Domain-Specific Retrieval for Retrieval-Augmented Generation: A Case Study on Pertussis Research (Student Abstract)

Hiroki Takabatake[1], Niken Prasasti Martono[1], Asaomi Kuwae[2], Toshihiko Iuchi[1], Hayato Ohwada[1]

[1]Dept. of Industrial and Systems Engineering, Tokyo University of Science, Noda, Chiba, Japan
[2]Graduate School of Infection Control Sciences, Kitasato University, Minato-ku, Tokyo, Japan
7424527@ed.tus.ac.jp, niken@rs.tus.ac.jp, kuwae@lisci.kitasato-u.ac.jp, toshihiko_iuchi@rs.tus.ac.jp, ohwada@rs.tus.ac.jp

Abstract

Integrating knowledge from scientific literature is essential in biomedical research. However, the rapid growth of scientific literature makes staying up to date increasingly challenging. Retrieval-Augmented Generation (RAG) offers a promising framework, but its effectiveness in specialized biomedical domains remains unclear. In this work, we propose a two-stage retrieval pipeline for RAG, with a focus on *Bordetella pertussis* as a case study. Our method first applies hard filtering with synonym expansion to eliminate irrelevant passages, and then performs hybrid search, followed by reranking. We evaluate our approach using a dataset of 58 pertussis-related queries with automatic relevance judgments from multiple large language models(LLMs). Experimental results show that our pipeline improves MAP@10 by 13.4-20.4 points compared with existing methods and achieves the highest MRR@10. Furthermore, consistent improvements across different LLMs highlight the effectiveness of our approach.

Introduction

Integrating knowledge from scientific literature is essential for discovering new research directions, refining methodologies, and supporting evidence-based decision making. In medicine, it plays a particularly important role in drug discovery, clinical therapies, and pathological research(Luo et al. 2022). For example, in pertussis research, the development of therapeutic strategies requires identifying proteins that interact with those produced by *Bordetella pertussis*. However, due to the high experimental cost, it is necessary to narrow down candidates with reference to the scientific literature. Nevertheless, this task is further complicated by the fact that relevant scientific literature are dispersed across multiple fields, and the volume of research continues to grow rapidly, staying up to date has become challenging. Given these challenges, computational methods that can efficiently leverage scientific literature are crucial. Recent advances in Retrieval-Augmented Generation (RAG)(Lewis et al. 2020; Ram et al. 2023) offer a promising direction. RAG enables language models to retrieve and integrate up-to-date, domain-specific knowledge from external sources at inference. The overall performance of RAG depends largely on the search results retrieved from external knowledge sources. Consequently, many recent studies have focused on innovations in retrieval methods. However it remains unclear whether these approaches are effective in specialized, terminology-heavy medical texts. In this work, we contribute a domain-specific retrieval system tailored for scientific literature, with *B. pertussis* as a case study. Our goal is to demonstrate how improved retrieval enhances knowledge integration in terminology-heavy biomedical contexts.

Methods

We propose a two-stage retrieval pipeline, which consists of hard filtering and hybrid retrieval with a reranker.

Hard filtering. The initial stage of the retrieval pipeline removes irrelevant passages using keywords extracted from the query and their synonyms. Specifically, keywords are extracted from the query using LLMs, and the terms are limited through rule-based criteria. Synonyms for these terms are then generated by LLMs. By incorporating synonyms into hard filtering, this process helps mitigate the risk of missing relevant passages due to variations in keyword expression.

Hybrid search with a reranker. We construct the hybrid search by combining vector search and keyword search. Following HyDE(Gao et al. 2023), we generate three hypothetical documents for vector search and extract keywords from them for keyword search. The results of the two searches are then integrated using Reciprocal Rank Fusion (RRF), where k is fixed to 60. To evaluate relevance scores accurately, we apply a reranker using GPT-4.1, which selects the top-N passages as the final retrieval results.

Datastore. The retrieval sources in our pipeline consist of 76,496 abstracts collected from PubMed with regard to the medical bacteriology area, especially focusing on *Bordetella* including causative agency of whooping cough. In this work, we use two datastores: one for keyword search, which stores documents as plain text, and the other for vector search where documents are chunked and converted into embeddings. Specifically, each passage consists of five sentences, with an overlap of two sentences. We then generate embeddings for each passage using the OpenScholar_Retriever(Asai et al. 2024).

Evaluation. Following PubMedQA(Jin et al. 2019), we constructed a test dataset from PubMed under the same datastore conditions, consisting of 58 queries with Gold-standard labels—Yes, No, and Maybe—annotated by do-

main experts. To evaluate passage relevance, we introduce an automatic evaluation approach using GPT-4o, Gemini 2.5 Pro, and Claude Sonnet 4. The evaluation prompts are based on UMBRELA(Upadhyay et al. 2024) and are designed to output a 0-3 relevance score. Finally, the scores are binarized and mapped to labels of relevant and non-relevant. Furthermore, we report MAP@10 as the retrieval performance metric. However, due to LLM-based evaluation is too costly for the entire corpus, we redefine MAP@k by fixing the denominator to k. Formally, redefined MAP@k can be represented as:

$$MAP@k = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{\sum_{i=1}^{k} P_q(i) \cdot rel_q(i)}{k} \quad (1)$$

This adjustment results in underestimation of the absolute MAP@10 scores, but it remains suitable for relative performance comparison between different models.

Results and Discussion

Comparison of retrieval performance. The comparison results with existing methods such as HyDE, RAG-Fusion(Raudaschl 2023), and a hybrid search are shown in Table 1 (top). In terms of MAP@10, our proposed method achieved an improvement of 13.4 to 20.4 points. These results confirm that our proposed two-stage retrieval pipeline can lead to notable performance improvements. Regarding MRR@10, our approach achieved the highest score compared with existing methods. However, hard filtering is limited, as it risks discarding relevant passages in cases where only a small number of such passages exist for a query. As shown in Table 1 (bottom), MAP@10 changed little while MRR@10 decreased, but excluding queries with zero relevant passages revealed greater gains in MAP@10 and improved MRR@10 over the no-filtering setting. These findings indicate that hard filtering functions effectively for the majority of queries, and appropriate tuning of the filtering conditions could lead to further performance improvements.

Performance on downstream tasks. The comparison results with each LLM are shown in Table 2. The results clearly demonstrate that our approach consistently achieves the highest scores across all LLMs. These results suggest that the improvements achieved by our method extend beyond the retrieval stage, having beneficial effects on downstream tasks as well. Since the observed performance differences across models were relatively small, it is indicated that our proposed approach does not rely on a specific LLM and can be applied effectively across a wide range of models.

	MAP@10	MRR@10
Performance across all queries		
HyDE	18.7	67.1
RAG-Fusion	21.5	77.7
Hybrid Search	25.7	89.0
Ours w/o hard filtering	38.9	**96.3**
Ours	**39.1**	92.2
Excluding zero-result queries		
Ours w/o hard filtering	40.1	98.8
Ours	**42.0**	**99.1**

Table 1: Comparison of Retrieval Performance

	Acc.	Macro-F1	Weighted-F1
GPT-4.1			
HyDE	48.3	43.4	52.3
RAG-Fusion	55.1	48.1	58.6
Hybrid Search	55.2	49.9	58.1
Ours	**60.3**	**54.9**	**63.9**
Claude Sonnet 4			
HyDE	48.3	40.2	50.3
RAG-Fusion	56.9	36.4	57.4
Hybrid Search	56.9	34.8	57.2
Ours	**62.1**	**48.0**	**62.0**
Gemini 2.5 Pro			
HyDE	48.3	41.9	51.1
RAG-Fusion	56.9	48.1	58.9
Hybrid Search	55.2	48.4	58.6
Ours	**61.0**	**57.0**	**65.0**

Table 2: Performance on Downstream Tasks for each LLM (Acc. denotes Accuracy).

Conclusion

In this work, we proposed a retrieval system designed for retrieval-augmented generation to support knowledge integration from scientific literature. Our approach introduced a two-stage retrieval pipeline consisting of hard filtering and hybrid search with a reranker.

The experimental results showed that our method significantly improved retrieval performance compared to existing methods, achieving the highest scores in both MAP@10 and MRR@10. While hard filtering may exclude relevant passages, it was shown to be effective in the majority of cases. Furthermore, our approach consistently outperformed all baseline methods across multiple LLMs in downstream tasks. This suggests that the improvements in retrieval stage contribute positively to downstream tasks. In addition, the relatively small performance differences between LLMs indicate that our method is not dependent on any specific language model and can be effectively applied across a wide range of language models.

Finally, we utilized PubMedQA by using paper titles as user queries in our evaluation. Nevertheless, it is still unclear whether this setting accurately reflects expert questions in real-world scenarios. Furthermore, several limitations exist in this work. First, while hard filtering enhances overall effectiveness, it may also inadvertently exclude relevant passages, and its impact requires more detailed investigation. Second, although our approach has the potential to be applied beyond the pertussis domain to other medical fields, the dataset used in this work is limited in scale, and therefore the generalization performance needs to be carefully considered.

References

Asai, A.; He, J.; Shao, R.; Shi, W.; Singh, A.; Chang, J. C.; Lo, K.; Soldaini, L.; Feldman, S.; D'arcy, M.; et al. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*.

Gao, L.; Ma, X.; Lin, J.; and Callan, J. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1762–1777.

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W. W.; and Lu, X. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; and Liu, T.-Y. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Ram, O.; Levine, Y.; Dalmedigos, I.; Muhlgay, D.; Shashua, A.; Leyton-Brown, K.; and Shoham, Y. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11: 1316–1331.

Raudaschl, A. H. 2023. Forget RAG, the Future is RAG-Fusion. *Towards Data Science*.

Upadhyay, S.; Pradeep, R.; Thakur, N.; Craswell, N.; and Lin, J. 2024. Umbrela: Umbrela is the (open-source reproduction of the) bing relevance assessor. *arXiv preprint arXiv:2406.06519*.