Corpus-Steered Query Expansion with Large Language Models

Anonymous ACL submission

Abstract

Recent studies demonstrate that query expansion, generated by large language models (LLMs), considerably enhances information retrieval systems by generating hypothetical documents that answer the queries as expansions. However, challenges arise from misalignments between the expansions and the retrieval corpus, resulting in issues like hallucinations and outdated information due to the limited intrinsic knowledge of LLMs. Inspired by Pseudo Relevance Feedback (PRF), we introduce Corpus-Steered Query Expansion (CSQE) to promote the incorporation of authentic knowledge embedded within the corpus. CSQE utilizes the relevance assessing capability of LLMs to systematically identify pivotal sentences in the initially-retrieved documents. These corpus-originated texts are subsequently used to expand the query together with LLM-knowledge empowered expansions, bolstering the relevance between the query and the target documents. Extensive experiments reveal that CSQE exhibits remarkable performance without necessitating any training.

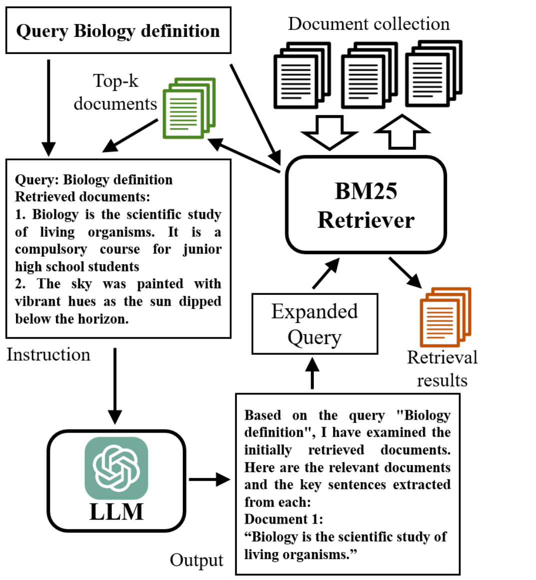

Figure 1: Overview of CSQE. Given a query *Biology definition* and the top-2 retrieved documents, CSQE utilizes an LLM to identify relevant document 1 and extract the key sentences from the corpus that contribute to the relevance. The query is then expanded by both these corpus-originated texts and LLM-knowledge empowered expansions (i.e., hypothetical documents that answer the query) to obtain the final results.

1 Introduction

Query expansions enhance the effectiveness of information retrieval systems by incorporating additional texts into the original query. Traditional methods often employ pseudo-relevance feedback (Amati and Van Rijsbergen, 2002; Robertson, 1990) or leverage external lexical knowledge sources (Bhogal et al., 2007; Qiu and Frei, 1993). Recent studies (Gao et al., 2022; Wang et al., 2023; Jagerman et al., 2023; Mackie et al., 2023) show query expansions generated by LLMs are able to significantly boost retrieval effectiveness, especially in zero-shot scenarios. For instance, Gao et al. (2022) demonstrates the effectiveness of utilizing LLMs to generate hypothetical documents answering the original query as additional terms to augment the query. Mackie et al. (2023) show the efficacy of applying pseudo-relevance feedback upon the LLM-generated answers for expansion. Despite variations in prompts or expansion methods, a common foundational element across these approaches is the reliance on the intrinsic knowledge of LLMs.

Despite their effectiveness, generations that rely on the intrinsic parametric knowledge within LLMs encounter various issues. These include hallucination (Zhang et al., 2023), inability to update (Kasai et al., 2022), and a deficiency in long-tail knowledge (Kandpal et al., 2023). Such generations may introduce irrelevant or misleading terms, adversely affecting the retrieval performance (Weller et al., 2023).

To this end, we propose Corpus-Steered Query

Expansion (CSQE). Unlike previous methods that rely on the intrinsic parametric knowledge of LLMs, CSQE exclusively leverages the remarkable relevance assessing capability of LLMs (Faggioli et al., 2023; Thomas et al., 2023). As illustrated in Figure 1, given a query and its initially retrieved documents, CSQE utilizes LLMs to first identify relevant documents to the query and then extract pivotal sentences that contribute to their relevance. These corpus-originated texts are then combined together with LLM-knowledge empowered expansions to expand the original query. By incorporating query expansions that strictly originate from the corpus, CSQE balances out the limitations commonly found in LLM-knowledge empowered expansions.

To sum up, our contributions are 3-fold:

1) We propose CSQE, which exclusively exploits the relevance assessing capability of LLMs to overcome the hinder posed by LLM-knowledge empowered expansions.

2) Experimental results reveal that CSQE combined with a simple BM25 model, without necessitating any training, outperform not only LLM-knowledge empowered expansion methods but also the SOTA supervised Contriever$^{\text{FT}}$ model across two high-resource web search datasets and six low-resource BEIR datasets.

3) Further analysis demonstrates the advantages of BM25 over dense retrieval with query expansion from LLMs, as well as query expansion over large-scale fine-tuning upon Contriever.

2 Method

In this section, we first describe how we implement a Knowledge Empowered Query Expansion baseline based on LLMs (KEQE), then present the details of CSQE to enhance BM25.

KEQE Inspired by recent works that directly generate hypothetical documents to answer the query via LLMs for boosting retrieval (Gao et al., 2022; Wang et al., 2023; Jagerman et al., 2023; Mackie et al., 2023), we implement a KEQE baseline in a similar pattern for fair comparison. Given a query q, we use LLMs to generate the hypothetical answer a via a task-agnostic prompt shown in Table 1. The concatenation of q and a is then used as the expanded query to BM25 to retrieve the final results.

It is worth noting that these hypothetical documents are inevitably susceptible to issues like hallucination, due to the limitation of LLMs' inherent knowledge, and then adversely affect the retrieval performance. To mitigate such problems, we propose CSQE to incorporate corpus-originated expansions with authentic knowledge embedded in the corpus.

CSQE Given a query q and the document collection $\mathcal{D}$, we first retrieve top-k documents $\{d_1, d_2, \ldots, d_k\}$ using BM25. Then we elicit large language models to directly generate the pseudo-relevance feedback via one-shot prompting as shown in Table 2, where the current retrieved documents are integrated. The learning context in the prompt is constructed from the TREC DL19 dataset for constraining the structure of generated texts. Noting that such a prompt remains unchanged for all tasks, we can therefore consider our method with minimal relevance supervision and being a zero-shot approach for all datasets excluding DL19.

Based on the above prompting, the generation of LLMs will contain (1) the indices of documents that are identified as relevant to the query and (2) the key sentences that contribute to their relevance, denoted as $S = \{s_1, s_2, \ldots, s_n\}$. Then we expand the query by concatenating q, all sentences in S, and generations from KEQE as a new query for BM25 retrieval, where the results in this turn are regarded as the final retrieved documents. Since these key sentences are identical to the existing texts in the corpus, [1] they are susceptible to issues such as hallucinations and shortness of long-tail knowledge and can balance out the limitations of KEQE expansions.

To increase diversity, we sample N generations from the LLM for expansion. For KEQE, $N = 5$. As CSQE involves both KEQE and corpus-originated expansions, we sample $N = 2$ for both KEQE and corpus-originated expansions, in total only 4 generations for fair comparison.

KEQE Prompt
Please write a passage to answer the question Question: $\{q\}$ Passage:

Table 1: Prompt of KEQE. $\{\cdot\}$ denotes the placeholder for the corresponding text.

[1] In our preliminary study, we found 830 out of 1000 key sentences extracted by GPT-3.5-Turbo are identical to sentences in the initially-retrieved documents.

CSQE Prompt

Query: "how are some sharks warm blooded"
Retrieved documents:
1. Most sharks are cold-blooded. Some, like the Mako and the Great white shark, are partially warmblooded (they are endotherms)…
2. Are sharks cold-blooded or warm-blooded? Sharks have a reputation as cold-blooded and despite how negative that term is…
3. Great white sharks are some of the only warm blooded sharks. This allows them to swim in colder waters in addition to warm, tropical waters…
You will begin by examining the initially retrieved documents and identifying the ones that are relevant, even partially, to the query. Once the relevant documents are identified, you will extract the key sentences from each document that contribute to their relevance.

Based on the query "how are some sharks warm blooded", I have examined the initially retrieved documents. Here are the relevant documents and the key sentences extracted from each:
Document 1:
"Most sharks are cold-blooded. Some, like the Mako and the Great white shark, are partially warm-blooded (they are endotherms)."
Document 3:
"Great white sharks are some of the only warm-blooded sharks."

Query: "$\{q\}$"
Retrieved documents:
1. $\{d_1\}$
2. $\{d_2\}$
…
$\{k\}$. $\{d_k\}$
You will begin by examining the initially retrieved documents and identifying the ones that are relevant, even partially, to the query. Once the relevant documents are identified, you will extract the key sentences from each document that contribute to their relevance.

Table 2: Prompt of CSQE. $\{\cdot\}$ denotes the placeholder for the corresponding text. Refer to Appendix A.1 for the complete prompt.

3 Experiments

3.1 Setup

Datasets Following Gao et al. (2022), the datasets for evaluation are (1) Two web search datasets: TREC DL19 (Craswell et al., 2020) and TREC DL20 (Craswell et al., 2021), which are based on the high-resource MS-MARCO dataset (Bajaj et al., 2018); and (2) Six low-resource retrieval datasets from the BEIR dataset (Thakur et al., 2021) covering a variety of domains (e.g., medical and finance) and query types (e.g., news headlines and arguments).

Baselines The baselines we consider are within two categories: PRF methods and query expansion methods using LLMs. The PRF method we include is **BM25+RM3** (Jaleel et al., 2004). The query expansion methods with LLMs include: (1) **Contriever+HyDE**, a KEQE method that employs hypothetical documents generated by LLMs to enhance unsupervised Contriever (Izacard et al., 2022) model; (2) **BM25+GPR** (Mackie et al., 2023), a query expansion method that applies PRF upon LLM-knowledge empowered hypothetical texts. GPR is a strong baseline that outperforms multiple SOTA PRF methods; and (3) **BM25+KEQE**.

Moreover, we also include three supervised dense retrievers that are trained with over 500k human-labeled data of MS-MARCO for reference: (1) **DPR**; (2) **ANCE**, which involves sophisticated negative mining; and (3) **Contriever$^{\mathbf{FT}}$**, which is the fine-tuned version of Contriever.

Implementation We utilize GPT-3.5-Turbo[2] as our serving LLM for the trade-off between performance and cost. We sample from the LLM with a temperature of 1.0. The BM25 retrieval and RM3 query expansion are performed using Prserini (Lin et al., 2021) toolkit with default hyper-parameters. CSQE utilize the top-10 retrieved documents, with each truncated to at most 128 tokens. To increase diversity, for each API call, we sample N generations. For KEQE, $N = 5$. As CSQE involves both KEQE and corpus-originated expansions, we sample N = 2 for both KEQE and corpus-originated expansions, in total only 4 generations for fair comparison. The expanded query of each generation is further concatenated together to form the final query.

3.2 Web Search Results

Table 3 shows the retrieval results on TREC DL19 and DL20. CSQE is able to bring substantially larger improvement over BM25 compared to the strong PRF baseline RM3. Despite utilizing fewer LLM generations for expansion, CSQE surpasses KEQE on 5/6 metrics, showing the effectiveness of our corpus-steered approach. Moreover, CSQE consistently outperforms GPR on 5/6 metrics, which employs PRF on KEQE expansions, emphasizing the necessity of corpus-steered expansions. Without any training, CSQE with simple BM25 is able to beat the SOTA Contriever$^{\text{FT}}$ model across all metrics by a substantial margin.

3.3 Low-Resource Retrieval Results

The results on 6 low-resource BEIR datasets are shown in Table 4. Applying RM3 leads to performance drops on 5/6 datasets, while CSQE is robust to domain shifts and is able to consistently improve BM25 on all datasets. Although KEQE can achieve similar results as Contriever$^{\text{FT}}$, CSQE is able to outperform both KEQE and Contriever$^{\text{FT}}$ by a large margin, demonstrating the strong generalizability of CSQE.

4 Analysis

4.1 CSQE on Dense Retrieval

To test the versatility of CSQE, we apply CSQE on the unsupervised Contriever in Table 5. Fol-

[2]In our preliminary study, updating HyDE's LLM from Text-Davinci-003 to GPT-3.5-Turbo cannot improve results.

	DL19			DL20		
	map	ndcg@10	recall@1k	map	ndcg@10	recall@1k
w/o relevance judgement						
BM25	30.1	50.6	75.0	28.6	48.0	78.6
BM25+RM3	34.2	52.2	81.4	30.1	49.0	82.4
Contriever+HyDE	41.8	61.3	88.0	38.2	57.9	84.4
BM25+GRF	44.1	62.0	79.7	**48.6**	60.7	87.9
BM25+KEQE	45.0	65.9	**88.8**	42.8	60.5	88.3
BM25+CSQE	**47.2**	**67.3**	88.5	46.5	**66.2**	**89.1**
reference. w/ relevance judgement						
DPR	36.5	62.2	76.9	41.8	65.3	81.4
ANCE	37.1	64.5	75.5	40.8	64.6	77.6
Contriever^{FT}	41.7	62.1	83.6	43.6	63.2	85.8

Table 3: Results on TREC DL19 and DL20 datasets. In-domain supervised models DPR, ANCE and Contriever^{FT} are trained on the MS-MARCO dataset and listed for reference. **Bold** indicates the best result across all models.

	Scifact	Arguana	Trec-Covid	FiQA	DBPedia	TREC-NEWS	Avg.
			nDCG@10				
w/o relevance judgement							
BM25	67.9	39.7	59.5	23.6	31.8	39.5	43.7
BM25+RM3	64.6	38.0	59.3	19.2	30.8	42.6	42.4
Contriever+HyDE	69.1	46.6	59.3	27.3	36.8	44.0	47.2
BM25+KEQE	70.5	40.7	66.6	22	38.8	48.3	47.8
BM25+CSQE	**69.6**	40.3	**74.2**	25	40.3	**48.7**	**49.7**
reference. w/ relevance judgement							
DPR	31.8	17.5	33.2	29.5	26.3	16.1	25.7
ANCE	50.7	41.5	65.4	30.0	28.1	38.2	42.3
Contriever^{FT}	67.7	**44.6**	59.6	**32.9**	**41.3**	42.8	48.2

Table 4: Results on low-resource retrieval datasets. **Bold** indicates the best result across all models.

lowing Gao et al. (2022), we encode each query expansion separately into dense embeddings and average their embeddings with the original query embedding as the final embedding. Similar to the impact of CSQE on BM25, CSQE is able to improve Contriever significantly. Interestingly, it is worth noting that in all cases, Contriever performs worse than BM25. Surprisingly, query expansion (Contriever+CSQE) is proven to be more effective than fine-tuning the model using 500K human-labeled data (Contriever^{FT}).

Model	map	ndcg@10	recall@1k
Contriever	24.0	44.5	74.6
+KEQE	41.7	62.2	87.4
+CSQE	44.0	65.6	88.6
BM25	30.1	50.6	75.0
+KEQE	45.0	65.9	88.8
+CSQE	47.6	68.6	89.0
Contriever^{FT}	41.7	62.1	83.6

Table 5: Results of CSQE on Contriever on DL19.

4.2 Corpus-Originated expansion on Different LLMs

We apply different LLMs for corpus-originated expansion in Table 6. Consistent with findings in Sun et al. (2023), we find LLM-based expansion is able to bring consistent improvements and more powerful models are able to bring bigger improvement. Considering the trade-off between performance and cost, we choose GPT-3.5-Turbo as our serving LLM.

Model	map	ndcg@10	recall@1k
BM25	30.1	50.6	75.0
w/ LLAMA-2-7B-Chat	35.8	54.3	82.5
w/ Text-Davinci-003	37.9	55.8	80.8
w/ GPT-3.5-Turbo	41.9	63.9	82.9
w/ GPT-4	42.9	67.0	84.8

Table 6: Corpus-originated expansion with different LLMs on DL19.

5 Conclusion

In this paper, we propose CSQE, which utilizes the relevance assessing ability of LLMs to balance out limitations associated with the intrinsic knowledge of LLMs. Experimental evaluation demonstrates CSQE's superiority over the LLM-knowledge empowered expansion methods and SOTA supervised Contriever^{FT} model across various datasets.

Limitations

We acknowledge two limitations in our work: computational overhead and reliance on closed-source models. The utilization of OPENAI LLMs necessitates API calls, resulting in increased processing time and latency. However, in retrieval tasks where latency is less crucial, such as legal case retrieval, our method may offer benefits. Moreover, our approach does not necessitate training, making it more accessible to researchers and practitioners without extensive GPU resources. Additionally, the unavailability of the LLMs' source models and training data restricts our ability to conduct thorough analysis. There may exist social biases (Zhao et al., 2017) in LLM generations and thus have the risk of offending people from under-represented groups.

We utilize ChatGPT to correct the grammar in our paper and ensure that none of the text was directly generated by ChatGPT.

References

Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset.

J. Bhogal, A. Macfarlane, and P. Smith. 2007. A review of ontology based query expansion. *Information Processing Management*, 43(4):866–886.

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

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the trec 2019 deep learning track.

Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '23, page 39–50, New York, NY, USA. Association for Computing Machinery.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

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

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*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2022. RealTime QA: What's the answer right now?

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2356–2362, New York, NY, USA. Association for Computing Machinery.

Iain Mackie, Shubham Chatterjee, and Jeffrey Dalton. 2023. Generative relevance feedback with large language models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2026–2031, New York, NY, USA. Association for Computing Machinery.

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 Robertson. 1990. On term selection for query expansion. *Journal of Documentation*, 46:359–364.

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.

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)*.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large language models can accurately predict searcher preferences.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *EMNLP*.

Orion Weller, Kyle Lo, David Wadden, Dawn J Lawrie, Benjamin Van Durme, Arman Cohan, and Luca Soldaini. 2023. When do generative query and document expansions fail? a comprehensive study across methods, retrievers, and datasets. *ArXiv*, abs/2309.08541.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

A Appendix

A.1 Instruction of PRF-LLM

Query: "how are some sharks warm blooded"
Retrieved documents:
1. Most sharks are cold-blooded. Some, like the Mako and the Great white shark, are partially warmblooded (they are endotherms). Cold blooded although if you've ever seen a Great White Shark hunt sea lions you'd be thinking they would have to be hotblooded. Actually the Salmon Shark is a warm blooded shark.
2. Are sharks cold-blooded or warm-blooded? Sharks have a reputation as cold-blooded and despite how negative that term is, it is not entirely inaccurate. Sharks are by no means evil, vicious killers like that quote suggests. Nonetheless, sharks are, for the most part anyways, efficient ectothermic predators. Endo vs Ecto.
3. Great white sharks are some of the only warm blooded sharks. This allows them to swim in colder waters in addition to warm, tropical waters. Great White sharks can be found asâ¦ north as Alaska and as south as the southern tip of South America. They exist worldwide, everywhere in-between. 5 people found this useful.
4. Sharks' blood gives them turbo speed. Several species of shark and tuna have something special going on inside their bodies. For a long time, scientists have known that some fish species appear warm-blooded. Salmon sharks can elevate their body temperatures by up to 20 degrees compared to the surrounding water, for example.

You will begin by examining the initially retrieved documents and identifying the ones that are relevant, even partially, to the query. Once the relevant documents are identified, you will extract the key sentences from each document that contribute to their relevance.

Based on the query "how are some sharks warm blooded", I have examined the initially retrieved documents. Here are the relevant documents and the key sentences extracted from each:
Document 1:
"Most sharks are cold-blooded. Some, like the Mako and the Great white shark, are partially warm-blooded (they are endotherms)."
"Actually, the Salmon Shark is a warm-blooded shark."
Document 3:
"Great white sharks are some of the only warm-blooded sharks."
"This allows them to swim in colder waters in addition to warm, tropical waters."
Document 4:
"Salmon sharks can elevate their body temperatures by up to 20 degrees compared to the surrounding water, for example."

Query: "$\{q\}$"
Retrieved documents:
1. $\{d_1\}$
2. $\{d_2\}$
$\ldots$
$\{k\}$. $\{d_k\}$
You will begin by examining the initially retrieved documents and identifying the ones that are relevant, even partially, to the query. Once the relevant documents are identified, you will extract the key sentences from each document that contribute to their relevance.

A.2 Dataset Statistics

Details about the retrieval datasets are shown in Table 7.

Dataset	#Test	#Corpus
DL19	43	8,841,823
DL20	50	8,841,823
Scifact	300	5183
Arguana	1406	8674
Trec-Covid	50	171,332
FiQA	648	57,638
DBPedia	400	4,635,922
TREC-NEWS	57	594,977

Table 7: **Dataset Statistics**