Christopher C Vogt and Garrison W Cottrell. 1999. Fusion via a linear combination of scores. *Information retrieval* 1, 3 (1999), 151–173.

[36] E Voorhees, Narendra K Gupta, and Ben Johnson-Laird. 1995. The collection fusion problem. *NIST SPECIAL PUBLICATION SP* (1995), 95–95.

[37] Ellen M Voorhees, Ian Soboroff, and Jimmy Lin. 2022. Can Old TREC Collections Reliably Evaluate Modern Neural Retrieval Models? *arXiv preprint arXiv:2201.11086* (2022).

[38] Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 921–930.

[39] Shuai Wang, Shengyao Zhuang, and Guido Zuccon. 2021. Bert-based dense retrievers require interpolation with bm25 for effective passage retrieval. In *Proceedings of the 2021 ACM SIGIR international conference on theory of information retrieval*. 317–324.

[40] Shengli Wu. 2012. Linear combination of component results in information retrieval. *Data & Knowledge Engineering* 71, 1 (2012), 114–126.

[41] Shengli Wu, Yaxin Bi, Xiaoqin Zeng, and Lixin Han. 2009. Assigning appropriate weights for the linear combination data fusion method in information retrieval. *Information Processing & Management* 45, 4 (2009), 413–426.

[42] Zequ Wu, Yi Luan, Hannah Rashkin, D. Reitter, and Gaurav Singh Tomar. 2021. CONQRR: Conversational Query Rewriting for Retrieval with Reinforcement Learning. In *Conference on Empirical Methods in Natural Language Processing*.

[43] Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing Conversational Search: Large Language Model-Aided Informative Query Rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5985–6006. https://doi.org/10.18653/v1/2023.findings-emnlp.398

[44] Emine Yilmaz, Nick Craswell, Bhaskar Mitra, and Daniel Campos. 2020. On the Reliability of Test Collections for Evaluating Systems of Different Types. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2101–2104. https://doi.org/10.1145/3397271.3401317

[45] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 1933–1936.

[46] Hamed Zamani, Johanne R. Trippas, Jeffrey Dalton, and Filip Radlinski. 2022. Conversational Information Seeking. *Found. Trends Inf. Retr.* 17 (2022), 244–456. https://api.semanticscholar.org/CorpusID:246210119