MILL: Mutual Verification with Large Language Models for Zero-Shot Query Expansion

Pengyue Jia[1], **Yiding Liu**[2], **Xiangyu Zhao**[1]*, **Xiaopeng Li**[2]
Changying Hao[2], **Shuaiqiang Wang**[2], **Dawei Yin**[2]
[1]City University of Hong Kong, [2]Baidu Inc.
{jia.pengyue,xiaopli2-c}@my.cityu.edu.hk
{liuyiding.tanh,cyhaocn,shqiang.wang}@gmail.com
xianzhao@cityu.edu.hk, yindawei@acm.org

Abstract

Query expansion, pivotal in search engines, enhances the representation of user information needs with additional terms. While existing methods expand queries using retrieved or generated contextual documents, each approach has notable limitations. Retrieval-based methods often fail to accurately capture search intent, particularly with brief or ambiguous queries. Generation-based methods, utilizing large language models (LLMs), generally lack corpus-specific knowledge and entail high fine-tuning costs. To address these gaps, we propose a novel zero-shot query expansion framework utilizing LLMs for mutual verification. Specifically, we first design a query-query-document generation method, leveraging LLMs' zero-shot reasoning ability to produce diverse sub-queries and corresponding documents. Then, a mutual verification process synergizes generated and retrieved documents for optimal expansion. Our proposed method is fully zero-shot, and extensive experiments on three public benchmark datasets are conducted to demonstrate its effectiveness over existing methods. Our code is available online at https://github.com/Applied-Machine-Learning-Lab/MILL to ease reproduction.

1 Introduction

Query expansion is a critical technique in search systems, aiming to effectively capture and represent users' information needs (Zhu et al., 2023; Efthimiadis, 1996). Search engines employ query expansion to resolve ambiguities in queries and align the vocabulary of queries and documents. Central to this task is the development of contextual documents, comprising additional query terms, to enhance effectiveness (Azad and Deepak, 2019).

Specifically, existing research predominantly falls into two categories: retrieval-based and generation-based methods. Retrieval-based methods (Lv and Zhai, 2010; Yan et al., 2003; Li et al., 2022) typically construct contextual documents from the targeted corpus, assuming that the top-retrieved documents (i.e., pseudo-relevance feedback (PRF)) are reasonable expansions of a given query. Generation-based methods (Jagerman et al., 2023; Mao et al., 2023; Wang et al., 2023a) often utilize advanced generative models, such as Large Language Models, as an external knowledge base for producing contextual documents.

However, both methods have clear limitations. For retrieval-based methods, it has been observed in practice that the documents retrieved with the original query do not align well with the information needs, particularly when the original query itself is brief and ambiguous (Cao et al., 2008; Jagerman et al., 2023). For generation-based methods, directly using off-the-shelf LLMs in a few-shot or zero-shot manner can hardly align the model with a specific corpus (Wang et al., 2023a). In contrast, the LLMs could easily generate useless out-of-domain information.

To this end, we propose a novel query expansion framework based on Large Language Models (LLMs), integrating both retrieved and generated documents to mitigate their respective limitations. First, to improve contextual document generation, we design a query-query-document prompt that leverages an LLM as a zero-shot reasoner to decompose a query into multiple sub-queries during contextual document generation. This helps the LLM generate diverse contextual information that is more likely to cover the underlying search intent.

Next, we propose a mutual verification framework that exploits generated and retrieved contextual documents for query expansion. To be more specific, we propose to filter out the uninformative generated documents via comparing their relevance with the top-retrieved documents. By doing this, the selected generated documents are intuitively more aligned with the target corpus. Conversely, we also filter out the noisy retrieved documents

*Corresponding author

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2498–2518
June 16-21, 2024 ©2024 Association for Computational Linguistics

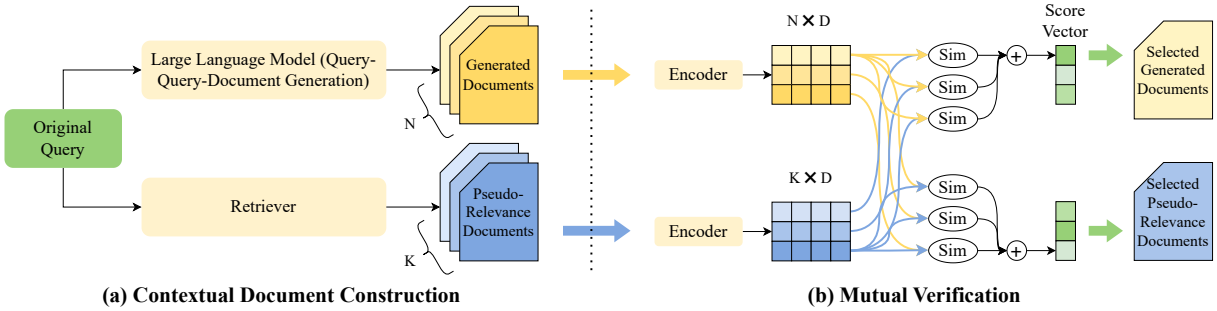

Figure 1: Overview of MILL.

via comparing their relevance with the generated documents. The external contextual knowledge embedded in the generated documents can facilitate the retrieved documents to more accurately reveal search intent. We evaluate the proposed method on the downstream information retrieval task in a zero-shot manner. The results on three public datasets demonstrate that our proposed method significantly outperforms the state-of-the-art baselines. Overall, the contributions can be summarized as follows:

- We propose a **M**utual Ver**I**fication method with **L**arge **L**anguage model (denoted as MILL), a novel framework that combines generated and retrieved context for query expansion. MILL is able to mitigate the limitations of generated and retrieved context, and thus can provide more high-quality context for query expansion.

- To improve the generated contextual documents, we design a query-query-document prompting method, which elicits richer and more diverse knowledge from LLMs to cover the underlying search intents and information needs of users.

- MILL can perform high-quality query expansion in a zero-shot manner. We conduct extensive experiments on the downstream information retrieval task on three public datasets. The results demonstrate that MILL significantly outperforms existing retrieval and generation-based methods.

2 Problem Definition

Given a user query q, query expansion is to apply a function f to expand q with additional contextual information: $q' = f_\theta(q)$, where θ represents the parameters. Using the expanded query q' should be able to achieve better downstream retrieval performance compared to the original query q. More formally, such an objective can be defined as

$$\underset{\theta}{\operatorname{argmax}}\ \mathcal{M}(q', R), \text{ where } q' = f_\theta(q). \quad (1)$$

where $\mathcal{M}$ denotes the evaluation metric of the retrieval performance (e.g., recall, NDCG), and R denotes the retrieval model.

3 Methodology

In this section, we introduce our proposed query expansion method in detail. Specifically, we give an overview of MILL in Section 3.1, elaborate the query-query-document generation in Section 3.2, and introduce the mutual verification framework in Section 3.3.

3.1 Overview

The overall workflow of MILL is depicted in Figure 1, which comprises two steps, i.e., contextual document construction (Figure 1(a)) and mutual verification (Figure 1(b)). In particular, the two steps focus on the diversity and quality of contextual documents, respectively.

In the **contextual document construction** stage, we aim to construct diverse sets of contextual documents via both retrieval and generation. To create diverse contextual documents via generation, we propose a query-query-document prompt, which instructs an LLM to generate sub-queries and contextual documents in a step-by-step manner. This can better leverage the reasoning ability of LLMs that decompose a given query expansion task into multiple sub-tasks, where the generated documents could be more diverse and informative.

In the **mutual verification** stage, we aim to identify those high-quality contextual documents constructed in the first stage. In particular, the mutual verification leverages the strengths of generated documents in implying the search intent and

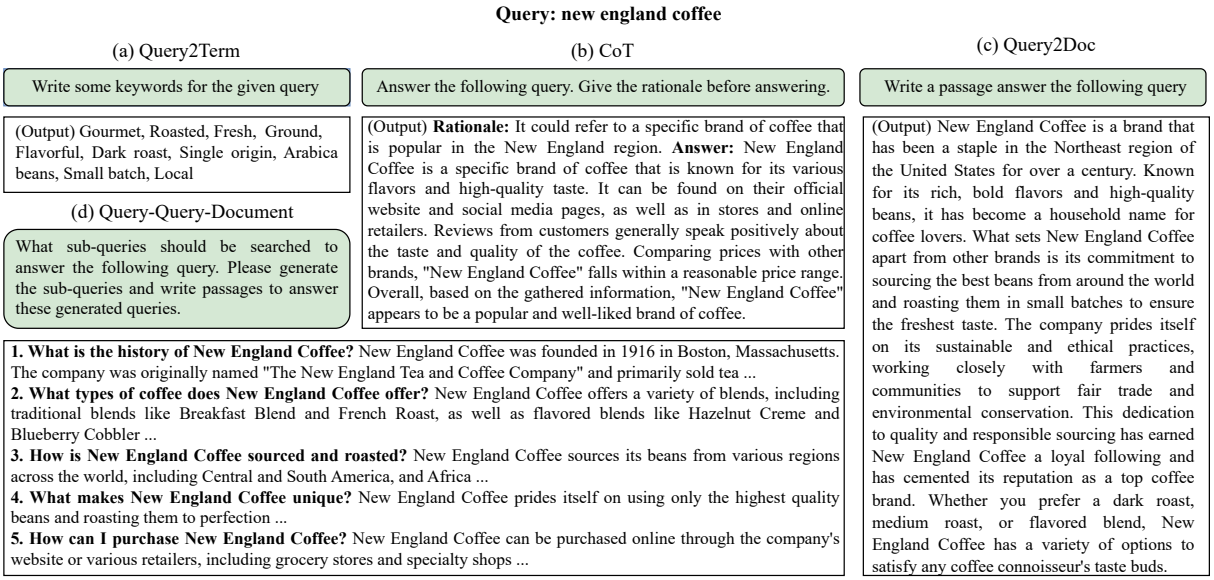

Figure 2: Query-query-document prompt compared to Query2Term, CoT, and Query2Doc. Query-query-document instructs the LLM to expand the original query from multiple perspectives by inferring the sub-queries and generating corresponding contextual documents.

the domain-specific nature of PRF documents, enabling a reciprocal selection between the two types of contextual documents. As a result, the finalized documents are more high-quality query expansion to be applied in downstream retrieval tasks.

3.2 Query-Query-Document Generation

Recently, a handful of studies (Wang et al., 2023a; Jagerman et al., 2023) have explored using Large Language Models (LLMs) to expand queries and gain initial success. However, most of them use a rather simple prompt for document generation, e.g., "write a passage that answers the given query". For a brief or ambiguous query that has multiple possible intents, the generation results could easily miss the real search intent. Motivated by this, we design a novel zero-shot prompt, particularly for the query expansion task. This method can exploit the reasoning ability of LLMs to first decompose the original query into multiple sub-queries before document generation. This improves generation diversity, and the contextual documents are more likely to cover the real search intent.

As shown in Figure 2(d), we use the instruction "what sub-queries should be searched to answer the following query: {query}." to generate sub-queries that further clarify the original query. At the same time, we instruct the language model to generate contextual documents for each sub-query through "Please generate the sub-queries and write passages to answer these generated queries." By doing this, we finally have multiple sub-queries and their corresponding contextual documents, which are more likely to cover the user's search intent. Note that the proposed method is zero-shot, which can be easily extended to few-shot.

3.3 Mutual Verification

Next, we elaborate on the mutual verification framework, where we leverage the aforementioned generated documents and pseudo-relevance documents (i.e., the retrieval-based contextual documents) to improve the overall quality of query expansion. The intuition is to leverage two types of information to complement each other, which are 1) the corpus-specific domain information of retrieved pseudo-relevance documents, and 2) the generated information of LLM reasoning that is more likely to uncover real search intent.

More specifically, the inputs of mutual verification have two sets of contextual documents:

$$\mathcal{D}^{\text{LLM}} = \{d_n^{\text{LLM}}\} = \text{LLM}(p, q),\ n \in (0, N] \quad (2)$$

$$\mathcal{D}^{\text{PRF}} = \{d_k^{\text{PRF}}\} = R_r(q),\ k \in (0, K] \quad (3)$$

where $\mathcal{D}^{\text{LLM}}$ represents the N LLM-generated documents with query-query-document prompt (denoted as p), and $\mathcal{D}^{\text{PRF}}$ represents the K documents retrieved by a vanilla PRF method (denoted as R_r), e.g., BM25 retrieval. Note that each generated document comprises a series of sub-queries and their

corresponding passages.

Next, we aim to rerank the documents in $\mathcal{D}^{\text{LLM}}$ and $\mathcal{D}^{\text{PRF}}$. In specific, we first use an off-the-shelf dense representation model to compute the representation (i.e., $\mathbf{x}_n^{\text{LLM}}$ or $\mathbf{x}_k^{\text{PRF}}$) of each document (i.e., d_n^{LLM} or d_k^{PRF}) as

$$\mathbf{x}_n^{\text{LLM}} = \text{Encoder}(d_n^{\text{LLM}}), \quad (4)$$

$$\mathbf{x}_k^{\text{PRF}} = \text{Encoder}(d_k^{\text{PRF}}), \quad (5)$$

where $\mathbf{x}_n^{\text{LLM}}$ denotes the vector for n-th generated document and $\mathbf{x}_k^{\text{PRF}}$ denotes the vector for k-th pseudo-relevance documents.

Then, we compute the semantic relevance between every pair of d_n and d_k with cosine similarity (denoted as $\text{sim}(\cdot)$), and assign a score to every document as

$$s_n^{\text{LLM}} = \sum_{k=1}^{K} \text{sim}(\mathbf{x}_n^{\text{LLM}}, \mathbf{x}_k^{\text{PRF}}), \quad (6)$$

$$s_k^{\text{PRF}} = \sum_{n=1}^{N} \text{sim}(\mathbf{x}_k^{\text{PRF}}, \mathbf{x}_n^{\text{LLM}}). \quad (7)$$

Here, we score every generated document d_n^{LLM} via aggregating its semantic relevance scores with all pseudo-relevance documents. Therefore, the score s_n^{LLM} can be interpreted as how well d_n^{LLM} is aligned with the target corpus. On the other hand, the score s_k^{PRF} can be viewed as how well the retrieved document d_k^{PRF} is likely to be a reasonable context judged by the reasoning results of LLM.

Finally, we select the top-scored documents in both sets as the final contextual documents as

$$\begin{aligned} \mathcal{D}_s^{\text{LLM}} &= \{d_n^{\text{LLM}}\},\ n \in \{n \,|s_n^{\text{LLM}} \in TopN'(s^{\text{LLM}})\}, \\ \mathcal{D}_s^{\text{PRF}} &= \{d_k^{\text{PRF}}\},\ k \in \{k \,|s_k^{\text{PRF}} \in TopK'(s^{\text{PRF}})\}, \end{aligned} \quad (8)$$

where $\mathcal{D}_s^{\text{LLM}}$ and $\mathcal{D}_s^{\text{PRF}}$ are the final selected document sets.

3.4 Query Expansion for Retrieval

After mutual verification, we integrate the selected generated documents and pseudo-relevance documents with the original query to perform the final retrieval task. In particular, we concatenate them as the new query q' as:

$$q' = \text{concat}(q,\ q,\ q,\ q,\ q,\ \mathcal{D}_s^{\text{PRF}},\ \mathcal{D}_s^{\text{LLM}}) \quad (9)$$

We repeat the original query 5 times following papers (Wang et al., 2023a; Jagerman et al., 2023) to emphasize its significance. It is worth noting that the proposed query expansion method does not need any additional labeled data and model fine-tuning. Such a zero-shot method with off-the-shelf LLM and retriever has huge potential to be applied in various search systems.

4 Experiments

4.1 Datasets and Metrics

To evaluate the effectiveness of our proposed method, we conduct extensive experiments on the following public datasets: TREC-DL-2019, TREC-DL-2020, and BEIR.

- **TREC-DL-2019&2020 (Craswell et al., 2021).** TREC-DL-2019 and TREC-DL-2020 [1] are the datasets used in the TREC Deep Learning Track. We conduct passage retrieval tasks on the datasets, each of which contains 200 queries and 8.84 million passages.
- **BEIR (Thakur et al., 2021).** BEIR[2] is a heterogeneous benchmark for comprehensive zero-shot evaluation of methods in various information retrieval tasks. We select 9 datasets with small test or dev sets from the 18 available datasets.

Following previous work (Claveau, 2021; Jagerman et al., 2023; Mao et al., 2023), we use the NDCG@N, MAP@N, Recall@N, and MRR@N as the evaluation metrics, each of which is reported with N $\in \{10, 100, 1000\}$. Additional experiments on MSMARCO are provided in Appendix A.3.

4.2 Baselines

We conduct comparative experiments with the following baselines, which can be divided into three categories: (1) **Traditional query expansion methods**: Bo1 (Amati and Van Rijsbergen, 2002), KL (Amati and Van Rijsbergen, 2002), RM3 (Abdul-Jaleel et al., 2004), and AxiomaticQE (Fang and Zhai, 2006; Yang and Lin, 2019). (2) **LLM-based expansion methods**: Query2Term (Jagerman et al., 2023), Query2Term-FS (the few-shot version of Query2Term), Query2Term-PRF (PRF document augmented Query2Term), Query2Doc (Wang et al., 2023a), Query2Doc-FS, Query2Doc-PRF, CoT (Jagerman et al., 2023), CoT-PRF. (3) **Ensembled expansion methods**: These are the variants of the LLM-based expansion methods by additionally concatenating top-retrieved PRF documents to the query. They are denoted as Query2Term*, Query2Term-FS*, Query2Term-PRF*, Query2Doc*, Query2Doc-FS*, Query2Doc-PRF*, CoT*, and CoT-PRF*. The details of the baselines and their prompts are introduced in Appendix A.1 and Appendix A.2.

[1] https://microsoft.github.io/msmarco/

[2] https://github.com/beir-cellar/beir

Table 1: Overall comparison on TREC-DL-2019 and TREC-DL-2020. The optimal results are highlighted in bold, while the suboptimal results are underscored. The results are reported on NDCG@N, AP@N, Recall@N, and MRR@N with N $\in \{10, 100, 1000\}$. The improvements are all significant (i.e., two-sided t-test with $p < 0.05$) between the optimal and suboptimal results.

	Metrics	NDCG			AP			Recall			MRR		
		@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
TREC-DL-2019	No expansion	47.95	48.74	59.34	10.14	29.07	37.00	12.23	44.22	73.62	79.44	79.49	79.50
	Traditional expansion methods												
	Bo1	50.86	50.01	61.09	10.98	31.08	39.99	12.85	45.42	75.11	78.75	78.81	78.81
	KL	50.57	49.84	60.82	10.95	30.94	39.77	12.84	45.26	74.66	78.44	78.50	78.50
	RM3	51.56	50.41	61.23	10.78	31.70	40.45	13.14	46.37	75.43	78.94	79.01	79.01
	AxiomaticQE	47.95	48.74	59.34	10.14	29.07	37.00	12.23	44.22	73.62	79.44	79.49	79.50
	LLM-based expansion methods												
	Query2Term	44.17	42.95	55.21	9.08	22.61	29.91	11.06	37.10	68.81	71.61	71.77	71.77
	Query2Term-FS	50.38	49.54	61.67	11.30	29.40	37.51	12.52	43.83	76.13	75.22	75.73	75.73
	Query2Term-PRF	48.56	48.08	57.63	11.05	30.78	37.18	12.24	43.69	70.20	80.10	80.17	80.19
	Query2Doc	62.77	61.45	71.75	13.68	39.28	49.04	14.78	52.25	84.21	90.89	91.04	91.04
	Query2Doc-FS	**63.83**	61.42	72.02	14.30	39.54	49.65	15.44	52.57	83.75	90.55	90.55	90.56
	Query2Doc-PRF	59.00	57.47	68.23	12.15	35.29	44.56	14.39	50.29	82.12	86.63	86.63	86.63
	CoT	63.44	59.57	70.94	13.43	35.53	45.67	14.97	49.91	83.43	**92.61**	**92.61**	**92.61**
	CoT-PRF	61.63	56.81	67.85	13.13	34.84	44.73	14.82	49.02	80.37	91.47	91.72	91.72
	Ensembled expansion methods												
	Query2Term*	57.26	55.93	67.18	13.07	36.18	45.48	14.60	50.12	81.13	83.53	83.86	83.86
	Query2Term-FS*	54.16	54.46	65.14	12.38	35.76	44.74	14.08	49.58	79.03	78.88	79.10	79.11
	Query2Term-PRF*	52.17	51.84	61.49	11.94	33.93	41.94	13.69	47.26	74.36	79.07	79.25	79.25
	Query2Doc*	63.59	61.74	72.41	13.98	40.81	51.31	15.37	53.94	84.78	91.28	91.49	91.49
	Query2Doc-FS*	64.05	62.10	72.79	13.88	41.02	51.55	15.47	54.23	84.90	92.29	92.29	92.29
	Query2Doc-PRF*	62.34	61.78	72.35	13.84	41.22	51.44	15.19	**54.28**	85.44	89.05	89.24	89.24
	CoT*	64.77	61.30	72.08	14.05	39.08	49.56	15.73	52.35	84.23	92.19	92.19	92.19
	CoT-PRF*	56.37	55.05	65.42	12.78	36.11	45.07	14.63	49.49	78.08	82.56	82.93	82.93
	MILL	63.80	**62.50**	**73.74**	**14.75**	**41.96**	**53.11**	**16.17**	54.26	**85.92**	91.69	91.81	91.81
TREC-DL-2020	No expansion	49.36	50.26	59.81	14.27	31.42	35.87	17.61	50.47	75.12	80.21	80.21	80.21
	Traditional expansion methods												
	Bo1	49.47	53.25	63.11	14.79	34.43	39.67	17.74	54.66	79.48	80.83	80.99	80.99
	KL	49.27	53.20	63.01	14.68	34.31	39.53	17.66	54.70	79.39	80.83	80.99	80.99
	RM3	50.43	54.02	63.47	14.93	35.13	40.22	17.89	55.80	79.94	78.49	78.59	78.59
	AxiomaticQE	49.36	50.26	59.81	14.27	31.42	35.87	17.61	50.47	75.12	80.21	80.21	80.21
	LLM-based expansion methods												
	Query2Term	50.12	52.43	62.27	13.12	33.06	38.49	17.39	54.61	79.07	78.74	78.77	78.78
	Query2Term-FS	47.80	49.16	60.50	13.33	30.16	35.59	15.82	50.22	78.76	79.38	79.83	79.83
	Query2Term-PRF	47.76	48.92	59.57	12.32	29.03	33.70	14.70	49.29	76.68	78.97	79.29	79.29
	Query2Doc	61.22	60.13	69.97	19.06	41.31	47.03	21.57	57.58	83.38	88.27	88.44	88.44
	Query2Doc-FS	61.45	59.30	69.40	18.94	39.75	45.27	21.65	56.30	82.57	90.32	90.37	90.38
	Query2Doc-PRF	55.28	57.60	67.09	17.00	38.21	43.49	19.74	58.50	82.57	84.22	84.49	84.49
	CoT	58.39	56.74	67.02	18.15	37.32	42.34	21.51	54.02	80.11	88.02	88.02	88.03
	CoT-PRF	60.81	58.41	67.47	19.02	39.27	44.04	21.71	56.84	80.49	89.00	89.00	89.00
	Ensembled expansion methods												
	Query2Term*	53.17	55.14	65.08	14.53	36.30	41.51	18.07	56.79	81.61	83.89	84.10	84.10
	Query2Term-FS*	50.95	52.11	62.42	13.80	33.47	38.34	16.98	53.55	79.38	81.77	81.81	81.81
	Query2Term-PRF*	50.80	53.44	63.68	14.09	34.12	39.29	17.41	55.25	81.14	79.89	80.14	80.14
	Query2Doc*	60.96	60.65	70.56	17.68	41.02	47.01	22.03	**59.88**	85.25	91.31	91.34	91.34
	Query2Doc-FS*	59.95	60.67	70.26	17.88	41.63	47.31	21.33	59.84	84.52	91.33	91.37	91.37
	Query2Doc-PRF*	**62.43**	60.59	70.53	18.32	41.54	47.44	**22.35**	59.63	84.93	91.42	91.45	91.45
	CoT*	59.90	59.15	69.35	17.16	39.59	45.50	20.57	57.50	84.22	92.19	92.19	92.21
	CoT-PRF*	59.75	58.41	68.81	17.75	38.89	44.56	20.63	56.09	83.67	91.20	91.27	91.27
	MILL	61.79	**61.15**	**71.23**	**19.05**	**41.76**	**48.17**	21.61	59.40	**85.27**	**92.61**	**92.71**	**92.72**

4.3 Implementation Details

We implement MILL and the baselines with PyTerrier (Macdonald and Tonellotto, 2020), a Python library helps conduct information retrieval experiments. For the BM25 retriever, we use the default parameters ($b = 0.75, k_1 = 1.2, k_3 = 8.0$) provided by PyTerrier (Macdonald and Tonellotto, 2020). For MILL and all the LLM-based baselines, we use the GPT-3.5-turbo-Instruct API (Brown et al., 2020) provided by OpenAI to generate contextual documents. The generation parameters are set as temperature $= 0.7$ and top_p $= 1$. We use the text-embedding-ada-002 provided by OpenAI as the text encoder, where the length of the returned vector is 1536. For other hyperparameters, we set the selection number of generated documents and PRF documents as 3, and the number of candidates as 5. To conduct a fair comparison for the LLM-

Table 2: Overall comparison on 9 datasets in BEIR on NDCG@1000. The optimal results are highlighted in bold, while the suboptimal results are underscored. The improvements are all significant (i.e., two-sided t-test with $p < 0.05$) between the optimal and suboptimal results.

Datasets	TREC-COVID	TOUCHE	SCIFACT	NFCORPUS	DBPEDIA	FIQA-2018	SCIDOCS	ARGUANA	CLIMATE-FEVER
No expansion	42.04	55.32	70.27	30.02	38.70	35.28	25.14	39.93	21.73
Traditional expansion methods									
Bo1	44.73	56.62	68.34	37.01	39.05	34.97	26.14	39.42	23.11
KL	44.88	56.72	67.83	37.18	38.87	35.12	26.15	39.31	23.07
RM3	44.54	55.79	65.28	37.27	38.11	33.14	25.91	38.14	20.71
AxiomaticQE	42.06	55.32	70.28	30.02	38.70	35.28	25.14	39.88	21.75
LLM-based expansion methods									
Query2Term	42.48	52.95	69.57	33.82	33.51	32.12	25.11	39.33	27.23
Query2Term-FS	41.13	57.10	71.39	38.57	39.36	35.78	26.18	39.72	24.32
Query2Term-PRF	39.90	53.72	60.79	38.21	34.83	31.50	24.97	38.68	23.98
Query2Doc	47.19	60.32	71.19	38.76	44.79	37.63	27.40	39.84	**32.39**
Query2Doc-FS	46.34	59.99	71.89	38.09	45.11	37.96	27.18	39.92	32.05
Query2Doc-PRF	43.87	56.84	67.82	39.41	39.85	34.09	26.16	38.85	26.90
CoT	49.32	60.77	71.63	38.88	43.05	37.28	27.50	40.00	30.25
CoT-PRF	46.53	59.03	73.65	39.84	40.43	38.04	26.23	40.01	25.78
Ensembled expansion methods									
Query2Term*	48.20	59.46	68.12	41.20	39.12	35.22	26.67	39.45	26.86
Query2Term-FS*	46.54	58.03	67.96	41.12	37.78	34.60	26.01	39.58	24.97
Query2Term-PRF*	45.61	56.74	64.80	39.33	36.85	33.27	25.88	39.20	23.98
Query2Doc*	50.26	61.87	71.49	41.33	44.06	37.05	27.49	39.49	30.67
Query2Doc-FS*	50.42	62.10	72.03	41.24	44.22	37.05	27.41	39.34	29.78
Query2Doc-PRF*	50.55	61.74	71.86	41.20	44.47	36.82	27.47	38.49	26.25
CoT*	50.69	61.69	72.33	41.08	42.29	38.13	27.62	39.59	29.74
CoT-PRF*	47.29	59.31	70.88	40.43	38.95	36.18	26.53	39.15	25.48
MILL	**52.53**	**62.15**	**74.14**	**41.75**	**46.39**	**39.23**	**28.36**	**40.11**	30.66

based baselines, we generate 3 expanded queries for each baseline and concatenate them as the final expansion result. The number of PRF documents for ensembled expansion methods is 3.

4.4 Main Results

Tables 1, 2, and 3 show the experimental results. The full results for the 9 selected datasets in BEIR are listed in Appendix A.4. We can draw the following key findings:

- Traditional query expansion methods exhibit positive effects for retrieval, while these carefully designed methods are outperformed by Query2Doc and CoT variants by a large margin. This implies that LLM-based methods are more promising for the query expansion task.

- Among LLM-based methods, CoT and Query2Doc variants are superior than Query2Term variants. The reason is that generated documents contain more contextualized information than discrete keywords.

- Using pseudo-relevance documents and few-shot examples as instructions in LLM-based methods does not necessarily yield positive gains. For instance, Query2Doc-PRF is worse than Query2Doc in TREC-DL-2019 and TREC-DL-2020. This shows that the query expansion task is non-trivial to be aligned to a specific corpus with straightforward prompting techniques.

- Ensembled expansion methods (e.g., Query2Doc*) are usually better than LLM-based expansion methods (e.g., Query2Doc), which demonstrates the importance of PRF documents in query expansion. Moreover, MILL is able to outperform the ensembled baselines on most metrics and datasets, as it adopts a more effective combination of generated and PRF documents.

- MILL is more effective than all the baselines in general, it always achieves either the best or the second best performance on all metrics and datasets in Table 1 and 2. It is also worth noting that MILL is a zero-shot method that is more applicable in various real-world applications.

4.5 Ablation Study

We design the following variants of MILL to conduct the ablation study:

- **w/o PRF**: Using QQD to generate expansion directly, without any PRF documents in expansion.

- **w/o MV**: Concatenating PRF documents to the QQD expansion We directly use K' top-retrieved documents of the original query as $\mathcal{D}_s^{\text{PRF}}$, without reranking and selection using generated documents $\mathcal{D}^{\text{LLM}}$.

Table 3: Overall comparison on 9 datasets in BEIR on Recall@1000. The optimal results are highlighted in bold, while the suboptimal results are underscored. The improvements are all significant (i.e., two-sided t-test with $p < 0.05$) between the optimal and suboptimal results.

Datasets	TREC-COVID	TOUCHE	SCIFACT	NFCORPUS	DBPEDIA	FIQA-2018	SCIDOCS	ARGUANA	CLIMATE-FEVER
No expansion	40.52	85.05	97.00	36.06	63.61	77.42	55.04	98.58	57.63
Traditional expansion methods									
Bo1	43.64	86.00	97.67	54.38	64.90	79.18	57.47	**98.65**	60.22
KL	43.63	**86.14**	97.67	54.79	64.71	78.84	57.38	**98.65**	60.01
RM3	43.71	85.79	97.67	56.12	64.37	78.82	57.88	98.08	58.18
AxiomaticQE	40.53	85.05	97.00	36.06	63.61	77.42	55.04	98.58	57.66
LLM-based expansion methods									
Query2Term	40.82	77.24	99.00	58.82	58.90	78.22	60.00	98.51	66.59
Query2Term-FS	40.34	85.33	98.33	61.72	65.67	81.84	60.15	98.51	62.87
Query2Term-PRF	39.50	83.29	97.50	60.55	61.11	76.31	59.25	**98.65**	63.79
Query2Doc	45.42	84.08	99.00	61.09	70.29	82.72	61.63	98.51	72.98
Query2Doc-FS	44.66	83.95	99.33	59.55	70.04	83.46	61.33	98.36	73.01
Query2Doc-PRF	42.53	83.50	99.00	62.50	66.41	79.14	59.50	98.58	67.15
CoT	47.27	84.42	98.67	60.63	69.24	83.56	60.90	98.44	69.86
CoT-PRF	44.93	84.37	98.67	59.87	66.06	82.14	58.72	98.58	64.26
Ensembled expansion methods									
Query2Term*	47.04	84.92	98.67	64.66	64.77	80.36	60.82	98.44	69.09
Query2Term-FS*	45.43	85.34	99.00	64.31	64.48	79.85	58.96	98.44	66.94
Query2Term-PRF*	44.67	85.42	98.83	60.80	63.37	78.84	59.61	98.44	63.79
Query2Doc*	48.43	85.49	99.33	64.23	69.95	82.47	61.32	**98.65**	**73.51**
Query2Doc-FS*	48.70	85.21	99.33	64.70	70.23	82.40	61.05	**98.65**	72.82
Query2Doc-PRF*	48.92	84.94	99.33	63.97	70.19	82.43	61.81	98.58	66.42
CoT*	48.87	85.16	99.33	63.82	67.63	83.32	61.69	98.51	71.64
CoT-PRF*	45.97	85.76	99.00	62.34	64.55	81.35	59.35	98.58	64.30
MILL	**50.55**	85.21	**99.67**	**64.95**	**71.13**	**84.23**	**61.86**	98.44	71.09

Table 4: Ablation study of MILL on TREC-DL-2020, TREC-COVID and SCIFACT.

Methods	NDCG @1000	AP @1000	Recall @1000	MRR @1000
TREC-DL-2020				
w/o PRF	70.65	48.10	**85.97**	89.10
w/o MV	70.28	46.73	85.11	90.75
w/o QQD	69.46	47.39	83.98	87.69
MILL	**71.23**	**48.17**	85.27	**92.72**
TREC-COVID				
w/o PRF	51.17	27.35	49.09	**92.40**
w/o MV	51.73	28.44	50.00	87.40
w/o QQD	50.84	27.30	49.16	89.08
MILL	**52.53**	**29.30**	**50.55**	91.17
SCIFACT				
w/o PRF	73.01	65.58	**99.67**	66.54
w/o MV	72.43	64.79	**99.67**	65.71
w/o QQD	71.13	62.96	**99.67**	63.98
MILL	**74.14**	**66.88**	**99.67**	**68.09**

- **w/o QQD**: Replacing QQD prompt in MILL with Query2Doc prompt.

Table 4 shows the results of the ablation study on three datasets, where we can draw the following conclusions: (1) All the three components of MILL have significant contributions to the final performance, (2) **MILL** is better than **w/o QQD**, which demonstrates the effectiveness of our proposed QQD prompt. This shows that QQD prompt can effectively leverage the reasoning capabilities of LLMs, assisting LLMs to reveal more diverse and specific search intent, (3) **MILL** is superior to **w/o MV**, which verifies the effectiveness of the mutual verification. By mutually selecting the generated and pseudo-relevance documents, it effectively mitigates the corpus unalignment problem of LLMs and compensates for the inaccurate search intent of conventional pseudo-relevance documents, and (4) Compared to **w/o PRF** and **w/o MV**, **MILL** shows a more significant improvement on BEIR datasets than on TREC-DL-2020. It may indicate that, in specialized domains, mutual verification can more effectively enhance query expansion performance through the use of PRF documents. More results can be found in Appendix A.5.

4.6 Varying the Number of Documents

In the aforementioned experiments, the default number of candidate (i.e., both generated and retrieved) documents is set to $K = N = 5$, and the number of final selected documents is set to $K' = N' = 3$. In this subsection, we vary the number of candidates and selected documents and report the performance of MILL on TREC-COVID, w.r.t. NDCG@1000 and MRR@1000. More details and results on additional datasets can be found in Appendix A.6.

From Figure 3, we have observations: (1) More selected pseudo-relevance documents decrease MRR@1000 dramatically. This shows that

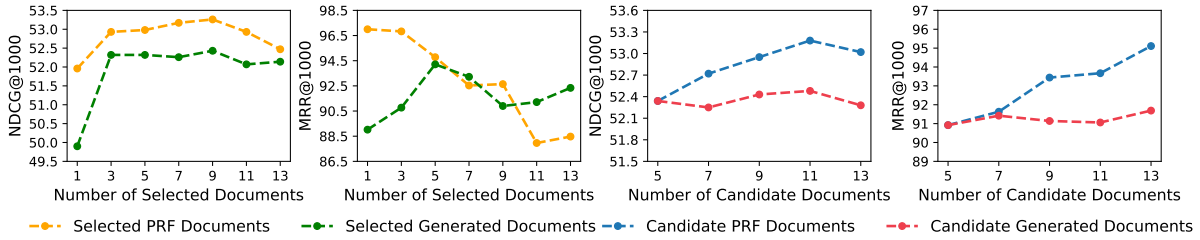

Figure 3: Varying the number of candidate and selected documents on TREC-COVID.

Table 5: A query expansion example from the TREC-DL-2020 dataset using MILL.

Query	What is a nonconformity earth science
Expansion Content	Pseudo-relevance document: There are three types of unconformities(below): angular **unconformity**, **nonconformity**, and disconformity. Anangular **unconformity** (left) is an erosional **surface** separating steeply dipping **rock** layers below from gently dipping layers above. **Unconformity** is the general term for **missing** pages of Earth history. LLM-generated document: **1.** What is the definition of a **nonconformity** in earth science? A **nonconformity** in earth science refers to a type of **unconformity**, which is a **break in the geologic record** where layers of **rock** are **missing**. Specifically, a **nonconformity** is where **sedimentary rocks** are deposited on top of **igneous or metamorphic rocks**. **2.** How are **nonconformities** formed inearth science? **Nonconformities** are formed throughthe process of **erosion**, where layers of **rock** are gradually wornaway. This exposes the underlying **igneous or metamorphic rocks**, which are then **covered by** new **sedimentary rocks** over time. **3.** What are the different types of **nonconformities** in earth science? There are three main types of **nonconformities**: angular, **nonconformity**, and disconformity. Angular **nonconformities occur** when **sedimentary rocks** are deposited at an angle on top of tilted or folded **igneous or metamorphic rocks**. **Nonconformities occur** when **sedimentary rocks** are deposited on top of **eroded igneous ormetamorphic rocks**. Disconformities **occur** when there is a gap in the **sedimentary rock** layers, indicating **a period of erosion** or non-deposition. **4.** What are some examples of **nonconformities found** in the Earth's geology? **One well-known example** of a **nonconformity** is the **Grand Canyon** in Arizona, USA. The **sedimentary layers** seen in the **canyon** were deposited on top of tilted and folded igneous and **metamorphic rocks**, indicating a long history of **erosion and deposition**.
Ground Truth	**Nonconformities** are **unconformities** that separate **igneous or metamorphic rocks** from overlying **sedimentary rocks**. They usually indicate that along **period of erosion occurred** prior to **deposition of the sediments** (several km of **erosion** necessary). They are a feature of stratified **rocks**, and are therefore usually found in **sediments** (but may also occur in stratified volcanics). They are **surfaces** between two **rock** bodies that constitute a substantial **break (hiatus) in the geologic record** (sometimes people say inaccurately that time is **missing**). **Nonconformity**. When **igneous or metamorphic rocks** are **eroded** and then **covered by** younger **sedimentary rocks**, the contact is called a **nonconformity**. **One of the most famous** of these is **found in the Grand Canyon**, where the oldest **sedimentary rocks** are more than a billion years younger than the 1.6 billion-year-old **metamorphic rocks** on which they rest.
Filtered-out PRF document (ranked #1 by BM25 with the original query)	Definition of nonconformance in the AudioEnglish.org Dictionary. Meaning of non-conformance. What does nonconformance mean? Proper usage of the word nonconformance. Information about nonconformance in the AudioEnglish.org dictionary, synonyms, and antonyms.
Filtered-out LLM-generated document	... **7.** What other geological features are commonly associated with nonconformities in earth science? Nonconformities are often found alongside other geological features, such as faults, folds, and intrusions, which can all provide additional information about the Earth's history and the processes that have shaped it. **8.** How can nonconformities in earth science be identified in the field? Nonconformities can be identified by looking for the distinct contact between two different rock types, as well as the difference in age between the two layers. Geologists may also use specialized tools, such as radiometric dating, to determine the age of the rocks. **9.** Are nonconformities only found on land in earth science? No, nonconformities can also be found underwater in the oceans, where layers of sedimentary rock are exposed and show similar

more selected pseudo-relevance documents usually bring more noise to query expansion. In contrast, the generated documents are rather robust, where more selections do not significantly undermine the performance. (2) When we introduce more candidate documents, the mutual verification framework is able to effectively select pseudo-relevance documents, where both NDCG@1000 and MRR@1000 increase. This shows that LLM-generated documents are very useful for filtering out noisy pseudo-relevance documents. On the other hand, more generated candidate documents do not bring further performance gain, when the number of selected documents is fixed.

4.7 Case Study

We show an illustrative example in Table 5, which contains the original query, the expansion content, and the ground truth (i.e., the human-labeled relevant document). The words of ground truth passage that appear in the pseudo-relevance document are highlighted in bold, and those in the generated documents of different sub-queries are marked with different colors. We can see that the generated document is able to provide more useful information for identifying the ground truth passage. We also show the filtered-out PRF document, and the filtered-out LLM-generated document in the table, from which we can observe that the filtered-out

documents seem to be 1) PRF documents with limited information and 2) LLM-generated documents with too much extension of the original query. The mutual verification process can filter out these noisy or uninformative documents for MILL.

5 Related Work

Query Expansion. Query expansion is a prevalent technique in search platforms, which restructures the original query to more accurately express search intent and enhance the alignment with corpus (Bhogal et al., 2007; Carpineto and Romano, 2012; Efthimiadis, 1996). Early studies employed lexical knowledge bases (Qiu and Frei, 1993; Voorhees, 1994) or Pseudo-relevance Feedback (PRF) (Amati and Van Rijsbergen, 2002; Robertson, 1990; Rocchio Jr, 1971; Lv and Zhai, 2010; Yan et al., 2003; Li et al., 2022) for expanding the query with additional information. PRF documents can supplement information for any query, but they also encounter the issue of misalignment with the original query (Jagerman et al., 2023).

Recently, the integration of LLMs with information retrieval has emerged as a prominent area of research (Li et al., 2023a; Dong et al., 2023; Sun et al., 2023; Ni et al., 2021; Zhuang et al., 2023; Dai et al., 2022; Bonifacio et al., 2022; Muennighoff, 2022), where LLM-based query expansion methods have also been proposed. In particular, Query2Doc (Wang et al., 2023a) proposes a query-document prompt, leveraging the semantic understanding and generative capabilities of LLMs to extend the original query. Another recent study (Jagerman et al., 2023) applies LLMs directly for query expansion across multiple datasets, finding that employing the chain of thoughts (CoT) (Wei et al., 2022b) approach achieves the best results. Moreover, LLMCS (Mao et al., 2023) applies LLMs for query expansion in conversational search, constructing the context search intents as a prompt and combining the chain of thoughts and self-consistency techniques to enhance search performance. In our paper, we focus on alleviating the limitations of both PRF-based and generation-based method. We propose a query-query-document generation method and a mutual verification framework to effective leverage both retrieved and generated contextual documents.

Large Language Models. LLMs have strong and robust abilities in language understanding and generation (Kojima et al., 2022; Huang et al., 2022; Liu et al., 2023; Wang et al., 2023b; Xu et al., 2024a; Wang et al., 2024; Liu et al., 2024; Li et al., 2023b; Xu et al., 2024b, 2023; Fan et al., 2023), especially with increased model parameters (Zhao et al., 2023; Jagerman et al., 2023; Wei et al., 2022a). LLMs have the instruction-following ability (Longpre et al., 2023; Wei et al., 2021) and can be boasted through a few contexts (Min et al., 2022; Dong et al., 2022), enhancing the performance of LLMs in downstream specific tasks. Moreover, these methods are straightforward and effective, for they require minimal human effort to provide instructions or in-context examples but reach good results. For example, Flan-T5 (Chung et al., 2022) achieves remarkable results in various NLP downstream tasks by instruction tuning the base model. Recently, many studies (Wei et al., 2022b; Besta et al., 2023; Yao et al., 2023; Wang et al., 2022) explored the reasoning capabilities of LLMs and discovered that LLMs are powerful zero-shot reasoners.

6 Conclusion

In this paper, we propose a novel zero-shot LLMs-based framework for query expansion. First, we design a QQD prompt scheme that allows LLMs to generate diverse contextual documents via zero-shot reasoning. Next, we introduce a mutual verification method that allows retrieved and generated contextual documents to complement each other as query expansion. The experimental results show that our method is superior to the state-of-the-art baselines on three public datasets.

Acknowledgments

This research was partially supported by Research Impact Fund (No.R1015-23), APRC - CityU New Research Initiatives (No.9610565, Start-up Grant for New Faculty of CityU), CityU - HKIDS Early Career Research Grant (No.9360163), Hong Kong ITC Innovation and Technology Fund Midstream Research Programme for Universities Project (No.ITS/034/22MS), Hong Kong Environmental and Conservation Fund (No. 88/2022), and SIRG - CityU Strategic Interdisciplinary Research Grant (No.7020046, No.7020074), Ant Group (CCF-Ant Research Fund, Ant Group Research Fund), Huawei (Huawei Innovation Research Program), Tencent (CCF-Tencent Open Fund, Tencent Rhino-Bird Focused Research Program), CCF-BaiChuan-Ebtech Foundation Model Fund, and Kuaishou.

7 Limitations

One limitation of our work is the retrieval efficiency. On one hand, during retrieval, MILL needs to perform multiple autoregressive generations for each query based on the query-query-document prompt, and then use mutual verification methods with PRF documents to obtain selected documents. On the other hand, the extended length of the query increases the time required to search the inverted index. To address the issue of multi-round autoregressive generation, N generated documents can be produced in parallel, which will improve generation efficiency. Regarding the issue of extended query length, we can further utilize simple rule-based filtering methods (e.g., deleting words with limited semantic information or truncating documents with word counts) to compress the query.

In addition, from the experiments conducted on the BEIR datasets, we can observe that MILL does not perform well on some metrics for the ARGUANA and CLIMATE-FEVER datasets. This may indicate the limitations of MILL in some scenarios. For ARGUANA, we notice that the queries have 193 words on average, which is roughly 10 to 20 times more words than other BEIR datasets. Thus, it might not necessarily need query expansion, which limits the improvement of MILL. For CLIMATE-FEVER, we observe that the queries are often declarative sentences, rather than specific questions. In such cases, the QQD approach is more likely to generate off-the-topic subqueries, which undermines the effectiveness of the final query expansion. These observations suggest that MILL could have different performances on different kinds of queries, which will be more comprehensively studied in the future.

References

Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. 2004. Umass at trec 2004: Novelty and hard. Computer Science Department Faculty Publication Series, page 189.

Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS), 20(4):357–389.

Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query expansion techniques for information retrieval: a survey. Information Processing & Management, 56(5):1698–1735.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. arXiv preprint arXiv:2308.09687.

Jagdev Bhogal, Andrew MacFarlane, and Peter Smith. 2007. A review of ontology based query expansion. Information processing & management, 43(4):866–886.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. arXiv preprint arXiv:2202.05144.

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.

Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. 2008. Selecting good expansion terms for pseudo-relevance feedback. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 243–250.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. Acm Computing Surveys (CSUR), 44(1):1–50.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416.

Vincent Claveau. 2021. Neural text generation for query expansion in information retrieval. In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pages 202–209.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the trec 2020 deep learning track. arXiv preprint arXiv:2102.07662.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. arXiv preprint arXiv:2209.11755.

Qian Dong, Yiding Liu, Qingyao Ai, Zhijing Wu, Haitao Li, Yiqun Liu, Shuaiqiang Wang, Dawei Yin, and Shaoping Ma. 2023. Aligning the capabilities of

large language models with the context of information retrieval via contrastive feedback. arXiv preprint arXiv:2309.17078.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. arXiv preprint arXiv:2301.00234.

Efthimis N Efthimiadis. 1996. Query expansion. Annual review of information science and technology (ARIST), 31:121–87.

Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms).

Hui Fang and ChengXiang Zhai. 2006. Semantic term matching in axiomatic approaches to information retrieval. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 115–122.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. arXiv preprint arXiv:2210.11610.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. arXiv preprint arXiv:2305.03653.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213.

Hang Li, Shengyao Zhuang, Ahmed Mourad, Xueguang Ma, Jimmy Lin, and Guido Zuccon. 2022. Improving query representations for dense retrieval with pseudo relevance feedback: A reproducibility study. In European Conference on Information Retrieval, pages 599–612. Springer.

Xiaopeng Li, Lixin Su, Pengyue Jia, Xiangyu Zhao, Suqi Cheng, Junfeng Wang, and Dawei Yin. 2023a. Agent4ranking: Semantic robust ranking via personalized query rewriting using multi-agent llm.

Xinhang Li, Chong Chen, Xiangyu Zhao, Yong Zhang, and Chunxiao Xing. 2023b. E4srec: An elegant effective efficient extensible solution of large language models for sequential recommendation.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2023. Moelora: An moe-based parameter efficient fine-tuning method for multi-task medical applications.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Zijian Zhang, Feng Tian, and Yefeng Zheng. 2024. Large language model distilling medication recommendation model.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. arXiv preprint arXiv:2301.13688.

Yuanhua Lv and ChengXiang Zhai. 2010. Positional relevance model for pseudo-relevance feedback. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pages 579–586.

Craig Macdonald and Nicola Tonellotto. 2020. Declarative experimentation ininformation retrieval using pyterrier. In Proceedings of ICTIR 2020.

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. arXiv preprint arXiv:2303.06573.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In EMNLP.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. arXiv preprint arXiv:2202.08904.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. arXiv preprint arXiv:2108.08877.

Yonggang Qiu and Hans-Peter Frei. 1993. Concept based query expansion. In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pages 160–169.

Stephen E Robertson. 1990. On term selection for query expansion. Journal of documentation, 46(4):359–364.

Joseph John Rocchio Jr. 1971. Relevance feedback in information retrieval. The SMART retrieval system: experiments in automatic document processing.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. arXiv preprint arXiv:2304.09542.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Ellen M Voorhees. 1994. Query expansion using lexical-semantic relations. In SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University, pages 61–69. Springer.

Hanbing Wang, Xiaorui Liu, Wenqi Fan, Xiangyu Zhao, Venkataramana Kini, Devendra Yadav, Fei Wang, Zhen Wen, Jiliang Tang, and Hui Liu. 2024. Rethinking large language model architectures for sequential recommendations.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models. arXiv preprint arXiv:2303.07678.

Maolin Wang, Yao Zhao, Jiajia Liu, Jingdong Chen, Chenyi Zhuang, Jinjie Gu, Ruocheng Guo, and Xiangyu Zhao. 2023b. Large multimodal model compression via efficient pruning and distillation at antgroup.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023. Large language models for generative information extraction: A survey.

Derong Xu, Ziheng Zhang, Zhenxi Lin, Xian Wu, Zhihong Zhu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. 2024a. Multi-perspective improvement of knowledge graph completion with large language models.

Derong Xu, Ziheng Zhang, Zhihong Zhu, Zhenxi Lin, Qidong Liu, Xian Wu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. 2024b. Editing factual knowledge and explanatory ability of medical large language models.

Rong Yan, Alexander Hauptmann, and Rong Jin. 2003. Multimedia search with pseudo-relevance feedback. In Image and Video Retrieval: Second International Conference, CIVR 2003 Urbana-Champaign, IL, USA, July 24–25, 2003 Proceedings 2, pages 238–247. Springer.

Peilin Yang and Jimmy Lin. 2019. Reproducing and generalizing semantic term matching in axiomatic information retrieval. In Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41, pages 369–381. Springer.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. arXiv preprint arXiv:2303.18223.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. arXiv preprint arXiv:2308.07107.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2308–2313.

A Appendix

A.1 Baselines

Traditional query expansion methods

- **Bo1 (Amati and Van Rijsbergen, 2002).** The Bose-Einstein 1 (Bo1) weighting approach is a method that reconstructs the query based on the frequency of terms found in the feedback documents associated with each query.
- **KL (Amati and Van Rijsbergen, 2002).** This method rewrites the queries similar to Bo1 but based on Kullback Leibler divergence.
- **RM3 (Abdul-Jaleel et al., 2004).** A method used for query expansion in information retrieval, which finds the most relevant terms to the query by using the top-ranked documents returned from the initial query and adds these terms to the original query to create an expanded query.
- **AxiomaticQE (Fang and Zhai, 2006; Yang and Lin, 2019).** Axiomatic query expansion (AxiomaticQE) rewrites and expands the origin query by axiomatic semantic term matching.

LLM-based expansion methods

- **Query2Term.** It uses LLMs to generate related terms to the origin query in a zero-shot manner. The zero-shot prompts only contain task instructions and the original query.
- **Query2Term-FS.** The few-shot version of Query2Term. The few-shot prompts are built upon zero-shot prompts by adding a few examples. In particular, Query2Term-FS expands upon Query2Term by incorporating additional sets of query-keywords examples.
- **Query2Term-PRF.** It uses the top-3 documents retrieved by the original query as context information to instruct the LLMs to expand the original query.
- **Query2Doc.** The zero-shot version of query2doc (Wang et al., 2023a), whose structure is similar to Query2Term. It uses LLMs to generate related passages to the origin query.
- **Query2Doc-FS.** The few-shot version of query2doc (Wang et al., 2023a). The prompt structure is similar to Query2Term-FS.
- **Query2Doc-PRF.** It constructs the prompt with pseudo-relevance feedback in a zero-shot manner based on Query2Doc-ZS, like the Query2Term-PRF.
- **CoT.** Chain-of-Thought (CoT) (Jagerman et al., 2023) instructs LLMs to generate text step by step, providing a detailed thought process before generating the final answer.
- **CoT-PRF.** A pseudo-relevance feedback based version of CoT similar to Query2Term-PRF.

Ensembled expansion methods

The ensembled expansion methods contain Query2Term*, Query2Term-FS*, Query2Term-PRF*, Query2Doc*, Query2Doc-FS*, Query2Doc-PRF*, CoT*, CoT-PRF*. They are the variants to the corresponding LLM-based expansion methods by directly concatenating the top-k PRF documents to the expanded query.

A.2 Prompts

Figure 6 shows the prompts for the variants of Query2Term. The core prompt is "Write some keywords for the given query: {query}."

Table 6: Prompts for Query2Term and its variants.

Method	Prompt
Query2Term	Write some keywords for the given query: {query}
Query2Term-FS	Write some keywords for the given query: Context: query:{query1} keywords:{keywords1} query:{query2} keywords: {keywords2} query: {query3} keywords:{keywords3} query: {query} keywords:
Query2Term-PRF	Write some keywords for the given query: Context: {PRF doc 1} {PRF doc 2} {PRF doc 3} query: {query} keywords:

Figure 7 shows the prompts for the Query2Doc variants. The main prompts are the sentence: "Write a passage answer the following query: {query}."

For the CoT and its variants, their prompts are in Figure 8. The prompts ask LLMs to give the rationale before answering.

Table 7: Prompts for Query2Doc and its variants.

Method	Prompt
Query2Doc	Write a passage answer the following query: {query}
Query2Doc-FS	Write a passage answer the following query: Context: query:{query1} passage:{passage1} query:{query2} passage: {passage2} query: {query3} passage:{passage3} query: {query} passage:
Query2Doc-PRF	Write a passage answer the following query: Context: {PRF doc 1} {PRF doc 2} {PRF doc 3} query: {query} passage:

Table 8: Prompts for CoT and its variants.

Method	Prompt
CoT	Answer the following query: {query} Give the rationale before answering.
CoT-PRF	Answer the following query: Context: {PRF doc 1} {PRF doc 2} {PRF doc 3} query: {query} Give the rationale before answering.

A.3 Results on MSMARCO

Table 9 shows the experimental results on MSMARCO dataset. MSMARCO[3] (Nguyen et al., 2016) is a collection of datasets constructed to advance the development of deep learning in the search field. We choose the passage dataset as our experimental scenario and take the first 100 queries from the dev group as the test queries. Results in Table 9 are based on the LLM text-davinci-003 provided by OpenAI.

A.4 More Results on BEIR

In this section, we list the full results for the 9 selected datasets from BEIR. Specifically, they are TREC-COVID, TOUCHE, SCIFACT, NFCORPUS, DBPEDIA, FIQA-2018, SCIDOCS, ARGUANA, and CLIMATE-FEVER. The optimal results are highlighted in bold, while the suboptimal results are underscored. The results are reported on NDCG@N, AP@N, Recall@N, and MRR@N with N (10, 100, 1000)

[3]https://microsoft.github.io/msmarco/

A.5 More Results for Ablation Experimetns

From the results shown in Table 19, we can draw some findings that MILL consistently achieves better performance than w/o PRF, w/o MV, and w/o QQD on all three datasets. This validates the effectiveness of both QQD and mutual verification across different datasets.

A.6 More Results for Experiments with Various Numbers of Documents

In this subsection, we will supplement the results on other metrics for the experiments with various numbers of documents. We use the gpt-3.5-turbo-instruct API provided by OpenAI to conduct these experiments.

The experiments concerning the number of selected documents are shown in Figure 4 and Figure 5. When the number of selected generated documents changes, the number of candidate generated documents remains 15, and the number of PRF candidate documents and the number of selected PRF documents remain 5 and 3. When the number of selected PRF documents changes, the number of candidate PRF documents remains 15, and the number of generated candidate documents and the number of selected generated documents remain 5 and 3. We can find that the trends of selected PRF documents in NDCG, AP, and Recall are consistent, yet contrary to that of MRR. This is due to the fact that NDCG, AP, and Recall are more comprehensive indicators, whereas MRR only considers the ranking of the topmost relevant document retrieved.

In the experiments regarding the number of candidate documents, as shown in Figure 6 and Figure 7, we can observe a similar trend across different metrics: as the number of generated document candidates increases, the metrics remain relatively stable. However, with an increase in the number of PRF document candidates, there is a noticeable growth in the metrics. This suggests that a specific number of generated documents, such as 5, can almost entirely cover the additional information provided by the generation process to aid in understanding the search intent of the original query. Meanwhile, PRF documents, derived from searches based on the original query, suggest that more PRF document candidates can cover a wider range of possible search intents, thereby enhancing the effectiveness of query expansion.

Table 9: Overall comparison on MSMARCO. The optimal results are highlighted in bold, while the suboptimal results are underscored. The results are reported on NDCG@N, AP@N, Recall@N, and MRR@N with N $\in \{10, 100, 1000\}$. The improvements are all significant (i.e., two-sided t-test with $p < 0.05$) between the optimal and suboptimal results.

Metrics	NDCG			AP			Recall			MRR		
	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
No expansion	28.69	34.02	36.23	23.56	24.65	24.72	44.50	69.00	86.50	22.65	23.76	23.83
Traditional expansion methods												
Bo1	29.18	33.44	35.89	23.61	24.33	24.43	46.50	67.50	86.50	24.07	24.82	24.91
KL	29.20	33.59	36.17	23.93	24.73	24.83	45.50	66.50	86.50	**24.39**	**25.22**	**25.31**
RM3	26.93	32.23	34.34	21.81	22.87	22.94	42.50	67.00	83.50	22.25	23.33	23.41
AxiomaticQE	28.69	34.02	36.23	23.56	24.65	24.72	44.50	69.00	86.50	22.65	23.76	23.83
LLM-based expansion methods												
Query2Term	23.28	29.50	32.00	19.74	21.01	21.08	34.17	63.17	83.67	19.91	21.17	21.24
Query2Term-FS	24.26	29.76	32.07	20.41	21.43	21.50	36.33	62.50	81.33	20.78	21.87	21.94
Query2Term-PRF	21.56	27.02	29.26	16.04	17.05	17.12	38.67	64.83	83.33	16.04	17.11	17.17
Query2Doc	25.83	31.31	33.82	20.27	21.33	21.42	43.50	69.00	88.83	20.39	21.50	21.58
Query2Doc-FS	28.23	33.22	35.89	23.10	23.99	24.09	44.67	68.83	89.50	23.00	23.94	24.04
Query2Doc-PRF	25.45	29.99	32.36	20.31	21.25	21.33	41.44	62.50	81.17	20.45	21.35	21.43
CoT	26.13	31.84	34.25	21.38	22.44	22.54	41.00	68.33	86.83	21.47	22.55	22.64
CoT-PRF	28.93	34.17	36.32	23.51	24.52	24.60	46.12	70.87	87.50	23.64	24.69	24.77
MILL	**29.99**	**34.92**	**37.26**	**24.01**	**24.98**	**25.07**	**48.67**	**71.67**	**89.83**	24.02	25.02	25.10

Table 10: Overall experimental results on TREC-COVID.

Metrics	NDCG			AP			Recall			MRR		
	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
No expansion	62.59	47.41	42.04	1.46	8.16	19.79	1.74	11.91	40.52	83.37	83.37	83.37
Traditional expansion methods												
Bo1	64.82	49.5	44.73	1.56	8.8	22.01	1.77	12.48	43.64	86.62	86.77	86.77
KL	65.8	49.93	44.88	1.59	8.9	22.26	1.79	12.51	43.63	86.62	86.79	86.79
RM3	64.05	48.5	44.54	1.55	8.62	21.87	1.78	11.22	43.71	82.96	83.06	83.06
AxiomaticQE	62.74	47.45	42.06	1.47	8.17	19.81	1.74	11.91	40.53	84.37	84.37	84.37
LLM-based expansion methods												
Query2Term	65.66	48.53	42.48	1.58	8.50	19.79	1.80	11.73	40.82	84.83	84.83	84.86
Query2Term-FS	57.78	44.80	41.13	1.39	8.35	20.32	1.60	11.60	40.34	77.31	77.59	77.61
Query2Term-PRF	56.55	41.64	39.90	1.37	6.73	17.90	1.56	9.89	39.50	80.95	81.34	81.34
Query2Doc	70.95	53.17	47.19	1.77	9.89	23.78	1.98	13.25	45.42	88.79	88.79	88.79
Query2Doc-FS	68.38	51.29	46.34	1.72	9.48	22.75	1.94	12.86	44.66	86.03	86.12	86.12
Query2Doc-PRF	63.98	49.41	43.87	1.54	8.79	21.86	1.76	12.07	42.53	81.40	81.55	81.55
CoT	**76.31**	56.54	49.32	1.98	10.87	25.51	2.18	14.24	47.27	89.38	89.38	89.38
CoT-PRF	68.23	52.46	46.53	1.71	9.49	23.31	1.91	12.88	44.93	90.20	90.20	90.20
Ensembled expansion methods												
Query2Term*	68.04	54.26	48.20	1.71	10.12	25.17	1.94	13.66	47.04	82.90	82.90	82.90
Query2Term-FS*	66.34	52.03	46.54	1.64	9.54	23.91	1.88	13.14	45.43	83.07	83.07	83.07
Query2Term-PRF*	65.00	50.49	45.61	1.60	8.95	22.86	1.81	12.41	44.67	85.40	85.40	85.40
Query2Doc*	72.73	57.78	50.26	1.83	11.28	27.08	2.09	14.58	48.43	86.87	86.87	86.87
Query2Doc-FS*	73.13	57.78	50.42	1.84	11.20	27.10	2.08	14.56	48.70	87.67	87.78	87.78
Query2Doc-PRF*	71.73	57.51	50.55	1.79	11.24	27.25	2.05	14.54	48.92	87.07	87.07	87.07
CoT*	73.90	57.91	50.69	1.89	11.31	27.02	2.14	14.62	48.87	88.47	88.47	88.47
CoT-PRF*	68.97	53.70	47.29	1.73	9.90	24.36	1.95	13.49	45.97	86.92	86.92	86.92
MILL	75.30	**60.24**	**52.53**	**2.03**	**12.22**	**29.30**	**2.22**	**15.40**	**50.55**	**91.17**	**91.17**	**91.17**

Table 11: Overall experimental results on TOUCHE.

Metrics	NDCG			AP			Recall			MRR		
	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
No expansion	34.28	45.48	55.32	13.06	20.96	22.47	20.69	54.92	85.05	62.28	62.71	62.71
Traditional expansion methods												
Bo1	35.62	46.98	56.62	14.19	22.19	23.69	21.35	56.47	86.00	63.54	64.07	64.07
KL	35.52	46.96	56.72	14.00	22.18	23.68	20.99	56.78	**86.14**	63.98	64.51	64.51
RM3	34.66	46.54	55.79	13.72	22.00	23.42	22.03	57.79	85.79	56.73	57.09	57.09
AxiomaticQE	34.28	45.48	55.32	13.06	20.96	22.47	20.69	54.92	85.05	62.28	62.71	62.71
LLM-based expansion methods												
Query2Term	34.51	44.05	52.95	13.11	19.88	21.13	20.05	49.98	77.24	65.60	66.13	66.14
Query2Term-FS	35.10	47.93	57.10	14.97	23.28	24.66	21.71	57.88	85.33	57.71	58.23	58.23
Query2Term-PRF	31.83	44.19	53.72	12.60	19.78	21.22	20.16	53.83	83.29	54.77	55.45	55.45
Query2Doc	42.36	51.12	60.32	17.44	25.51	26.91	23.80	56.10	84.08	75.63	75.97	75.97
Query2Doc-FS	40.71	51.30	59.99	16.91	25.72	27.02	23.01	57.46	83.95	70.84	71.06	71.06
Query2Doc-PRF	37.21	47.43	56.84	14.78	22.39	23.81	21.11	54.30	83.50	69.59	69.95	69.97
CoT	41.91	51.57	60.77	17.28	25.61	27.03	23.18	56.42	84.42	75.00	75.09	75.09
CoT-PRF	39.33	50.08	59.03	16.66	24.54	25.93	23.30	57.10	84.37	69.45	69.58	69.58
Ensembled expansion methods												
Query2Term*	40.25	50.78	59.46	15.96	24.86	26.25	24.10	58.54	84.92	69.54	69.54	69.54
Query2Term-FS*	37.69	49.41	58.03	15.27	23.93	25.31	23.81	58.67	85.34	60.65	60.94	60.94
Query2Term-PRF*	35.25	48.32	56.74	13.61	22.60	23.86	21.48	58.74	85.42	59.10	59.29	59.29
Query2Doc*	44.44	53.33	61.87	17.82	26.60	27.91	24.71	58.96	85.49	76.59	76.59	76.59
Query2Doc-FS*	43.98	53.91	62.10	17.75	26.74	28.02	24.28	**60.09**	85.21	**78.74**	**79.00**	**79.00**
Query2Doc-PRF*	43.51	53.60	61.74	17.45	26.45	27.74	24.19	59.59	84.94	76.89	77.22	77.22
CoT*	43.48	53.24	61.69	17.28	26.32	27.65	24.06	58.77	85.16	78.66	78.80	78.80
CoT-PRF*	39.82	50.73	59.31	15.75	24.58	25.88	23.70	58.96	85.76	67.94	68.10	68.10
MILL	**45.35**	**54.00**	**62.15**	**18.05**	**27.02**	**28.33**	**25.44**	59.97	85.21	77.04	77.23	77.23

Table 12: Overall experimental results on SCIFACT.

Metrics	NDCG			AP			Recall			MRR		
	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
No expansion	67.22	69.66	70.27	62.11	62.67	62.7	81.43	92.27	97	63.24	63.66	63.68
Traditional expansion methods												
Bo1	65.14	67.63	68.34	59.3	59.92	59.95	81.59	92.2	97.67	60.42	60.87	60.89
KL	64.68	67.08	67.83	58.69	59.28	59.31	81.59	91.87	97.67	59.76	60.18	60.21
RM3	62.22	64.54	65.28	55.45	55.97	55.99	81.34	91.93	97.67	56.24	56.58	56.61
AxiomaticQE	67.22	69.66	70.28	62.11	62.68	62.7	81.43	92.27	97	63.24	63.66	63.68
LLM-based expansion methods												
Query2Term	66.13	68.87	69.57	60.54	61.18	61.21	81.7	93.73	99	61.6	62.14	62.16
Query2Term-FS	68.34	70.71	71.39	62.92	63.5	63.54	83.32	93.47	98.33	64.13	64.6	64.62
Query2Term-PRF	57.67	59.91	60.79	49.72	50.22	50.25	80.46	90.9	97.5	50.58	50.93	50.96
Query2Doc	67.92	70.6	71.19	62.59	63.24	63.27	82.82	94.43	99	63.81	64.34	64.36
Query2Doc-FS	68.61	71.39	71.89	63.37	64.02	64.04	83.17	95.43	99.33	64.55	65.07	65.08
Query2Doc-PRF	64.53	66.96	67.82	58.6	59.15	59.19	81.31	92.53	99	59.74	60.12	60.15
CoT	68.58	71.13	71.63	63.3	63.87	63.89	83.03	**94.77**	98.67	64.77	65.18	65.19
CoT-PRF	70.98	72.95	73.65	66.2	66.64	66.67	84.56	93.27	98.67	67.09	67.47	67.49
Ensembled expansion methods												
Query2Term*	65.01	67.37	68.12	58.82	59.38	59.41	82.62	92.80	98.67	59.57	60.03	60.05
Query2Term-FS*	64.66	67.13	67.96	58.52	59.16	59.19	82.07	92.47	99.00	59.30	59.78	59.81
Query2Term-PRF*	61.61	63.76	64.80	54.61	55.17	55.20	81.43	90.73	98.83	55.69	56.04	56.08
Query2Doc*	68.42	70.74	71.49	62.94	63.52	63.55	83.71	93.70	99.33	64.03	64.51	64.53
Query2Doc-FS*	69.02	71.40	72.03	63.69	64.25	64.28	83.93	94.53	99.33	64.69	65.14	65.16
Query2Doc-PRF*	68.84	71.25	71.86	63.45	64.02	64.04	83.93	94.60	99.33	64.52	64.97	64.99
CoT*	69.53	71.63	72.33	64.00	64.53	64.56	84.88	94.03	99.33	65.35	65.74	65.77
CoT-PRF*	68.05	69.99	70.88	62.51	62.98	63.02	83.69	92.13	99.00	63.58	63.93	63.96
MILL	**71.37**	**73.47**	**74.14**	**66.34**	**66.85**	**66.88**	**85.24**	94.5	**99.67**	**67.69**	**68.07**	**68.09**

Table 13: Overall experimental results on NFCORPUS.

Metrics	NDCG			AP			Recall			MRR		
	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
No expansion	32.22	27.29	30.02	12.08	14.36	14.89	14.78	24.38	36.06	53.44	53.82	53.83
Traditional expansion methods												
Bo1	33.49	30.21	37.01	12.73	15.98	17.09	16.26	29.71	54.38	52.74	53.24	53.28
KL	33.56	30.22	37.18	12.73	15.89	17.01	16.3	29.61	54.79	53.49	53.99	54.03
RM3	33.41	30.31	37.27	12.36	15.68	16.8	16.82	30.46	56.12	52.35	52.81	52.85
AxiomaticQE	32.22	27.29	30.02	12.08	14.36	14.89	14.78	24.38	36.06	53.44	53.82	53.83
LLM-based expansion methods												
Query2Term	25.79	24.94	33.82	8.3	10.89	12.04	12.29	27.27	58.82	44.79	45.63	45.68
Query2Term-FS	31.92	30.66	38.57	11.24	14.63	15.91	15.38	32.83	61.72	52.99	53.68	53.71
Query2Term-PRF	32.14	29.92	38.21	11.92	15.01	16.29	16.78	31.63	60.55	49.27	49.83	49.87
Query2Doc	33.47	30.41	38.76	12.54	15.31	16.54	16.68	30.96	61.09	54.61	55.19	55.23
Query2Doc-FS	33.41	30.1	38.09	12.59	15.32	16.45	16.27	30.22	59.55	54.08	54.64	54.7
Query2Doc-PRF	33.82	31.23	39.41	12.64	16.17	17.44	16.97	32.7	62.5	51.26	51.72	51.77
CoT	34.52	30.68	38.88	12.95	15.78	16.93	16.88	29.53	60.63	56.23	56.64	56.69
CoT-PRF	35.76	31.93	39.84	**13.95**	16.9	18.09	18.13	31.76	59.87	55.65	56.05	56.09
Ensembled expansion methods												
Query2Term*	35.61	32.96	41.20	13.44	17.05	18.34	18.03	**34.46**	64.66	54.99	55.57	55.62
Query2Term-FS*	35.27	32.84	41.12	13.49	17.12	18.43	17.87	34.23	64.31	54.09	54.63	54.68
Query2Term-PRF*	33.92	31.27	39.33	12.94	16.28	17.55	17.53	32.72	60.80	51.44	52.04	52.08
Query2Doc*	35.76	32.74	41.33	13.62	17.04	18.40	17.79	33.30	64.23	56.48	57.00	57.05
Query2Doc-FS*	35.88	32.58	41.24	13.62	16.94	18.30	18.05	33.21	64.70	55.95	56.40	56.44
Query2Doc-PRF*	35.83	32.60	41.20	13.50	16.94	18.32	17.80	32.87	63.97	55.82	56.24	56.29
CoT*	36.05	32.51	41.08	13.93	17.13	18.45	17.99	32.04	63.82	56.03	56.51	56.55
CoT-PRF*	35.23	32.26	40.43	13.70	17.08	18.33	17.83	33.14	62.34	54.23	54.81	54.86
MILL	**36.79**	**33.02**	**41.75**	13.81	**17.18**	**18.56**	**18.21**	32.42	**64.95**	**58.35**	**58.86**	**58.91**

Table 14: Overall experimental results on DBPEDIA.

Metrics	NDCG			AP			Recall			MRR		
	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
No expansion	26.59	32.45	38.7	11.59	17.71	18.89	17.2	42.15	63.61	51.7	52.37	52.39
Traditional expansion methods												
Bo1	26.59	32.59	39.05	11.65	18.03	19.24	17.32	42.67	64.9	50.47	51.17	51.2
KL	26.42	32.44	38.87	11.52	17.89	19.09	17.27	42.62	64.71	50.01	50.84	50.86
RM3	25.47	31.81	38.11	10.88	17.4	18.6	17.05	42.92	64.37	46.6	47.28	47.31
AxiomaticQE	26.59	32.45	38.7	11.59	17.71	18.89	17.2	42.15	63.61	51.7	52.37	52.39
LLM-based expansion methods												
Query2Term	22.1	26.59	33.51	9.16	13.54	14.54	14.11	34.63	58.9	46.54	47.16	47.2
Query2Term-FS	26.46	31.9	39.36	11.87	17.04	18.29	17.67	41.59	65.67	53.5	54.16	54.19
Query2Term-PRF	23.39	27.85	34.83	9.98	14.79	15.96	16.1	37.15	61.11	45.37	46.03	46.07
Query2Doc	32.31	37.72	44.79	14.27	20.65	21.97	20.13	46.37	70.29	61.82	62.32	62.34
Query2Doc-FS	32.87	37.99	45.11	14.65	20.86	22.16	19.65	45.85	70.04	63.35	63.82	63.84
Query2Doc-PRF	27.43	33.22	39.85	11.53	18.11	19.34	18.74	44.23	66.41	52.58	53.26	53.28
CoT	29.96	36.01	43.05	13.29	19.42	20.7	19.22	45.76	69.24	57.68	58.3	58.32
CoT-PRF	28.17	33.66	40.43	12.26	18.49	19.75	18.15	43.43	66.06	52.95	53.59	53.6
Ensembled expansion methods												
Query2Term*	27.10	32.43	39.12	11.83	17.76	19.02	18.48	42.73	64.77	50.67	51.44	51.47
Query2Term-FS*	25.36	30.73	37.78	11.06	16.74	17.98	17.65	40.92	64.48	46.52	47.17	47.22
Query2Term-PRF*	24.79	30.21	36.85	10.63	16.41	17.57	17.16	40.47	63.37	47.06	47.63	47.66
Query2Doc*	31.81	37.12	44.06	14.23	21.12	22.45	20.68	46.24	69.95	57.58	58.03	58.06
Query2Doc-FS*	31.88	37.39	44.22	14.08	21.13	22.48	20.85	46.98	70.23	57.77	58.30	58.32
Query2Doc-PRF*	32.43	37.70	44.47	14.53	21.53	22.82	21.19	47.03	70.19	58.81	59.34	59.37
CoT*	30.34	35.75	42.29	13.51	20.16	21.39	19.86	45.75	67.63	56.10	56.68	56.70
CoT-PRF*	26.93	32.41	38.95	11.66	17.96	19.18	18.22	42.59	64.55	49.81	50.38	50.41
MILL	**34.33**	**39.71**	**46.39**	**15.65**	**22.89**	**24.28**	**21.32**	**48.86**	**71.13**	**64.09**	**64.53**	**64.55**

Table 15: Overall experimental results on FIQA-2018.

Metrics	NDCG			AP			Recall			MRR		
	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
No expansion	25.26	31.74	35.28	19.4	20.86	21.04	30.97	55.92	77.42	31.03	32.11	32.18
Traditional expansion methods												
Bo1	24.36	31.21	34.97	18.71	20.3	20.49	30.21	56.25	79.18	29.37	30.51	30.58
KL	24.75	31.4	35.12	18.99	20.52	20.72	30.88	56.21	78.84	29.77	30.84	30.92
RM3	22.8	29.23	33.14	16.85	18.32	18.51	30.37	54.82	78.82	26.47	27.55	27.63
AxiomaticQE	25.26	31.76	35.28	19.4	20.87	21.04	30.97	56	77.42	31.03	32.11	32.18
LLM-based expansion methods												
Query2Term	21.72	28.1	32.12	16.15	17.45	17.65	28.42	54.12	78.22	25.82	26.83	26.91
Query2Term-FS	24.83	31.95	35.78	18.9	20.49	20.68	30.5	58.45	81.84	30.57	31.61	31.68
Query2Term-PRF	21.56	27.43	31.5	16.29	17.55	17.73	27.47	50.78	76.31	25.32	26.21	26.29
Query2Doc	27	33.92	37.63	20.46	22.15	22.34	34.26	60.11	82.72	32.64	33.73	33.78
Query2Doc-FS	27.23	34.46	37.96	20.37	22.15	22.33	34.8	61.94	83.46	33.14	34.23	34.29
Query2Doc-PRF	23.51	30.26	34.09	17.91	19.39	19.57	28.99	55.33	79.14	29.18	30.19	30.27
CoT	26.69	33.78	37.28	19.8	21.48	21.65	**34.88**	62.34	83.56	32.12	33.16	33.22
CoT-PRF	27.78	34.3	38.04	21.45	23.06	23.24	34.5	59.26	82.14	33.25	34.21	34.29
Ensembled expansion methods												
Query2Term*	24.88	31.48	35.22	18.74	20.28	20.46	31.83	57.39	80.36	29.70	30.64	30.71
Query2Term-FS*	24.13	30.76	34.60	18.17	19.70	19.89	30.84	56.18	79.85	29.09	30.07	30.15
Query2Term-PRF*	23.19	29.17	33.27	17.36	18.73	18.92	30.16	52.96	78.84	27.46	28.36	28.45
Query2Doc*	26.45	33.67	37.05	19.78	21.48	21.65	34.06	62.17	82.47	31.82	32.81	32.87
Query2Doc-FS*	26.56	33.66	37.05	19.79	21.45	21.62	34.51	61.77	82.40	31.65	32.63	32.68
Query2Doc-PRF*	26.02	33.33	36.82	19.43	21.19	21.37	33.52	61.30	82.43	31.37	32.42	32.48
CoT*	27.58	34.61	38.13	20.99	22.63	22.81	34.73	61.82	83.32	33.20	34.25	34.31
CoT-PRF*	25.71	32.49	36.18	19.47	21.11	21.29	32.43	58.55	81.35	31.11	32.07	32.14
MILL	**28.42**	**35.63**	**39.23**	**21.89**	**23.61**	**23.8**	34.63	**62.46**	**84.23**	**34.94**	**35.99**	**36.05**

Table 16: Overall experimental results on SCIDOCS.

Metrics	NDCG			AP			Recall			MRR		
	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
No expansion	14.71	20.91	25.14	8.36	9.73	9.94	15.84	34.48	55.04	25.37	26.41	26.48
Traditional expansion methods												
Bo1	15.1	21.82	26.14	8.73	10.29	10.51	16.43	36.39	57.47	25.31	26.41	26.48
KL	15.1	21.81	26.15	8.75	10.31	10.54	16.37	36.24	57.38	25.43	26.54	26.61
RM3	14.56	21.49	25.91	8.41	10.05	10.28	15.79	36.24	57.88	24.46	25.63	25.7
AxiomaticQE	14.71	20.91	25.14	8.36	9.73	9.94	15.84	34.48	55.04	25.37	26.41	26.48
LLM-based expansion methods												
Query2Term	13.04	20.02	25.11	7.32	8.84	9.1	14.3	35.08	60	22.34	23.66	23.73
Query2Term-FS	14.16	21.25	26.18	8.07	9.68	9.94	15.26	36.21	60.15	24.31	25.54	25.62
Query2Term-PRF	13.1	20.13	24.97	7.49	9.12	9.37	14.84	35.56	59.25	20.54	21.84	21.91
Query2Doc	15.09	22.63	27.4	8.57	10.34	10.59	16.13	38.31	61.63	26.21	27.49	27.55
Query2Doc-FS	15.06	22.35	27.18	8.43	10.16	10.43	16.49	37.94	61.33	25.83	27.01	27.08
Query2Doc-PRF	14.3	21.5	26.16	8.21	9.96	10.21	15.7	36.78	59.5	23.84	25.03	25.11
CoT	15.54	22.77	27.5	8.9	10.58	10.84	16.65	37.96	60.9	26.81	28.07	28.13
CoT-PRF	14.71	21.66	26.23	8.44	10.1	10.34	16.05	36.5	58.72	24.77	25.91	25.98
Ensembled expansion methods												
Query2Term*	14.77	21.93	26.67	8.45	10.14	10.39	16.39	37.65	60.82	24.21	25.32	25.41
Query2Term-FS*	14.59	21.41	26.01	8.38	9.99	10.22	16.21	36.47	58.96	23.73	24.85	24.93
Query2Term-PRF*	14.18	21.18	25.88	8.13	9.78	10.03	16.05	36.70	59.61	22.40	23.55	23.63
Query2Doc*	15.25	22.90	27.49	8.73	10.59	10.84	16.54	38.98	61.32	25.84	27.08	27.15
Query2Doc-FS*	15.33	22.76	27.41	8.78	10.59	10.84	16.75	38.57	61.05	25.75	26.97	27.03
Query2Doc-PRF*	15.06	22.69	27.47	8.63	10.51	10.77	16.36	38.62	61.81	25.35	26.65	26.71
CoT*	15.46	22.84	27.62	8.84	10.60	10.87	16.97	38.57	61.69	25.85	27.12	27.18
CoT-PRF*	15.02	22.01	26.53	8.64	10.33	10.56	16.64	37.27	59.35	24.70	25.80	25.88
MILL	**16.38**	**23.73**	**28.36**	**9.5**	**11.23**	**11.48**	**17.49**	**39.28**	**61.86**	**28.1**	**29.25**	**29.31**

Table 17: Overall experimental results on ARGUANA.

Metrics	NDCG			AP			Recall			MRR		
	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
No expansion	**34.24**	39.36	39.93	22.55	23.70	23.72	**71.27**	94.17	98.58	22.56	23.71	23.73
Traditional expansion methods												
Bo1	33.07	38.92	39.42	21.81	23.14	23.16	68.78	**94.74**	**98.65**	21.82	23.15	23.17
KL	32.90	38.79	39.31	21.66	23.01	23.03	68.49	94.59	**98.65**	21.68	23.02	23.04
RM3	30.81	37.29	38.14	20.75	22.20	22.23	62.52	91.61	98.08	20.76	22.21	22.24
AxiomaticQE	34.19	39.30	39.88	22.49	23.64	23.66	**71.27**	94.10	98.58	22.50	23.65	23.67
LLM-based expansion methods												
Query2Term	33.23	38.65	39.33	21.85	23.08	23.11	69.20	93.39	98.51	21.85	23.08	23.11
Query2Term-FS	33.94	39.17	39.72	22.29	23.46	23.49	70.84	94.31	98.51	22.31	23.47	23.50
Query2Term-PRF	32.36	38.06	38.68	20.97	22.27	22.29	68.49	93.88	**98.65**	20.96	22.26	22.29
Query2Doc	33.69	39.30	39.84	22.31	23.60	23.62	69.63	94.38	98.51	22.32	23.59	23.62
Query2Doc-FS	33.90	39.43	39.92	22.40	23.68	23.70	70.20	94.67	98.36	22.41	23.69	23.72
Query2Doc-PRF	32.45	38.32	38.85	21.11	22.46	22.48	68.42	94.45	98.58	21.15	22.48	22.50
CoT	34.21	39.48	40.00	22.57	23.77	23.79	71.05	94.52	98.44	22.60	23.78	23.80
CoT-PRF	34.12	39.51	40.01	22.51	23.74	23.76	70.77	**94.74**	98.58	22.50	23.74	23.76
Ensembled expansion methods												
Query2Term*	33.42	38.87	39.45	21.94	23.18	23.20	69.77	94.03	98.44	21.96	23.19	23.22
Query2Term-FS*	33.51	39.00	39.58	22.07	23.32	23.35	69.70	94.10	98.44	22.09	23.34	23.36
Query2Term-PRF*	33.24	38.63	39.20	21.66	22.88	22.90	69.99	94.10	98.44	21.68	22.89	22.91
Query2Doc*	33.24	38.96	39.49	21.87	23.18	23.20	69.27	94.59	**98.65**	21.88	23.19	23.22
Query2Doc-FS*	33.20	38.80	39.34	21.71	23.00	23.02	69.70	94.52	**98.65**	21.71	23.01	23.03
Query2Doc-PRF*	31.73	37.93	38.49	20.67	22.10	22.12	66.86	94.24	98.58	20.69	22.11	22.13
CoT*	33.61	39.08	39.59	22.04	23.29	23.31	70.34	94.67	98.51	22.05	23.31	23.33
CoT-PRF*	32.86	38.62	39.15	21.46	22.79	22.81	69.06	94.52	98.58	21.46	22.80	22.82
MILL	34.10	**39.56**	**40.11**	**22.65**	**23.91**	**23.94**	70.20	94.31	98.44	**22.67**	**23.92**	**23.95**

Table 18: Overall experimental results on CLIMATE-FEVER.

Metrics	NDCG			AP			Recall			MRR		
	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
No expansion	12.52	17.96	21.73	8.28	9.32	9.48	16.20	35.80	57.63	17.07	18.26	18.35
Traditional expansion methods												
Bo1	13.48	19.43	23.11	8.86	10.03	10.20	17.65	39.10	60.22	18.14	19.32	19.40
KL	13.48	19.42	23.07	8.90	10.06	10.23	17.50	38.92	60.01	18.25	19.43	19.51
RM3	11.24	17.17	20.71	6.80	7.99	8.15	16.39	37.64	58.18	13.99	15.21	15.28
AxiomaticQE	12.54	17.98	21.75	8.28	9.33	9.49	16.23	35.82	57.66	17.09	18.28	18.37
LLM-based expansion methods												
Query2Term	16.90	23.50	27.23	11.24	12.68	12.86	21.88	45.16	66.59	22.47	23.64	23.71
Query2Term-FS	14.11	20.53	24.32	9.27	10.65	10.83	18.47	41.00	62.87	18.78	19.98	20.06
Query2Term-PRF	13.90	20.48	23.98	8.49	9.93	10.09	20.64	43.62	63.79	16.56	17.86	17.92
Query2Doc	**21.62**	**28.64**	**32.39**	**14.64**	**16.30**	**16.50**	27.54	51.68	72.98	**28.41**	**29.60**	**29.65**
Query2Doc-FS	21.44	28.24	32.05	14.46	15.99	16.19	27.50	51.43	73.01	28.12	29.21	29.26
Query2Doc-PRF	16.83	23.33	26.90	10.66	12.11	12.29	23.92	46.82	67.15	20.57	21.71	21.77
CoT	19.62	26.61	30.25	13.45	14.99	15.18	24.66	48.97	69.86	26.04	27.31	27.36
CoT-PRF	15.71	22.24	25.78	10.26	11.70	11.87	21.19	43.97	64.26	20.19	21.44	21.50
Ensembled expansion methods												
Query2Term*	16.21	23.15	26.86	10.04	11.60	11.79	23.49	47.84	69.09	19.59	20.80	20.86
Query2Term-FS*	14.33	21.16	24.97	8.80	10.28	10.47	20.89	45.00	66.94	17.38	18.66	18.73
Query2Term-PRF*	13.90	20.48	23.98	8.49	9.93	10.09	20.64	43.62	63.79	16.56	17.86	17.92
Query2Doc*	19.64	27.06	30.67	12.57	14.36	14.55	**27.55**	**53.01**	**73.51**	24.07	25.32	25.36
Query2Doc-FS*	18.75	26.14	29.78	11.86	13.62	13.81	26.55	52.15	72.82	22.88	24.10	24.15
Query2Doc-PRF*	16.30	22.69	26.25	10.20	11.61	11.78	23.65	46.11	66.42	19.61	20.77	20.84
CoT*	18.94	26.08	29.74	12.27	13.93	14.13	25.95	50.85	71.64	23.72	24.90	24.95
CoT-PRF*	15.65	22.18	25.48	9.87	11.30	11.46	22.24	45.29	64.30	19.31	20.47	20.53
MILL	20.28	27.06	30.66	13.60	15.14	15.32	26.94	50.65	71.09	25.88	27.02	27.08

Table 19: More results of ablation experiments.

Methods	NDCG			AP			Recall			MRR		
	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
TREC-DL-2020												
w/o PRF	60.13	59.88	70.65	**19.63**	41.57	48.10	**22.47**	58.34	**85.97**	88.97	89.09	89.10
w/o MV	59.46	59.42	70.28	18.01	40.18	46.73	21.45	58.66	85.11	90.70	90.75	90.75
w/o QQD	60.26	59.89	69.46	17.96	41.56	47.39	21.19	58.55	83.98	87.65	87.69	87.69
MILL	**61.79**	**61.15**	**71.23**	19.05	**41.76**	**48.17**	21.61	**59.40**	85.27	**92.61**	**92.71**	**92.72**
TREC-COVID												
w/o PRF	**76.61**	58.53	51.17	2.01	11.43	27.35	2.20	14.72	49.09	**92.40**	**92.40**	**92.40**
w/o MV	74.34	59.40	51.73	1.87	11.79	28.44	2.14	15.15	50.00	87.40	87.40	87.40
w/o QQD	71.87	57.57	50.84	1.85	11.27	27.30	2.07	14.55	49.16	88.90	89.08	89.08
MILL	75.30	**60.24**	**52.53**	**2.03**	**12.22**	**29.30**	**2.22**	**15.40**	**50.55**	91.17	91.17	91.17
SCIFACT												
w/o PRF	69.75	72.35	73.01	64.92	65.55	65.58	83.27	94.43	**99.67**	66.01	66.52	66.54
w/o MV	69.53	71.70	72.43	64.26	64.76	64.79	84.38	94.03	**99.67**	65.29	65.68	65.71
w/o QQD	67.98	70.47	71.13	62.32	62.94	62.96	83.89	**94.53**	**99.67**	63.46	63.96	63.98
MILL	**71.37**	**73.47**	**74.14**	**66.34**	**66.85**	**66.88**	**85.24**	94.50	**99.67**	**67.69**	**68.07**	**68.09**

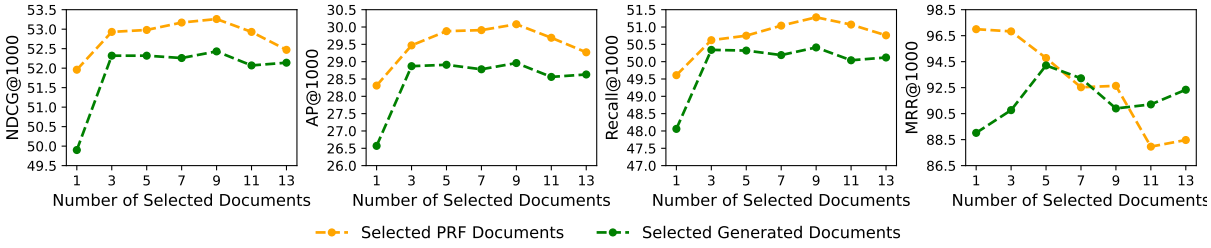

Figure 4: Hyperparameter analysis on the number of document selections on TREC-COVID. The x-axis denotes the number of documents selected, and the y-axis represents the metrics values (NDCG@1000, AP@1000, Recall@1000, and MRR@1000).

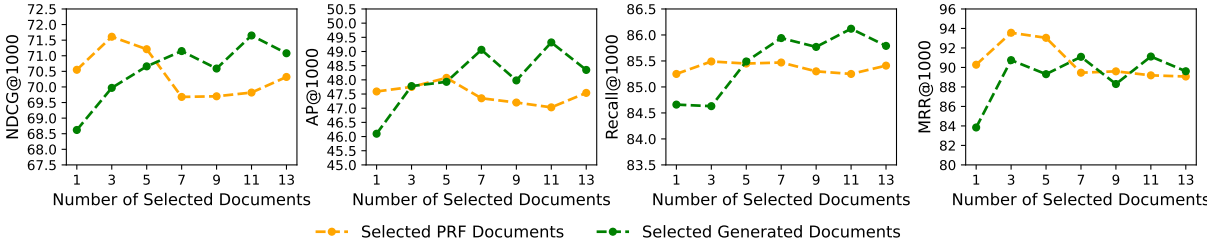

Figure 5: Hyperparameter analysis on the number of document selections on TREC-DL-2020. The x-axis denotes the number of documents selected, and the y-axis represents the metrics values (NDCG@1000, AP@1000, Recall@1000, and MRR@1000).

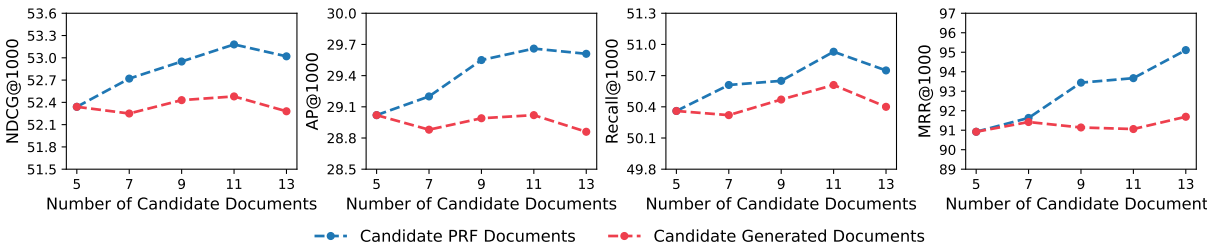

Figure 6: Hyperparameter analysis on the number of document candidates on TREC-COVID. The x-axis denotes the number of document candidates, and the y-axis represents the metrics values (NDCG@1000, AP@1000, Recall@1000, and MRR@1000).

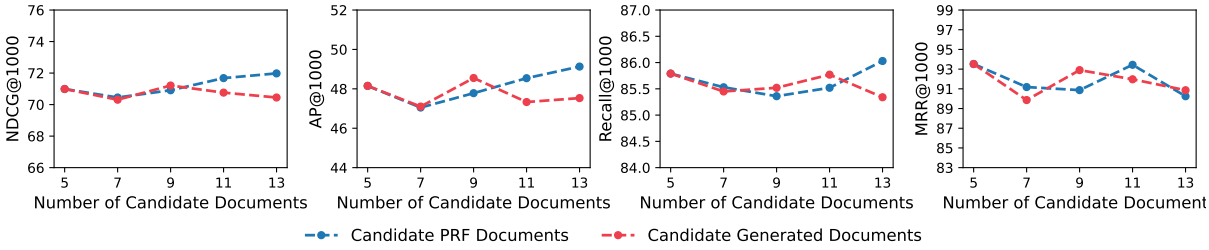

Figure 7: Hyperparameter analysis on the number of document candidates on TREC-DL-2020. The x-axis denotes the number of document candidates, and the y-axis represents the metrics values (NDCG@1000, AP@1000, Recall@1000, and MRR@1000).