GaQR: An Efficient Generation-augmented Question Rewriter

Oliver Young
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
oliveryoung200211@gmail.com

Yixing Fan
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
fanyixing@ict.ac.cn

Ruqing Zhang
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
zhangruqing@ict.ac.cn

Jiafeng Guo
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
guojiafeng@ict.ac.cn

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

Xueqi Cheng
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
cxq@ict.ac.cn

Abstract

Query understanding is an essential part in search systems to improve the recall. Unlike prior works focusing on word expansions, in this paper, we leverage the comprehension ability of large language models (LLMs) to generate detailed queries from a global semantic perspective. To this end, we introduce an efficient *generation-augmented question rewriter* (GaQR) to reformulate a question into several queries using chain of thought (CoT) and make it more efficient through knowledge distillation. We first prompt a teacher model to generate indicative queries by considering answer generation one step ahead. Then, we filter out low-quality queries by validating the effectiveness of all generated queries in retrieving useful passages. Finally, we distill a student rewriter based on the verified results to improve efficiency. Our experimental results demonstrate that the rewriter improves the retrieval performance by 3% to 15% on the Miracl and NFCorpus datasets and shows good generalisation ability across different retrieval methods. Moreover, the efficiency of the rewriter after knowledge distillation is improved by as much as 5 times. Code is available at https://github.com/youngbeauty250/GaQR.

CCS Concepts

• **Information systems → Query reformulation**.

Keywords

Question rewriting, Generation-augmented retrieval, Knowledge distillation

ACM Reference Format:
Oliver Young, Yixing Fan, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, and Xueqi Cheng. 2024. GaQR: An Efficient Generation-augmented Question Rewriter. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679930

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679930

1 Introduction

Search systems have developed into complex systems that including many modules, understanding the query intent constitutes a first step. Accurate intent comprehension is pivotal for search systems to be effective while a misinterpretation can affect all subsequent modules. In contrast to traditional keyword-based queries, real-world queries often involve complex, natural language descriptions of questions, for which it is usually difficult to retrieve relevant documents.

Previous research on query understanding has primarily focused on semantic or topical word expansions. One of the mainstream approaches is *global expansion* [2, 12, 19, 26], which directly expands and reconstructs the initial query based on external knowledge bases such as WordNet [17]. Another group of approaches is *local feedback*, which involves the correction and expansion of the initial query based on the top-k pseudo-documents retrieved from the initial query [1, 20, 22]. Though these approaches are practical in many applications, they either rely on pre-defined heuristic rules or heavily depend on the quality of the retrieved pseudo-documents, potentially causing semantic shifts in the rewritten queries and limiting its ability to capture the user's intent.

Recently, with the fast development of large language models (LLMs), there have been efforts to use the strong comprehension ability of LLM for understanding query intent [9, 14, 16, 30]. For instance, Query2doc [30] prompts LLMs to generate pseudo-documents and concatenates them with the original query as the rewritten query. Mao et al. [16] introduces LLM4CS by directly prompting LLMs to generate multiple query rewrites to help conversational search. Ma et al. [14] proposes a framework that combines the reinforcement learning with LLM rewrites, while Jagerman et al. [9] compare the effect of different prompts for LLM rewriting. There are also works trying to interleave between query rewriting and answer generation in retrieval-augmented generation (RAG) [6, 25], which take the previously generated answer to refine the query for next-step retrieval in the retrieval-generation loop. However, directly using LLMs for query rewriting may lead to undesirable results since LLMs are prone to hallucinate. More importantly, LLMs require substantial computational resources, posing a significant burden for real-time responses in the search system.

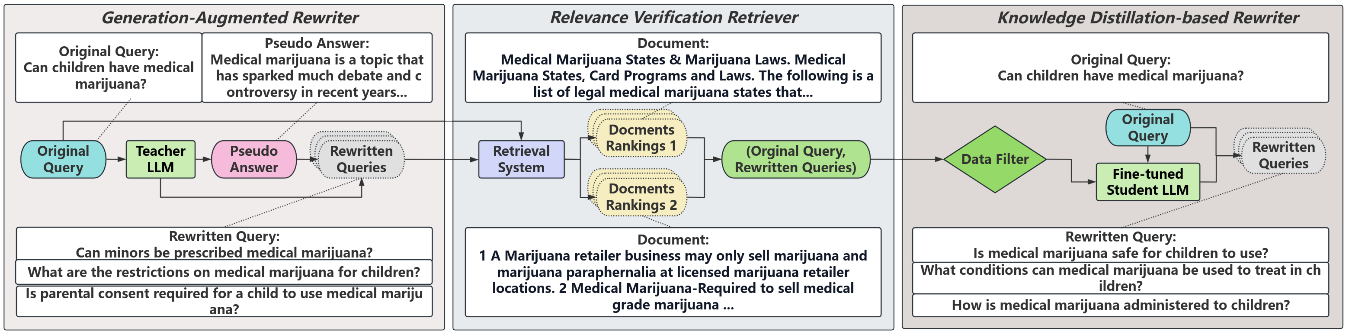
Generation-Augmented Rewriter
Original Query:
Can children have medical marijuana?
Pseudo Answer:
Medical marijuana is a topic that has sparked much debate and c ontroversy in recent years...
Original Query
Teacher LLM
Pseudo Answer
Rewritten Queries
Rewritten Query:
Can minors be prescribed medical marijuana?
What are the restrictions on medical marijuana for children?
Is parental consent required for a child to use medical mariju ana?
Relevance Verification Retriever
Document:
Medical Marijuana States & Marijuana Laws. Medical Marijuana States, Card Programs and Laws. The following is a list of legal medical marijuana states that...
Retrieval System
Docments Rankings 1
Docments Rankings 2
(Orginal Query, Rewritten Queries)
Document:
1 A Marijuana retailer business may only sell marijuana and marijuana paraphernalia at licensed marijuana retailer locations. 2 Medical Marijuana-Required to sell medical grade marijuana ...
Knowledge Distillation-based Rewriter
Original Query:
Can children have medical marijuana?
Data Filter
Original Query
Fine-tuned Student LLM
Rewritten Queries
Rewritten Query:
Is medical marijuana safe for children to use?
What conditions can medical marijuana be used to treat in ch ildren?
How is medical marijuana administered to children?

Figure 1: Overview of the generation-augmented question rewriter (GaQR). From left to right, we show its three main parts.

In this work, we propose *generation-augmented question rewriter* (GaQR), an efficient question rewriter to reformulate an input question into a set of queries by considering the anticipated answer. Specifically, we employ a teacher rewriter to first generate pseudo-answers and then rewrite the original question based on these pseudo-answers. Then, to enhance the efficiency of LLMs in rewriting, the rewritten queries and the original question are sent to a retrieval system, where the rewritten queries with higher recall are filtered out as verified results. Finally, these verified results are used as training data to fine-tune a student rewriter. Through knowledge distillation, the rewriting proficiency of the teacher is condensed onto the student, ensuring efficiency improvements without compromising effectiveness.

To evaluate our method, we conduct experiments on the MS MARCO [5] and Miracl [36] datasets, and SciFact [29], NFCorpus [3], ArguAna [28], FiQA-2018 [15] and TREC-COVID [27] datasets from the BEIR [23] benchmark. The teacher rewriter is based on GPT-3.5-instruct [4], and the base model for the student rewriter is Llama2-7b-chat [24]. The teacher rewriter exhibits performance improvements of around 2%–3% on MS MARCO and Miracl. By fine-tuning, the student rewriter learns rewriting capabilities from the teacher, achieving 2%–3% improvement on both datasets. In a completely zero-shot scenario, the student rewriter shows good generalisation across different retrieval methods, improving retrieval performance by about 2%–15% over BM25 on multiple datasets, such as FiQA-2018 for QA and NFCorpus for bio-medical IR. Furthermore, the inference time gets a 5-fold improvement compared to mainstream generative-relevance feedback (GRF) methods such as Query2Doc [30], positioning GaQR as a more economical approach for LLM query rewriting.

2 Methodology

In this section, we introduce the GaQR method for question rewriting. Our method consists of three major components, including *generation-augmented rewriter*, *relevance verification retriever*, *and knowledge distillation-based rewriter*, as illustrated in Figure 1. The *generation-augmented rewriter* thinks one-step further about the answer, and rewrites queries based on pseudo-answers. Then, the *relevance verification retriever* filters out low-quality queries according to their effectiveness in retrieving relevant documents. Finally, the *knowledge distillation-based rewriter* is built by fine-tuning a smaller LLM based on the verified results. We describe these components in detail in the follows.

2.1 Generation-Augmented Rewriter

The core of the *generation-augmented rewriter* is a special CoT [31], to employ a powerful LLM to generate high-quality search queries by thinking one-step further about the answer. The key idea is inspired by the traditional statistical language model [18], which assumes that the user would envision the ideal result before constructing the search query. Specifically, we prompts LLM to generate search queries with a two-step prompt. The first-step prompt requires the LLM to take the original question as input. Then, the second-step prompt combines the original question and the above answer as input, requires the LLM to generate effective queries in retrieving supporting evidences relevant to the answer. If fact, when we search a query, we first formulate the basic form of the answer in our minds. For example, when we ask "Which has a higher global production, apples or bananas?", we expect the answer to include the global production of apples, bananas, and a comparison of these two.

2.2 Relevance Verification Retriever

The *relevance verification retriever* aims to filter low-quality queries generated by the teacher model. The main reason to design this module is that we directly prompt the teacher rewriter to generate queries without fine-tuning, which may cause undesirable results due to the hallucination issue. In order to refine teacher's rewriting capabilities, we employ an external retriever to check its performance. Specifically, we use the retriever to perform document retrieval for the original question and rewritten queries, and then compare their recall performances on the returned document lists. Rewritten queries with higher recall are retained as a valid rewriting, verifying that they are more useful for the retriever.

2.3 Knowledge Distillation-based Rewriter

After obtaining a high-quality query rewriting dataset D, we fine-tune a student rewriter with much less number of parameters than the teacher. Specifically, we use D to train a LLM in a standard auto-regressive manner, with the objective showing as the Equation (1).

$$\mathcal{L} = \max_{M} \mathbb{E}_{(x,y)\sim D}\left[\log p_M(q' \mid q)\right], \tag{1}$$

where $\mathcal{L}$ represents the likelihood we are trying to maximize. M denotes the model parameters. The expectation $\mathbb{E}_{(x,y)} \sim D$ averages over our dataset D. $p_M(q' \mid q)$ is the probability of the model M generating the rewrites q' given the original query q.

Table 1: Main results in terms of R@1k of GaQR on 5 out-of-domain datasets.

	SciFact [29]		**NFCorpus [3]**		**ArguAna [28]**		**FiQA-2018 [15]**		**TREC-COVID [27]**	
Method	Origin	+GaQR	Origin	+GaQR	Origin	+GaQR	Origin	+GaQR	Origin	+GaQR
Lexical retrieval										
BM25	98.00	99.00$^{+1.00}$	33.80	51.08$^{+17.38}$	98.93	99.00$^{+0.07}$	73.04	75.11$^{+2.07}$	38.91	44.08$^{+5.17}$
QL [35]	98.00	**99.33**$^{+1.00}$	36.00	51.76$^{+15.76}$	99.15	98.86$^{-0.29}$	75.94	78.22$^{+2.28}$	41.28	**46.72**$^{+5.44}$
Sparse retrieval										
SPART [37]	96.27	97.17$^{+0.90}$	50.16	53.75$^{+3.59}$	97.72	98.43$^{+0.71}$	70.21	72.68$^{+2.47}$	32.63	35.37$^{+2.74}$
DocT5query [7]	98.00	99.00$^{+1.00}$	34.50	51.05$^{+16.55}$	98.93	99.08$^{+0.15}$	74.49	78.81$^{+4.32}$	36.35	39.98$^{+3.63}$
Dense retrieval										
TAS-B [8]	98.33	**99.33**$^{+1.00}$	58.53	58.63$^{+0.10}$	99.43	**99.50**$^{+0.07}$	**82.86**	82.52$^{-0.34}$	32.19	31.59$^{-0.60}$
ANCE [32]	95.67	96.33$^{+0.66}$	57.68	**59.27**$^{+1.59}$	99.00	98.86$^{-0.14}$	80.71	81.19$^{+0.48}$	34.23	34.88$^{+0.65}$
DPR [11]	91.43	92.83$^{+1.40}$	52.36	55.64$^{+3.28}$	94.45	95.45$^{+1.00}$	57.92	60.45$^{+2.53}$	15.78	17.16$^{+1.38}$

Table 2: Main results in terms of R@1k of GaQR and other methods on 5 out-of-domain datasets.

Methods	BM25	+RM3	+Q2E	+Q2D	+CoT	+GaQR
SciFact [29]	98.00	98.00	98.67	99.67	**100.00**	99.00
NFCorpus [3]	33.80	52.28	58.00	**58.17**	56.63	51.08
ArguAna [28]	98.93	98.43	99.00	98.93	98.51	**99.00**
FiQA-2018 [15]	73.04	**75.24**	71.10	71.37	70.53	75.11
TREC-COVID [27]	38.91	41.07	40.62	43.24	42.12	**44.08**
Average rank	4.6	4.2	3.6	2.6	3.6	**2.4**

3 Experimental Setup

3.1 Datasets & Model

We take three widely used benchmarks to test the effectiveness of our method, namely MS MARCO [5], Miracl [36] and BEIR [23]. To assess the generalisation capability of our model, we select a subset of BEIR from different domains, including SciFact [29], FiQA-2018 [15], ArguAna [28], NFCorpus [3] and TREC-COVID [27], as our test datasets. We take Recall@1000 (R@1K) as the main evaluation metric since we focus on the first-stage retrieval in search systems.

For the teacher rewriter described in Section 2.1, we adopt GPT-3.5-instruct. Without any fine-tuning or training, we prompt it to give a pseudo-answer $\hat{a}$ and rewrite the original question q as described in section 2.1. For the student rewriter defined in Section 2.3, we choose Llama2-7b-chat[1] as our backbone. For fine-tuning, we set the learning rate to 2e-5, epochs to 5, and temperature to 0.2.

We take the Anserini [13, 33, 34] implementation of BM25 [21] in 2.2 as our base retriever using the default parameters ($k = 0.9$ and $b = 0.4$). In addition, we include QL [35], SPART [37], DocT5query [7], TAS-B [8], ANCE [32] and DPR [11], as baselines to investigate the performance of our method, the parameters of which can be found in BEIR[23]. Also, to compare our method with other query expansion methods, we use BM25 with RM3 [10] and LLM-based methods mentioned in [9] including Q2E, Q2D and CoT as LLM baselines.

3.2 Main Results

To evaluate the generalisation ability of GaQR, we select a small number of queries (i.e., 6000 queries in total) from MS MARCO and Miracl datasets to fine-tune the GaQR. For each retrieval method, we test the performance of GaQR on five out-of-domain datasets mentioned in Section 3.1. The overall results are summarized in Table 1, as we can see: (i) The performance of different datasets varies significantly, where the SciFact dataset and the ArguAna dataset show limited improvements, while the other three datasets show large improvements. The main reason may be that the query length in these three tends to be shorter than SciFact and ArguAna. In this way, it is much more difficult to retrieve relevant documents with original queries, and GaQR shows great performance through query rewriting. (ii) For different baselines, lexical retrieval methods show strong generalisation ability on each dataset. Typically, both sparse retrieval methods and dense retrieval methods rely on the training dataset to boost their performances. There are also exceptions where TAS-B and ANCE show very good performances on NFCorpus and FiQA-2018 dataset. (iii) GaQR exhibits strong generalisation ability across different retrieval methods on a wide range of datasets, e.g., the improvements on NFCorpus over BM25 and QL are 17.38% and 15.76%, respectively. Besides, the improvements on lexical retrieval methods is relatively higher than on sparse retrieval and dense retrieval methods, which may be that we use BM25 as the *relevance verification retriever* in GaQR.

In addition, we compare our method with other expansion methods. For a fair comparison, we take the Llama2-7b-chat as the backbone for all LLM-based methods. The results are summarized in Table 2. As we can see, GaQR outperforms all the baselines on ArguAna, FiQA-2018 and TREC-COVID and shows good performance on SciFact, achieving the best average ranking. However, GaQR performs worse than other baselines on NFCorpus. The main reason may be that NFCorpus is a keyword-based dataset and GaQR tends to generate simple but precise rewritten queries. Thus GaQR may not be able to extend enough semantically related keywords as other baselines do. Although the Q2D method is close to our method in terms of average rankings, its computational latency is much higher, as we will show in Section 3.4. Overall, the improvements on different retrieval methods and comparisons with other query expansion methods demonstrate the effectiveness of GaQR.

[1] https://llama.meta.com/llama2/

Table 3: Ablation study in terms of R@100 and R@1k of GaQR on MS MARCO [5] and Miracl [36] .

	MS MARCO [5]		Miracl [36]	
Method	R@100	R@1k	R@100	R@1k
BM25	65.78	85.26	70.71	87.94
GaQR	65.57	87.12	**76.02**	90.97
without Q	59.89	82.38	71.45	87.85
without filter	65.51	86.38	75.22	91.08
without distill	**66.37**	**87.46**	75.94	**91.44**

Table 4: Latency analysis of our method. The LLM call is on a single A-800 GPU. We average the time(ms) for each query.

	Q2D	CoT	Q2E	GaQR
SciFact [29]	10243ms (4.43x)	11000ms (4.76x)	3620ms (1.57x)	2310ms (1x)
NFCorpus [3]	10658ms (7.51x)	13502ms (9.52x)	4165ms (2.94x)	1419ms (1x)
FiQA-2018 [15]	11463ms (6.62x)	13237ms (7.65x)	6232ms (3.60x)	1731ms (1x)
Miracl [36]	5092ms (3.62x)	6380ms (4.53x)	3504ms (2.49x)	1407ms (1x)

3.3 Ablation Study

To better understand the effectiveness of each component of GaQR, we conduct an ablation study on MS MARCO [5] and Miracl [36] by removing different parts from GaQR. From Table 3, we find see that: 1) When only using rewritten queries without original query Q for retrieval, the performance declines sharply on the two datasets, even worse than the BM25 baseline. In fact, the rewritten queries will enrich the semantics of the original query, but only using them may cause semantic shifts for retrieval. 2) Without *relevance verification retriever*, GaQR achieves lower performance than origin, demonstrating that it is important to filter the training data. 3) Interestingly, when we test the teacher rewriter without distilling, it outperforms its student rewriter GaQR in most conditions, but performs worse than GaQR in terms of R@100 on Miracl, which shows the effectiveness of fine-tuning in boosting small LLMs. At a higher level of efficiency, GaQR obtains results which approach or even surpass the teacher rewriter.

3.4 Latency Analysis

To evaluate the efficiency of GaQR, we conduct a latency analysis on four datasets. Table 4 shows the results of GaQR and different rewriting methods. First, we can see that all the methods are consistently more efficient in processing Miracl questions than on other datasets, which may be that LLM is more effective in understanding natural questions than keyword-based questions. Second, Q2E achieves the lowest latency among all baselines since it does not require the generation of answers or CoT. Finally, GaQR has about 4–9 times improvement compared to the CoT method, and 1–3 times improvement compared to the Q2E method. It suggests that GaQR refines LLM's ability to focus on rewriting well, which in turn speeds up the efficiency of its rewriting.

Table 5: Case analysis from the Miracl [36] dataset.

Query	Why is voting day on a Tuesday?
Rewritten queries	1. **What** is the **historical significance** behind making election day a Tuesday?(Explanation) 2. **How** does having voting day on a Tuesday **affect** voter turnout?(Process) 3. **Why** is it **important** for election day to fall on a **specific day** of the week?(Reason)

3.5 Case Study

To better understand how GaQR rewrites questions, in Table 5 we show a sampled question from the Miracl dataset along with three rewritten queries from our GaQR rewriter. As we can see, the original question is simple and wide, expressed in natural language. For search systems, documents retrieved may be more about the key term of "voting day" or "Tuesday," and less about the relationship between them. In contrast, the rewritten queries express more complex semantics than the original question, including three question types "What," "How," and "Why" which represent three possible intentions of the user. The documents retrieved by these rewritten queries can provide more comprehensive information for users.

4 Conclusion

We have proposed an efficient question rewriter named GaQR to improve the retrieval performance for search systems. Our results show that LLMs with additional answer generation could produce high-quality question rewrites. Moreover, GaQR obtains good generalisation performance across different retrieval datasets through knowledge distillation. Furthermore, our method gets a balance between efficiency and effectiveness, which makes it more practical to apply LLMs to search systems. Still, there are limitations to our work. First, we take search results as an indirect evaluation for rewriting, since there lacks question rewriting datasets. Second, rewritten queries generated by GaQR may be limited by the training corpus. Finally, we take all types of question in a unified rewriting process, whereas questions with same type may share common rewriting patterns, e.g., "When" questions and "Why" questions. It would be interesting to create a taxonomy of questions and rewrite them accordingly. We will investigate these topics in future work.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (NSFC) under Grants No. 62372431, 62302486, the Youth Innovation Promotion Association CAS under Grants No. 2021100, the Innovation Project of ICT, CAS under Grants No. E463010005, E361140, the Strategic Priority Research Program of the CAS under Grants No. XDB0680102, the project under Grants No. 2023YFA-1011602, JCKY2022130C039 and 2021QY1701, the Dutch Research Council (NWO) under project no. 024.004.022, NWA.1389.20.183, KICH3.LTP.20.006, and the European Union's Horizon Europe program under grant agreement No. 101070212. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

[1] Gianni Amati. 2003. Probability Models for Information Retrieval Based on Divergence from Randomness. https://api.semanticscholar.org/CorpusID:30208372

[2] Jing Bai, Jian-Yun Nie, Guihong Cao, and Hugues Bouchard. 2007. Using Query Contexts in Information Retrieval. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://api.semanticscholar.org/CorpusID:1356438

[3] Vera Boteva, Demian Gholipour Ghalandari, Artem Sokolov, and Stefan Riezler. 2016. A Full-Text Learning to Rank Dataset for Medical Information Retrieval. In *European Conference on Information Retrieval*. https://api.semanticscholar.org/CorpusID:14355670

[4] 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 (2020). https://api.semanticscholar.org/CorpusID:218971783

[5] Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *ArXiv* abs/1611.09268 (2016). https://api.semanticscholar.org/CorpusID:1289517

[6] Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. RQ-RAG: Learning to Refine Queries for Retrieval Augmented Generation. *arXiv preprint arXiv:2404.00610* (2024).

[7] David R. Cheriton. 2019. From doc2query to docTTTTTquery. https://api.semanticscholar.org/CorpusID:208612557

[8] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy J. Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021). https://api.semanticscholar.org/CorpusID:233231706

[9] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query Expansion by Prompting Large Language Models. *ArXiv* abs/2305.03653 (2023). https://api.semanticscholar.org/CorpusID:258546701

[10] Nasreen Abdul Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah S. Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. In *Text Retrieval Conference*. https://api.semanticscholar.org/CorpusID:16221853

[11] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. *ArXiv* abs/2004.04906 (2020). https://api.semanticscholar.org/CorpusID:215737187

[12] Michael E. Lesk. 1969. Word-word Associations in Document Retrieval Systems. *American Documentation* 20 (1969), 27–38. https://api.semanticscholar.org/CorpusID:5481961

[13] Jimmy J. Lin, Matt Crane, Andrew Trotman, Jamie Callan, Ishan Chattopadhyaya, John Foley, Grant Ingersoll, Craig Macdonald, and Sebastiano Vigna. 2016. Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge. In *European Conference on Information Retrieval*. https://api.semanticscholar.org/CorpusID:2741762

[14] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query Rewriting for Retrieval-Augmented Large Language Models. *ArXiv* abs/2305.14283 (2023). https://api.semanticscholar.org/CorpusID:258841283

[15] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. *Companion Proceedings of the The Web Conference 2018* (2018). https://api.semanticscholar.org/CorpusID:13866508

[16] Kelong Mao, Zhicheng Dou, Haonan Chen, Fengran Mo, and Hongjin Qian. 2023. Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:257495903

[17] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38 (1995), 39–41. https://api.semanticscholar.org/CorpusID:1671874

[18] Jay M. Ponte and W. Bruce Croft. 1998. A Language Modeling Approach to Information Retrieval. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://api.semanticscholar.org/CorpusID:14103653

[19] Yonggang Qiu and Hans-Peter Frei. 1993. Concept Based Query Expansion. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://api.semanticscholar.org/CorpusID:11711278

[20] Stephen E. Robertson. 1991. On Term Selection for Query Expansion. *J. Documentation* 46 (1991), 359–364. https://api.semanticscholar.org/CorpusID:1418295

[21] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Text Retrieval Conference*. https://api.semanticscholar.org/CorpusID:3946054

[22] Joseph John Rocchio. 1971. Relevance Feedback in Information Retrieval. https://api.semanticscholar.org/CorpusID:61859400

[23] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *ArXiv* abs/2104.08663 (2021). https://api.semanticscholar.org/CorpusID:233296016

[24] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv* abs/2307.09288 (2023). https://api.semanticscholar.org/CorpusID:259950998

[25] H. Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. *ArXiv* abs/2212.10509 (2022). https://api.semanticscholar.org/CorpusID:254877499

[26] Ellen M. Voorhees. 1994. Query Expansion Using Lexical-semantic Relations. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://api.semanticscholar.org/CorpusID:18126742

[27] Ellen M. Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. TREC-COVID. *ACM SIGIR Forum* 54 (2020), 1–12. https://api.semanticscholar.org/CorpusID:218581058

[28] Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the Best Counterargument without Prior Topic Knowledge. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:51880268

[29] David Wadden, Kyle Lo, Lucy Lu Wang, Shanchuan Lin, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. *ArXiv* abs/2004.14974 (2020). https://api.semanticscholar.org/CorpusID:216867133

[30] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:257505063

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv* abs/2201.11903 (2022). https://api.semanticscholar.org/CorpusID:246411621

[32] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *ArXiv* abs/2007.00808 (2020). https://api.semanticscholar.org/CorpusID:220302524

[33] Peilin Yang, Hui Fang, and Jimmy J. Lin. 2017. Anserini: Enabling the Use of Lucene for Information Retrieval Research. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2017). https://api.semanticscholar.org/CorpusID:1340183

[34] Peilin Yang, Hui Fang, and Jimmy J. Lin. 2018. Anserini: Reproducible Ranking Baselines Using Lucene. *ACM J. Data Inf. Qual.* 10 (2018), 16:1–16:20. https://api.semanticscholar.org/CorpusID:53112735

[35] ChengXiang Zhai and John D. Lafferty. 2001. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. *ACM SIGIR Forum* 51 (2001), 268 – 276. https://api.semanticscholar.org/CorpusID:52864147

[36] Xinyu Crystina Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages. *ArXiv* abs/2210.09984 (2022). https://api.semanticscholar.org/CorpusID:252968355

[37] Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2021. SPARTA: Efficient Open-Domain Question Answering via Sparse Transformer Matching Retrieval. In *Proc. of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 565–575. https://doi.org/10.18653/v1/2021.naacl-main.47