Unlocking Multi-View Insights in Knowledge-Dense Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

While Retrieval-Augmented Generation (RAG) plays a crucial role in the application of Large Language Models (LLMs), existing retrieval methods in knowledge-dense domains like law and medicine still suffer from a lack of multi-perspective views, which are essential for improving interpretability and reliability. Previous research on multi-view retrieval often focused solely on different semantic forms of queries, neglecting the expression of specific domain knowledge perspectives. This paper introduces a novel multi-view RAG framework, *MVRAG*, tailored for knowledge-dense domains that utilizes intention-aware query rewriting from multiple domain viewpoints to enhance retrieval precision, thereby improving the effectiveness of the final inference. Experiments conducted on legal and medical case retrieval demonstrate significant improvements in recall and precision rates with our framework. Our multi-perspective retrieval approach unleashes the potential of multi-view information enhancing RAG tasks, accelerating the further application of LLMs in knowledge-intensive fields.

1 Introduction

In the ever-evolving domain of natural language processing, retrieval-augmented generation (RAG) stands as a cornerstone technology that synergistically combines large language models (LLMs) with a vast corpus of external knowledge (Gao et al., 2023). This integration endows RAG systems with a remarkable capacity for nuanced language understanding and sophisticated information retrieval. Such capabilities are especially transformative in knowledge-dense domains like *law*, *medicine* and *biology*, where RAG not only augments contextual comprehension but also improves the interpretability and reliability of the models. Despite these advantages, current RAG implementations face substantial challenges in adequately serving the complicated and multifaceted information requisites inherent in these specialized fields.

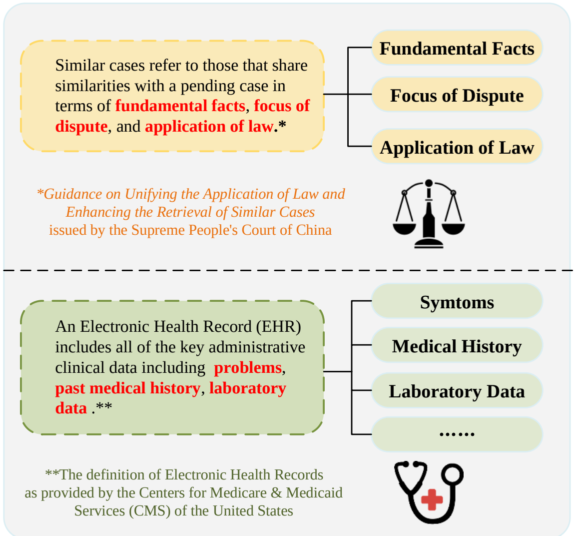

Figure 1: Examples of **professional perspectives** in legal and medical domain, constructed with the guidance of industry standards and norms.

Predominant among these challenges is the lack of multi-view domain information, leading into poor performance in some tasks, especially in knowledge-intensive domains. To be more specific, most RAG frameworks, mainly relying on vector-based similarity metrics for information retrieval, ignore the inner relationship and local similarity between queries and retrieved information. Figure 1 highlights the crucial multi-perspectives embedded in industry standards and norms. In the legal domain, for instance, the similarities in the *focus of dispute* among cases are more pivotal than superficial textual similarity; in medical scenarios, the pertinence of *medical history* or *symptoms* can be more diagnostic than literal similarity caused by a large amount of unimportant information. Traditional retrieval methods, hinged on full-text similarity, frequently fail to capture these essential professional nuances, leading to retrieval outcomes that may

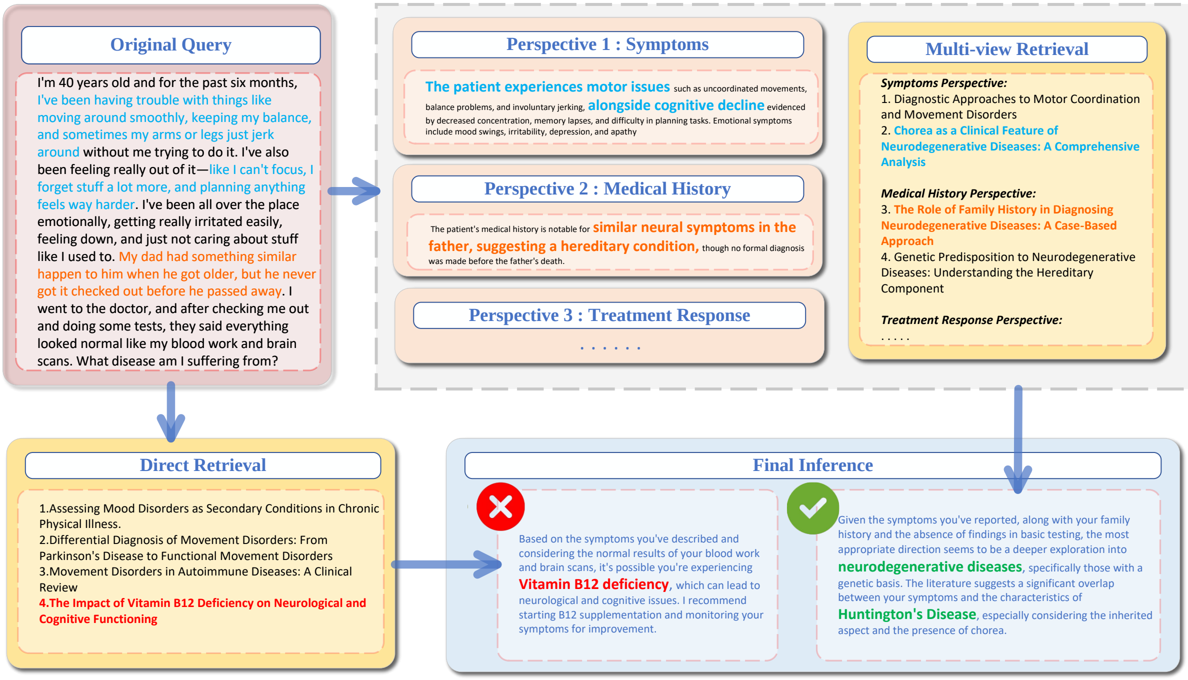

Figure 2: A case study showcasing the effectiveness of a multi-view retrieval framework in accurately diagnosing Huntington's disease. The model corrects an initial misdiagnosis of Vitamin B12 deficiency by refining the search criteria to focus on neurodegenerative symptoms and family medical history, thus demonstrating the importance of multi-view search strategies in medical diagnostics.

be imprecise, misleading, or lack of critical information. As is showcased in Figure 2, the absence of multi-view domain insights in retrieval systems can result in erroneous and potentially harmful outcomes, underscoring the critical need for a more nuanced and context-aware approach to information retrieval.

Previous work on multi-view RAG often focused on rewriting the original query in different textual forms, aiming to align better with search results across various semantic dimensions (Ma et al., 2023; Wang et al., 2023). However, such rewriting methods did not address the multi-view information inherent in the domain, failing to enhance the retrieval effectiveness in capturing local similarities. Although this approach has improved retrieval in general domains, it has not met the specialized retrieval needs of knowledge-dense domain LLMs.

To bridge this gap and address the limitations of previous methods, our study introduces a novel multi-view retrieval framework, *MVRAG*. This system is built around three pivotal stages: intention recognition, query rewriting, and retrieval augmentation. Initially, it utilizes a Large Language Model to identify the query's intent and assigns relevance scores to different professional perspectives, forming a *Perspective Vector*. This vector then informs the Query Rewriting stage, where the query is tailored to each identified perspective, ensuring the retrieval of contextually relevant documents. The final stage, Retrieval Augmentation, involves refining and integrating these documents to construct a comprehensive prompt, which is used to generate a context-aware, multi-faceted response.

Our multi-view query rewriting technique significantly improves the relevance and performance of information retrieval. It represents a paradigm shift from traditional methods, moving beyond mere lexical similarity to align closely with the multi-perspective intentions embedded in professional queries. This approach is expected to enhance the efficiency of RAG systems in specialized domains and potentially helps expand the application scope and interpretative depth of these models in complex environments. The contributions of this article mainly include the following three aspects:

- We unveil an advanced multi-perspective retrieval framework designed to navigate the complexities of knowledge-intensive fields. By leveraging intent recognition alongside multi-perspective question rewriting and matching rearrangement, our approach enriches retrieval outcomes with an array of information dimensions.

- We highlight the framework's transferability across various knowledge-dense domains, as evidenced by successful multi-domain case studies. These case studies validate the effectiveness of incorporating multi-perspective insights to achieve higher inference accuracy across diverse areas.

- Furthermore, we conduct detailed experiments to emphasize the crucial role of multi-perspective strategies in knowledge-danse RAG tasks and prove the effectiveness of multi-perspective information, thus paving the way for future investigations into multi-perspective methodologies.

2 Related Work

2.1 Retrieval-Augmented Generation (RAG)

Retrieval-augmented generation, commonly known as RAG (Izacard et al., 2022; Huo et al., 2023; Guu et al., 2020; Lewis et al., 2020) has become a prevalent technique to enhance LLMs. This approach integrates LLMs with retrieval systems, enabling them to access domain-specific knowledge and base their responses on factual information (Khattab et al., 2021). Additionally, RAG provides a layer of transparency and interpretability by allowing these systems to cite their sources (Shuster et al., 2021).

Recent studies have demonstrated enhancements in the quality of output results across various scenarios through techniques such as appending documents retrieved by RAG to the inputs of LLMs, and training unified embedded models. These methodologies have shown effective improvements in the performance of LLM outputs by leveraging RAG's ability to access and incorporate relevant information from external sources (Ram et al., 2023; Es et al., 2023).

2.2 Domain-Specific Large Language Models

Due to the general nature of LLMs, their expertise in specific domains such as *law*, *finance*, and *healthcare* is often limited. Recent research focuses on enhancing domain-specific expertise through Knowledge Enhancement techniques, introducing specific domain knowledge, and employing innovative training methods to address the issue of hallucinations in models. This approach has become a key strategy for improving the professionalism of vertical domain large models (Xi et al., 2023; Yao et al., 2023; Zhao et al., 2023; Zhu et al., 2023).

In the legal field, a number of well-known large language models have been born through methods such as secondary training, instruction fine-tuning, and RAG, such as wisdomInterrogatory[1], DISC-LawLLM[2], PKUlaw[3], and ChatLaw (Cui et al., 2023). These models have been specifically designed to address the unique requirements of the legal domain, providing specialized knowledge and expertise to enhance the quality and relevance of legal information retrieval and generation.

As for the medical domain, the development of domain-specific LLMs has been a key focus in recent research. Models like Huatuo[4], Zhongjing[5], and Doctor-Dignity[6] have been specifically designed to cater to the complex and nuanced requirements of the medical field, providing enhanced capabilities for medical information retrieval, diagnosis, and treatment planning.

2.3 Query Rewriting

In the process of RAG, challenges often arise from users' original queries being imprecisely phrased or lacking sufficient semantic information. Direct searches with such queries may lead Large Language Models to provide inaccurate or unanswerable responses. Thus, aligning the semantic space of user queries with that of document semantics is crucial. Query Rewriting techniques address this issue by refining the expression of queries, making them more precise and enriched, effectively bridging the gap and enhancing the accuracy and relevance of retrievals.

Gao et al. (2022) introduce an innovative approach predicated on the utilization of conjectural document embeddings, designed to maximize the congruence between query vectors and the corresponding authentic document representations, thereby augmenting the precision of query retrieval mechanisms. Concurrently, Wang et al. (2023) advocate for the employment of LLM as auxiliary tools in query formulation processes. Through the methodologies of query rewriting and the generation of pseudo-documents, among other modifica-

[1] https://github.com/zhihaiLLM/wisdomInterrogatory
[2] https://github.com/FudanDISC/DISC-LawLLM
[3] https://www.pkulaw.net/
[4] https://github.com/SCIR-HI/Huatuo-Llama-Med-Chinese
[5] https://github.com/SupritYoung/Zhongjing
[6] https://github.com/llSourcell/Doctor-Dignity

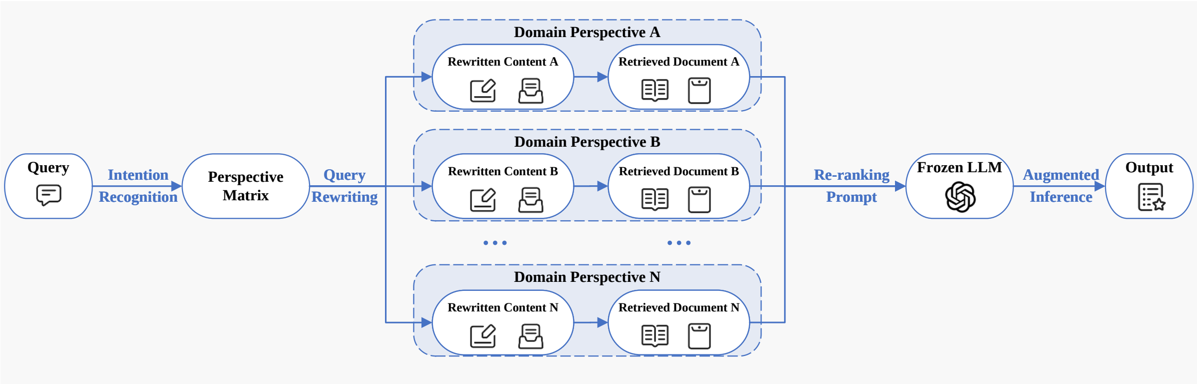

Figure 3: Framework of our Multi-View RAG System. This figure demonstrates the system's core processes: Intention Recognition, Query Rewriting, and Retrieval Augmentation, emphasizing the multi-view insights approach for intention-aware query rewriting

tions, their strategies have been empirically validated to deliver notable enhancements in retrieval outcomes (Ma et al., 2023).

Different from previous research, our work concentrates on the multi-perspective professional information within specific domains, which is a critical aspect often overlooked before. Instead of merely adjusting queries for semantic alignment, our novel multi-view retrieval framework captures the complex relationships and local nuances inherent to each domain, enhancing the retrieval process by focusing on the multi-perspective aspects of domain-specific knowledge.

3 Method

3.1 Overview

The method of our Multi-View Retrieval-Augmented Generation system consists of three main components: **Intention Recognition**, **Query Rewriting**, and **Retrieval Augmentation**. In the Intention Recognition phase, an LLM is used to identify and assign weights to various **professional perspectives** based on their relevance to a query, resulting in a *Perspective Vector*. This vector guides the Query Rewriting process, where *rewritten contents* are generated for each perspective. These contents are then used in a similarity-based search to retrieve relevant documents. Finally, in the Retrieval Augmentation stage, the retrieved documents are re-ranked based on their relevance score and then integrated into a structured prompt for generating the final response. This methodology ensures accurate, contextually relevant, and multi-perspective responses to complex queries. Our overall method is illustrated in Figure 3.

3.2 Intention Recognition

At the beginning of our method, we enlist the help of experts and scholars in the specific domain to identify the perspectives from which the query should be rewritten. This is done by consulting domain standards, aiming to satisfy task requirements and enhance interpretability.

As is described in Figure 1, domain experts are expected to draw upon the norms and regulations of specific domains to establish these perspectives effectively. This structured approach is vital for conducting a thorough analysis and ensuring the efficacy of the retrieval outcomes. For example, within the legal domain, reference to the authoritative opinions issued by the Supreme People's Court of China permits the identification of fundamental perspectives including *fundamental facts*, *focus of dispute* and *application of law*.

Building on this, our proposed system utilizes a Large Language Model to analyze the original query's underlying intention. Instead of a general quantitative function, we specifically employ $LLM_{score}(q, p_i)$ to directly obtain relevance scores from the LLM, assigning weight values to predefined professional perspectives based on their degree of relevance to the query. Let $P = \{p_1, p_2, ..., p_n\}$ be the set of predefined **professional perspectives**, where each p_i represents a distinct perspective. For a given query q, $LLM_{score}(q, p_i)$ evaluates the semantic alignment between q and each perspective p_i directly using the LLM. The weight value w_i assigned to each perspective is calculated using the following expression:

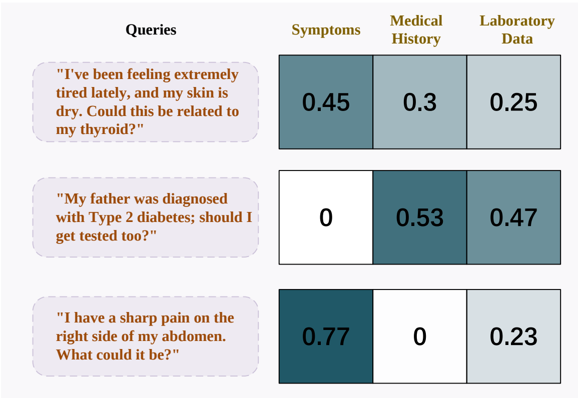

Figure 4: The visualization of Perspective Vectors, using varying shades of color to represent different normalized scores. The graph displays various scenarios corresponding to different types of queries, with scores below a threshold set to zero.

$$w_i = \begin{cases} LLM_{score}(q, p_i) & \text{if } LLM_{score}(q, p_i) > \theta \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $LLM_{score}(q, p_i)$ is the relevance score determined directly by the LLM for the alignment between the query q and the perspective p_i, and θ is a predefined threshold. If the relevance score $LLM_{score}(q, p_i)$ is greater than θ, the weight w_i is set to the relevance score itself. Otherwise, in cases where the score does not surpass the threshold, the weight w_i is set to 0. Figure 4 demonstrates the visualization of Perspective Vectors in several scenarios.

The result of this process is encapsulated in a *Perspective Vector* V_q for the query q, represented as:

$$V_q = \begin{bmatrix} w_1 & w_2 & \dots & w_n \end{bmatrix}.$$

This vector V_q quantitatively expresses the extent to which the query aligns with each professional perspective in the set P. The i-th element of the vector, w_i, represents the weight assigned to the i-th perspective p_i , reflecting the degree to which the query is related to the professional perspective p_i.

3.3 Query Rewriting and Alignment with Perspective Dimensions

After the acquisition of the *Perspective Vector* V_q, the focal task is to align the query with the identified dimensions of professional perspectives. This is achieved through a methodology termed *query rewriting*. Employing a large-scale language model, known as the **rewriter**, the system generates rewritten content C_i for each perspective p_i with an associated non-zero weight w_i in V_q.The function for generating rewritten content C_i can be represented as follows:

$$C_i = \text{Rewriter}(q, p_i) \quad \forall\, p_i \text{ with } w_i \neq 0 \tag{2}$$

where q represents the original query, p_i indicates a specific perspective, and w_i denotes the weight of perspective p_i in the Perspective Vector V_q. The **rewriter** model is designed to utilize its excellent natural language processing capabilities to generate rewritten content C_i that is contextually aligned with the professional perspective p_i, ensuring that the query is tailored to the specific nuances and requirements of the domain.

Upon generation of C_i, the system conducts a similarity-based search to extract a contextually relevant set of documents. The number of documents retrieved, denoted by k, is depended on the weight w_i. Higher weights lead to a broader retrieval scope, symbolizing a stronger relevance to the perspective in question. The resultant set of top k documents for perspective p_i is denoted as R_i.

$$R_i = \left\{ d_{ij} \;\middle|\; \begin{array}{l} d_{ij} \text{ ranks among the top } k \text{ by} \\ \text{Similarity}(d_{ij}, C_i), \forall j \in \{1, \dots, k\} \end{array} \right\}, \tag{3}$$

where d_{ij} represents the j-th document in the retrieved set R_i, and $\text{Similarity}(d_{ij}, C_i)$ denotes the cosine similarity score between the document d_{ij} and the rewritten content C_i.

This phase results in the assembly of a collection of document sets $\{R_1, R_2, ..., R_n\}$, each tailored to a unique professional perspective. This strategy ensures that the query is not only contextually aligned but also expansively addresses the multifaceted aspects of the query, thereby significantly enhancing the specificity and relevance of the information retrieval in specialized domains.

3.4 Retrieval Augmentation and Final Inference

Following the assembly of the retrieved document sets $\{R_1, R_2, \dots, R_n\}$, the system undertakes a refinement process to enhance the relevance and utility of the documents. This process is governed by the original *Perspective Vector* V_q and involves re-ranking and integrating the documents for final prompt generation.

The re-ranking process employs a ranking model, specifically designed to evaluate the alignment of each document d_{ij} in set R_i with the query's multi-perspectives. This evaluation considers both the similarity score obtained from the retrieval phase and the weight of the perspective p_i in the Perspective Vector V_q. Consequently, the relevance score $\text{Rel}(d_{ij})$ for each document d_{ij} is recalculated as follows:

$$\text{Rel}(d_{ij}) = \text{Similarity}(d_{ij}, C_i) \times w_i, \quad (4)$$

where $\text{Similarity}(d_{ij}, q)$ denotes the cosine similarity score between the document d_{ij} and rewritten content C_i, and w_i represents the weight of perspective p_i in V_q.

Documents are then re-ranked based on their recalculated relevance scores $\text{Rel}(d_{ij})$, ensuring that the final ordering reflects both the inherent relevance to the query and the significance of each perspective. Once re-ranked, the documents are integrated into a structured prompt, denoted as $\mathcal{P}$. This prompt is a composition of the original query q and the restructured document sets, where documents are concatenated with the query to form a single, cohesive prompt:

$$\mathcal{P} = q \bowtie \bigoplus_{i=1}^{n} \bigoplus_{j=1}^{k_i} d'_{ij}, \quad (5)$$

where d'_{ij} represents the re-ranked documents in set R_i and $\bowtie$ represents the concatenation operation.

This prompt is then fed into a large-scale reader model $\mathcal{M}$ for final inference. The reader model processes $\mathcal{P}$ and generates a response that encapsulates multi-perspective information, suitable for addressing the query's complexities.

By integrating a comprehensive retrieval augmentation process, the system ensures multi-perspective analysis and interpretation of the query, leading to the generation of nuanced and contextually rich responses. This approach is particularly effective in specialized domains where depth and diversity of perspectives are essential.

4 Experiment

4.1 Dataset

The dataset chosen for our experiment in the legal domain is the *LeCaRDv2* dataset (Li et al., 2023), distinguished within the legal case retrieval domain for its comprehensive collection of cases and meticulous construction methodology. To more effectively illustrate the application efficacy of our multi-perspective framework in addressing complex real-world issues, we have curated a selection of 90 uncommon instances from our dataset to serve as queries. These instances encompass 15 distinct and relatively infrequent types of legal cases, including *illegal deforestation* and *the falsification of official documents by state bodies*, thereby introducing elevated challenges for the task of analogous case retrieval within the legal domain.

As for the medical domain, we select the *PMC-Patients* dataset (Zhao et al., 2022), which represents a pioneering compilation, comprising 167, 000 patient summaries extracted from case reports within PubMed Central (PMC)[7]. Following the previous approach, we also selected 90 cases that performed poorly on the baseline model as the test set for experimentation, to better showcase the application effects in complex situations.

4.2 Baseline

To illustrate our framework's efficacy, we established baselines using prevalent pre-trained models within the *Dense* framework, adhering to dataset-specific settings. Here, we briefly summarize the baseline models for the English and Chinese datasets:

For the English PMC-patient dataset:

- **bge-large-en-v1.5**: A model with a transformer architecture for English text retrieval.[8]
- **all-MiniLM**: Converts text to dense vectors, ideal for semantic search tasks.[9]
- **gte-large**: Focused on relevance text pairs across diverse domains.[10]

For the Chinese LeCaRDv2 dataset:

- **bge-large-zh-v1.5**: Tailored for Chinese text retrieval with a transformer architecture.[11]
- **text2vec**: Utilizes CoSENT for Chinese semantic analysis.[12]

[7] https://www.ncbi.nlm.nih.gov/pmc/
[8] http://huggingface.co/BAAI/bge-large-en-v1.5
[9] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[10] https://huggingface.co/thenlper/gte-large
[11] http://huggingface.co/BAAI/bge-large-zh-v1.5
[12] https://huggingface.co/GanymedeNil/text2vec-large-chinese

Table 1: Retrieval Evaluation Results on Legal Case Retrieval Dataset (LeCaRDv2).

Model	Recall@100 (%)	Precision@100 (%)	F1 Score (%)
bge-m3 (Baseline)	3.125	0.938	1.442
bge-m3 (MVRAG)	**16.53**	**4.959**	**7.629**
bge-large-zh (Baseline)	7.343	2.204	3.391
bge-large-zh (MVRAG)	15.13	4.533	6.977
text2vec (Baseline)	3.959	1.187	1.827
text2vec (MVRAG)	7.675	2.303	3.542

Table 2: Retrieval Evaluation Results on Medicalal Case Retrieval Dataset (PMC-Patients)

Model	Recall@100 (%)	Precision@100 (%)	F1 Score (%)
gte-large (Baseline)	1.595	0.109	0.204
gte-large (MVRAG)	13.05	2.134	3.668
bge-large-en (Baseline)	8.791	1.672	2.809
bge-large-en (MVRAG)	**15.14**	**2.234**	**3.893**
all-MiniLM-L6 (Baseline)	8.285	0.995	1.776
all-MiniLM-L6 (MVRAG)	9.875	1.720	3.019

- **bge-m3**: Offers multi-lingual support and integrates various retrieval methods.[13]

The **GPT-4** model functions as our rewriter, leveraging its language understanding to align queries with identified perspectives, crucial for the query rewriting phase.

4.3 Results

Our results, as outlined in Tables 1 and 2, demonstrate the Multi-View framework's profound impact on enhancing RAG models. This integration markedly improves retrieval metrics across recall@100, precision@100, and F1 Score.

For the Legal Case Retrieval dataset, the framework significantly lifted the recall@100 for the bge-m3 model from 3.125% to 16.53%, evidencing its effectiveness. The bge-large-zh model saw recall@100 rise from 7.343% to 15.13%. Similarly, the text2vec model's recall@100 nearly doubled, showcasing consistent performance enhancements.

In the Medical Case Retrieval dataset, the gte-large model's recall@100 improved from 1.595% to 13.05%, while the bge-large-en model achieved a recall@100 of 15.14%. Even models with modest improvements, like all-MiniLM-L6, benefitted, showing enhanced recall@100 and F1 Scores.

For certain models that experience significant improvements, we posit that this is attributable to their superior grasp of global information. Upon rewriting, the information obtained is typically shorter than the original content, thus models that are better trained on aligning disparate lengths of text exhibit more substantial enhancements under our framework.

These outcomes affirm the multi-view framework's efficacy in improving recall, precision, and F1 Scores, emphasizing its value in refining retrieval processes for legal and medical domains. The framework's integration into RAG systems substantially increases document retrieval scope and accuracy, ensuring high-relevance information retrieval for domain-specific tasks.

4.4 Ablation Study

To quantify the impact of perspective selection strategies on retrieval performance, we executed ablation experiments on the LeCaRDv2 and PMC-patient datasets. In the legal domain, our original multi-view framework integrates four perspectives: *Fundamental Facts*, *Focus of Dispute*, *Application of Law*, and *Penalty*, while we utilize *Medical History, Symptoms, Laboratory Data* and *Treatment Response* for the medical task. We compared the full four-perspective approach against configurations where one perspective was omitted at a time, alongside a direct retrieval baseline.

For the law dataset, the multi-view framework achieved the highest recall@100 of 16.53%, significantly higher than the baseline's 3.11%. Excluding Fundamental Facts led to a noticeable drop in recall@100 to 12.11%, indicating its vital role. Omitting Penalty had the least impact, slightly reducing

[13] https://huggingface.co/BAAI/bge-m3

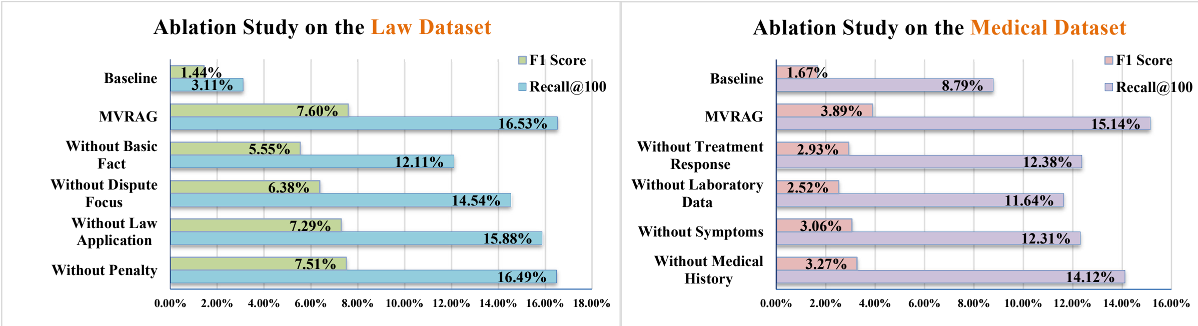

Figure 5: Ablation Study on impact of perspective selection strategies in our framework on Medical and Legal datasets. In the legal domain, the chart shows Recall@100 and F1 Score after excluding each perspective: *Fundamental Facts, Focus of Dispute, Application of Law, Penalty*. For the medical domain, it displays the effects of removing *Medical History, Symptoms, Laboratory Data, Treatment Response*. Each bar indicates the performance impact versus the full baseline and direct retrieval.

recall to 16.49%.

In the medical dataset, our approach attained a recall@100 of 15.14%. The absence of Medical History caused a considerable decline to 14.12%, demonstrating its importance. Similar to the law dataset, some perspectives had a lesser impact, implying potential information overlap.

These outcomes validate the multi-perspective model's robustness and highlight the nuanced contribution of each perspective towards enhancing retrieval accuracy in complex domains.

5 Case Study

To better showcase the effectiveness of our model, we conduct some case studies in several knowledge-dense domains, including *Biology*, *Geography*, *Literature*, *Political Science* and *Physics*. An extensive collection of detailed case studies across a wide range of disciplines are provided in the Appendix A for further exploration.

Within this section, we delve into the medical domain, employing the *PMC-Patients* dataset as the retrieval database. As illustrated in Figure 2, we selected a case of **Huntington's disease** from *PubMed* and transformed it into a query voiced by the patient, serving as our original query. In scenarios absent of our multi-view retrieval model, the retrieval system primarily fetched articles related to superficial symptoms such as emotional dysregulation and movement disorders. This surface-level matching, due to similar characterizations, erroneously led to articles on **Vitamin B12 deficiency**, consequently misdiagnosing the condition as a **Vitamin B12 deficiency**.

Conversely, our multi-view retrieval system initiates with intent recognition, prioritizing symptoms and medical history for query reformulation. Through individual matching and re-ranking, it yielded comparatively relevant search outcomes. Specifically, symptom keywords, refined with greater precision, directed the search towards articles on neurodegenerative diseases. Simultaneously, the emphasis on similar family medical history surfaced articles on genetic testing. With the aid of these references, the model adeptly identified the potential for **Huntington's disease**.

This case study emphatically demonstrates the superiority of our multi-view retrieval model. With the assistance of multi-view reformulation, it adeptly navigated towards more pertinent and informative references, thereby significantly enhancing the accuracy of the final diagnostic inference.

6 Conclusion

In this paper, we propose a multi-perspective approach to Retrieval-Augmented Generation tailored for knowledge-dense domains, aiming to incorporate the domain-specific insights missing from existing methods and enhance the reliability and interpretability of retrieval outcomes. By employing intent recognition, multi-perspective query rewriting, and document re-ranking, we have significantly improved retrieval performance in knowledge-dense areas such as law and medicine. Through experiments and case studies, we demonstrate the impact of integrating multi-perspective information in these domains, laying the groundwork for future incorporation of multi-perspective views into RAG systems.

Limitations

While our multi-view retrieval framework has shown promising results in enhancing retrieval performance, there are several limitations to consider. Firstly, the reliance on domain experts for perspective identification may introduce bias and limit the scalability of the model. Secondly, the lack of a unified evaluation metric for multi-perspective retrieval systems poses challenges in comparing performance across different domains. Lastly, the computational complexity of multi-perspective retrieval may hinder real-time applications in certain scenarios. Future research should focus on addressing these limitations and further exploring the potential of multi-perspective retrieval in specialized domains.

References

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels, 2022. *URL https://arxiv.org/abs/2212.10496*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Siqing Huo, Negar Arabzadeh, and Charles LA Clarke. 2023. Retrieving supporting evidence for llms generated answers. *arXiv preprint arXiv:2306.13781*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Relevance-guided supervision for openqa with colbert. *Transactions of the association for computational linguistics*, 9:929–944.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2023. Lecardv2: A large-scale chinese legal case retrieval dataset. *arXiv preprint arXiv:2310.17609*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

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

Yunjia Xi, Weiwen Liu, Jianghao Lin, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, Rui Zhang, and Yong Yu. 2023. Towards open-world recommendation with knowledge augmentation from large language models. *arXiv preprint arXiv:2306.10933*.

Jing Yao, Wei Xu, Jianxun Lian, Xiting Wang, Xiaoyuan Yi, and Xing Xie. 2023. Knowledge plugins: Enhancing large language models for domain-specific recommendations. *arXiv preprint arXiv:2311.10779*.

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*.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2022. Pmc-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems. *arXiv preprint arXiv:2202.13876*.

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*.

Appendix A Case Study Appendix

A.1 Biology

Domain	**Domain Perspectives**
Biology	Gene Expression Protein Function Species Evolution
Original Query	**Rewritten Queries**
Why do some species' pigment deposits change with environmental variations?	***Gene Expression Perspective***: "How do environmental factors affect gene expression in species leading to changes in pigment deposition?" ***Protein Function Perspective***: "What impact does environmental change have on the function and activity of pigment proteins in species?" ***Species Evolution Perspective***: "How do changes in pigment deposition reflect the adaptation and evolution of species to environmental changes?"
Direct Retrieval	**Multi-View Retrieval**
"Environmental Pollution and Animal Pigment Changes" "The Impact of Global Warming on Plant Colors" "The Effect of Light on Butterfly Wing Coloration"	***Gene Expression Perspective***: "Gene Expression Regulation and Pigment Accumulation under Environmental Stress" "Adaptive Gene Expression Changes in Response to Environmental Light Variations" ***Protein Function Perspective***: "The Role of Pigment Proteins in Environmental Adaptation" "Functional Diversity of Pigment Proteins under Environmental Pressure" ***Species Evolution Perspective***: "Molecular Mechanisms of Species Color Evolution Prompted by Environmental Changes" "Evolutionary Adaptations in Pigmentation: Evidence from Environmental Interactions"
Inference	**Multi-View Inference**

Based on our findings, it appears that changes in species' pigment deposits are primarily due to direct environmental impacts, such as pollution, global warming, and variations in light exposure. These external factors seem to straightforwardly affect the coloration without involving complex biological processes.	Upon reviewing comprehensive articles across multiple perspectives, we can conclude that the variations in pigment deposits among species in response to environmental changes are not merely direct reactions. Instead, these changes are the result of intricate biological mechanisms. Environmental factors influence gene expression and the functionality of pigment proteins, which in turn contribute to the species' evolutionary adaptation to their habitats. Specifically, temperature and light variations play crucial roles in modulating gene expression related to pigment deposition, highlighting the adaptive and evolutionary significance of these pigment changes.

A.2 Geography

Domain	**Domain Perspectives**
Geography	Climate Change Urban Development Biodiversity
Original Query	**Rewritten Queries**
How does the expansion of urban areas affect local climates and ecosystems?	***Climate Change Perspective***: "In what ways does urban expansion contribute to changes in local climate patterns?" ***Urban Development Perspective***: "How does the planning and growth of urban areas influence their surrounding ecosystems and climate?" ***Biodiversity Perspective***: "What impact does urban sprawl have on the biodiversity of adjacent natural habitats?"
Direct Retrieval	**Multi-View Retrieval**

"The Heat Island Effect in Mega Cities" "Urbanization and Its Impact on Rainfall Variability" "Effects of Concrete Surfaces on Urban Temperatures"	***Climate Change Perspective***: "Urban Expansion and Its Role in Altering Regional Climate Systems" "Greenhouse Gas Emissions from Urban Centers: A Climate Change Perspective" ***Urban Development Perspective***: "Sustainable Urban Planning: Balancing Growth and Environmental Preservation" "The Influence of Urban Landscape Design on Local Ecosystems" ***Biodiversity Perspective***: "Urbanization's Toll on Local Wildlife: Case Studies of Habitat Fragmentation" "Conservation Strategies for Mitigating Urbanization Effects on Biodiversity"

Inference	**Multi-View Inference**
The research primarily points to urban areas contributing to higher temperatures through the heat island effect and altering rainfall patterns due to extensive concrete surfaces. It suggests that the primary impact of urban expansion is a direct increase in local temperatures and changes in precipitation.	After analyzing data from various perspectives, it is evident that urban expansion affects local climates and ecosystems in multiple interconnected ways. Urban growth leads to increased greenhouse gas emissions, significantly influencing local climate patterns beyond just temperature increases, such as altering precipitation and wind patterns. From an urban development viewpoint, the layout and planning of cities can either harm or help preserve local ecosystems, highlighting the importance of sustainable development to mitigate negative impacts. Furthermore, urban sprawl severely affects biodiversity, leading to habitat fragmentation and loss. Therefore, the expansion of urban areas requires careful consideration of its broader environmental impacts, necessitating strategies that prioritize sustainability and biodiversity conservation.

A.3 Literature

Domain	**Domain Perspectives**
Literature	Narrative Techniques Historical Context Cultural Impact
Original Query	**Rewritten Queries**

How do modern novels reflect contemporary societal issues?	***Narrative Techniques Perspective***: "What narrative techniques are modern novelists using to explore and reflect contemporary societal issues?" ***Historical Context Perspective***: "How does the historical context of the early 21st century influence the themes and subjects of modern novels?" ***Cultural Impact Perspective***: "In what ways do modern novels influence and reflect the cultural attitudes and social issues of our time?"
Direct Retrieval	**Multi-View Retrieval**
"The Rise of the Digital Novel: Technology in Modern Literature" "Post-Modernism and Its Influence on 21st Century Literature" "The Evolution of Character Archetypes in Contemporary Fiction"	***Narrative Techniques Perspective***: "Stream of Consciousness and Nonlinear Narratives in Depicting Modern Complexities" "The Role of Metafiction in Critiquing Contemporary Societal Norms" ***Historical Context Perspective***: "Post-9/11 Literature: Terrorism and Its Aftermath in Modern Novels" "Economic Crises and Their Reflections in 21st Century Fiction" ***Cultural Impact Perspective***: "Literature as a Mirror to the #MeToo Movement: Narratives of Empowerment and Justice" "Climate Fiction: How Modern Novels Address Environmental Concerns"
Inference	**Multi-View Inference**

Based on the articles retrieved, it appears that modern novels primarily explore technological advancements and their implications, with a focus on the stylistic elements of post-modernism and the evolution of character archetypes. This suggests a significant emphasis on form and structure over content when reflecting contemporary societal issues.	Upon examining articles from diverse perspectives, it becomes clear that modern novels deeply engage with contemporary societal issues through various means. Through innovative narrative techniques, they offer nuanced explorations of complex issues like global terrorism, environmental crises, and social justice movements, such as #MeToo. The historical context of the early 21st century, marked by significant global events and economic turmoil, profoundly influences the thematic preoccupations of contemporary literature. Moreover, modern novels not only reflect but also actively shape cultural attitudes towards pressing social issues, demonstrating literature's power to influence societal change and public discourse. Therefore, modern novels serve as a crucial lens through which the multifaceted concerns and dynamics of contemporary society are examined and understood.

A.4 Political Science

Domain	**Domain Perspectives**
Political Science	Information Dissemination Voter Behavior Regulatory Policies
Original Query	**Rewritten Queries**
How do social media platforms influence political discourse and public opinion?	***Information Dissemination Perspective***: "What role do social media platforms play in the spread of political information and the formation of public opinion?" ***Voter Behavior Perspective***: "How are social media platforms affecting voter behavior and electoral outcomes in democratic societies?" ***Regulatory Policies Perspective***: "What regulatory measures are being implemented to ensure the integrity of political discourse on social media platforms?"
Direct Retrieval	**Multi-View Retrieval**

"The Rise of Social Media in Political Campaigns" "Social Media: The New Public Square for Political Discussion" "Echo Chambers and Filter Bubbles: The Polarization of Political Discourse on Social Media"	***Information Dissemination Perspective***: "Algorithmic Bias and News Feed Algorithms: Shaping Political Information on Social Media" "The Role of Social Media in Civic Engagement and Political Mobilization" ***Voter Behavior Perspective***: "Social Media's Impact on Voter Turnout and Political Participation: A Global Perspective" "The Influence of Online Political Advertising on Voter Preferences and Decisions" ***Regulatory Policies Perspective***: "Regulating Political Advertising on Social Media: Challenges and Approaches" "Social Media Governance: Balancing Free Speech with Political Integrity"

Inference	**Multi-View Inference**
Based on the directly retrieved articles, it might be concluded that social media primarily serves as a new platform for political campaigns, facilitating public political discussions and contributing to political polarization through echo chambers and filter bubbles. This perspective suggests a neutral to negative impact, focusing on the divisive aspects of social media's role in politics.	Integrating insights from articles across multiple perspectives reveals a nuanced understanding of social media's role in political discourse and public opinion. Social media platforms are crucial for disseminating political information, engaging citizens in political processes, and mobilizing voter participation. However, challenges such as algorithmic biases can shape political information in ways that may not always be transparent or equitable, potentially influencing voter behavior and preferences. Furthermore, the advent of online political advertising highlights the need for regulatory policies to ensure the integrity of political discourse, balancing the protection of free speech with the prevention of misinformation and the maintenance of political integrity. To address these complexities effectively, a comprehensive approach involving all stakeholders—platforms, policymakers, and the public—is essential for harnessing social media's potential as a force for democratic engagement while mitigating its risks.

A.5 Physics

Domain	**Domain Perspectives**
Physics	Quantum Mechanics Cryptography Technological Advancement
Original Query	**Rewritten Queries**

What are the implications of quantum computing for traditional encryption methods?	***Quantum Mechanics Perspective***: "How do principles of quantum mechanics underpin the operation of quantum computers and their impact on encryption?" ***Cryptography Perspective***: "What challenges does quantum computing pose to current cryptographic algorithms and traditional encryption methods?" ***Technological Advancement Perspective***: "Considering the advancements in quantum computing, what are the future prospects for secure communication technologies?"
Direct Retrieval	**Multi-View Retrieval**
"Introduction to Quantum Computing" "The Basics of Cryptography" "Quantum Computers and the Future of Encryption"	***Quantum Mechanics Perspective***: "Entanglement and Superposition: Pillars of Quantum Computing" "Quantum Algorithms: Redefining Computational Capabilities and Encryption" ***Cryptography Perspective***: "Quantum Computing vs. RSA Encryption: A Cryptographic Analysis" "Post-Quantum Cryptography: Preparing for the Quantum Challenge" ***Technological Advancement Perspective***: "The Race for Quantum Supremacy: Implications for Global Communication Security" "Innovations in Quantum-Resistant Encryption Methods for Future Technologies"
Inference	**Multi-View Inference**

From the directly retrieved articles, one might infer that quantum computing represents a futuristic advancement with the potential to enhance computational speed and efficiency, impacting encryption in undefined ways. This perspective might suggest a gradual and manageable transition from traditional to quantum-resistant encryption methods, underestimating the immediacy and severity of the challenge posed by quantum computing to current encryption standards.	Upon analyzing articles from a multi-perspective approach, it becomes evident that quantum computing introduces profound implications for traditional encryption methods. Principles of quantum mechanics, such as entanglement and superposition, enable quantum computers to potentially break current cryptographic algorithms, including RSA, in a fraction of the time taken by classical computers. This reality necessitates urgent development in post-quantum cryptography to safeguard secure communication technologies against quantum attacks. The ongoing advancements in quantum computing not only challenge existing cryptographic practices but also push the boundary of technological innovation, driving the creation of quantum-resistant encryption methods. Thus, the transition to quantum-safe encryption is not merely an incremental upgrade but a critical evolution required to maintain the confidentiality and integrity of digital communications in the quantum era.