Query pipeline optimization for cancer patient question answering systems

Maolin He, Rena Gao, Mike Conway, and Brian E. Chapman

***Abstract*— Retrieval-augmented generation (RAG) mitigates hallucination in Large Language Models (LLMs) by using query pipelines to retrieve relevant external information and grounding responses in retrieved knowledge. However, query pipeline optimization for cancer patient question-answering (CPQA) systems requires separately optimizing multiple components with domain-specific considerations. We propose a novel three-aspect optimization approach for the RAG query pipeline in CPQA systems, utilizing public biomedical databases like PubMed and PubMed Central. Our optimization includes: (1) document retrieval, utilizing a comparative analysis of NCBI resources and introducing Hybrid Semantic Real-time Document Retrieval (HSRDR); (2) passage retrieval, identifying optimal pairings of dense retrievers and rerankers; and (3) semantic representation, introducing Semantic Enhanced Overlap Segmentation (SEOS) for improved contextual understanding. On a custom-developed dataset tailored for cancer-related inquiries, our optimized RAG approach improved the answer accuracy of Claude-3-haiku by 5.24% over chain-of-thought prompting and about 3% over a naive RAG setup. This study highlights the importance of domain-specific query optimization in realizing the full potential of RAG and provides a robust framework for building more accurate and reliable CPQA systems, advancing the development of RAG-based biomedical systems.**

***Index Terms*— Biomedical computing, Oncology, Large language models**

I. Introduction

Question-answering (QA) tasks are crucial in the biomedical domain, where timely and accurate responses can impact human lives. With more than a million new citations added to PubMed annually [1], healthcare professionals and patients face an overwhelming influx of information, highlighting the need to quickly process, analyze and summarize the vast biomedical literature. Large Language Models (LLMs) have revolutionized QA tasks in diverse domains [2]. Unlike traditional search engines that rely on keyword matching, LLMs leverage transformer architectures to capture semantic relationships and nuances, enabling them to find semantically relevant information and process it into precise, coherent answers. Thus, LLM-based QA systems reduce users' need to synthesize the data manually. However, LLMs face a significant challenge: hallucination—producing fluent but unfaithful or nonsensical responses [3]. Further, LLMs rely on pre-trained data, which may lack domain-specific or real-time knowledge [2]. These issues are particularly acute in healthcare [4], where inaccuracy can have severe consequences [5], requiring the QA system to demonstrate accuracy, reliability, and currency.

Retrieval-Augmented Generation (RAG) [6] is a solution to these challenges by guiding LLMs in generating accurate responses by retrieving relevant external information, rather than relying solely on the model's neural weights. This approach can enhance performance in knowledge-intensive tasks [7] and open-domain QA. Especially in medical QA systems where questions are knowledge-intensive, LLMs excel as generators rather than knowledge databases [8]. Retrieval quality is crucial for RAG performance [9] due to the "distraction phenomenon," where irrelevant retrieval results in the prompt degrade response quality [10]. Prior work on retrieval quality has explored multi-step retrieval methods, such as recursive [11], iterative [12], and multi-hop retrieval [13]. These methods repeatedly use the query pipeline to retrieve potentially query-related evidence, but irrelevant retrieval results in any step can cause cascading errors [14], highlighting the need for query pipeline optimization. However, systematic optimization of the query pipeline in a RAG-based Cancer Patient Question Answering (CPQA) system remains underexplored.

Query pipeline data for CPQA requires reliability and accessibility. Thus, we used two knowledge resources provided by the National Center for Biotechnology Information (NCBI) through E-Utilities (APIs for search and download) [15]: PubMed, which provides literature abstracts [16], and PubMed Central (PMC), which offers full-text articles for a subset of the biomedical literature [17]. While most studies rely on PubMed [18], [19], we explored the utility of different NCBI resources for document retrieval. Additionally, we optimized three key query pipeline components. Specifically, we investigated three document retrieval methods, explored combinations of four dense retrievers (embedding models) and two reranker models for two-stage passage retrieval, and developed a novel text segmentation technique for precise semantic representation. By keeping the generation module fixed, the accuracy of the generated answers on our comprehensive

Maolin He is with the School of Computing and Information Systems, University of Melbourne, Parkville, VIC 3052 Australia (e-mail: maolinh@student.unimelb.edu.au)

Rena Gao is with the School of Computing and Information Systems, University of Melbourne, Parkville, VIC 3052, Australia (e-mail: wegao@student.unimelb.edu.au).

Mike Conway is with the School of Computing and Information Systems, University of Melbourne, Parkville, VIC 3052 Australia (e-mail: mike.conway@unimelb.edu.au).

Brian E. Chapman is with the School of Computing and Information Systems, University of Melbourne, Parville, VIC 3052, Australia (e-mail: chapmanbe@gmail.com).

cancer QA dataset can be used to identify optimal knowledge sources and assess the performance changes due to different query pipeline components. Our contributions include:

- We make the first comparative analysis of different NCBI sources for CPQA, showing that PMC reviews have a higher retrieval value than non-review PMC papers, while PubMed abstracts dominate retrieval data sources.
- For NCBI resources, we propose **H**ybrid **S**emantic **R**eal-time **D**ocument **R**etrieval (HSRDR), the first work to combine real-time Boolean search (via E-Utilities with LLM-rewritten queries) and MedCPT [20] search.
- For passage retrieval, we investigate optimal pairings of dense retrievers and rerankers, revealing rerankers varied impacts on dense retrievers and the superior performance of domain-specific over general models.
- We propose **S**emantic **E**nhanced **O**verlap **S**egmentation (SEOS), a novel text segmentation method integrating sentence semantics and embedding model impacts while utilizing chunk overlap for richer context.

II. Background

A. Data Source and Methods of Document Retrieval

Document retrieval identifies candidate documents for answer generation [21]. Utilizing PubMed and PMC for CPQA presents challenges: many articles lack valid abstracts [2], and older publications are often irrelevant to contemporary cancer queries. NCBI online searches' performance relies on effectively using Medical Subject Heading (MeSH) terms and Boolean operators. E-Utilities can employ its built-in translator to incorporate relevant MeSH terms, lexical variants, and common synonyms for medical term expansion to enhance query coverage [15], making it suitable for online search patient queries. However, the E-Utilities tool remains limited by the need to convert original queries to Boolean operators and its term-based orientation, which constrains its capacity for semantic processing. While semantic search improves query understanding, it can miss recent publications and require more overhead and extensive preprocessing. These issues highlight the need for innovative document retrieval approaches.

B. Two-Stage Passage Retrieval

The two main retrieval methods are (i) sparse retrieval (such as BM25 [22]), which struggles with lexicon mismatches and capturing semantic relationships, and (ii) dense retrieval, which improves semantic matching by encoding text into dense vector representations and calculating vector similarity. Due to vocabulary and semantic shifts, effective dense retrieval in the medical domain requires adapting embedding models (dense retrievers) via domain-specific continuous pre-training or fine-tuning [20], which is expensive and time-consuming. Therefore, it is practical to use off-the-shelf, domain-specific dense retrievers. Retrieval quality improves when rerankers, trained on bi-encoders or cross-encoders [23], perform detailed similarity assessments and then reorder the top-k results from dense retrievers [24]. This two-stage retrieval approach requires compatibility between the embedding models and the reranking models. Exploring the optimal pairing of these two components is therefore crucial.

C. Precise Semantic Representation

Precise semantic representation is essential for information retrieval, requiring two key elements: precise text representation such as the word embeddings used in dense retrievers, and optimized text chunking that determines the granularity of document segmentation. Naive chunking methods, splitting documents into passages with fixed chunk size (the maximum tokens per chunk) [25], risk truncating sentences mid-way and losing ordering information between chunks. To mitigate these issues, the Sentence Splitter in LlamaIndex, for example, parses text with a preference for complete sentences and introduces chunk overlap, where adjacent chunks share a specified number of tokens [26]. An advanced variant, the Sentence Window Splitter, expands each sentence-group chunk to include a fixed-size window of surrounding sentences. However, these methods' reliance on predefined fixed parameters (chunk size and window size) can lead to inflexible chunk boundaries. The text tiling algorithm [27] uses lexical similarity between sentence groups to identify more natural topic boundaries, but its reliance on simple lexical matching limits its effectiveness in domains with rich semantically equivalent but lexically different terms. The critical limitations across existing chunking methods are their ignorance of semantic information and the the fact that different embedding models have distinct chunk configurations [25].

III. Methods

A. Data collection

Prior studies primarily evaluated QA systems using datasets covering broad medical topics [28]–[30]. This approach may fail to capture the nuances of cancer-specific inquiries. For this project we constructed a cancer-specific datset by applying a MeSH-based filter to existing biomedical QA datasets to identify cancer-specific questions. Specifically, this study used all terms and synonyms under the 'neoplasm' MeSH subtree to identify cancer-related questions, ensuring comprehensive coverage. We applied our MeSH-based filter to

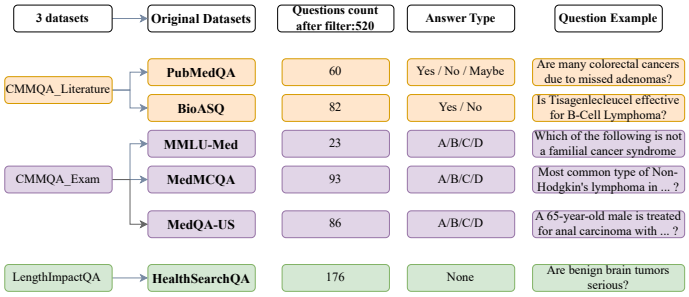

Fig. 1. Description of filtered cancer QA datasets used in this study.

six widely used medical QA datasets to create cancer-related evaluation datasets (Figure 1). For performance assessment, using multiple-choice questions simplifies evaluation, eliminates biases from text similarity computations or human annotation, and aligns with large-scale medical QA systems evaluations [28]–[30]. Therefore, we constructed the **C**ancer-related **M**ultiple-choice **M**edical **QA** Dataset (CMMQA) from

five multiple-choice medical QA datasets: PubMedQA [31] and BioASQ [2] are biomedical QA datasets with answers from biomedical research. MMLU-Med [29] is a QA subset of six biomedical tasks in Massive Multitask Language Understanding (MMLU) [32]. MedQA [33] collected questions from the US Medical Licensing Examination (USMLE). For MedMCQA, which contains questions from Indian medical entrance exams [34], only the Dev set with the ground truth was utilized. Another dataset, the "Health" dataset [35], comprising general consumer search queries, lacks definitive answers and is therefore unsuitable for accuracy assessment. However, its varied question lengths make it suitable for investigating the impact of question length on document retrieval methods.

B. Data Source and Methods of Document Retrieval

We created an LLM Rewrite Module to convert the complex, natural language questions into Boolean expressions for querying PubMed and PMC. Considering original consumer questions contain orthographic and grammatical errors and vary in length [36], [37], the LLM Rewrite Module operates in two steps: 1) query processing: the LLM employs traditional text normalization processes (like lowercase conversion and tokenization) [38] after error correction, then analyzes the processed questions to identify key concepts, relationships, and the overall intents. 2) Boolean expression generation: based on the above step, the module constructs a series of Boolean expressions from highly specific to more general formulations, allowing for flexible search via E-Utilities. For robustness, we implement a fallback strategy that iteratively applies the generated Boolean expressions, starting from the most specific and progressively moving to more relaxed versions (e.g., replacing AND with OR or removing highly specific terms while retaining core concepts) until a sufficient number of relevant documents are retrieved. This approach balances the precision of specific queries and broader searches, ensuring comprehensive results even for complex or unusual queries.

To address the semantic limitations of term-based search, we proposed Hybrid Semantic Real-time Document Retrieval (HSRDR), combining our enhanced term-based real-time search with semantic similarity-based search using the off-the-shelf MedCPT transformer model, which has been designed for zero-shot semantic retrieval of PubMed content [20]. MedCPT is trained for query-article retrieval and achieves a balance between performance and efficiency, making it ideal for semantic document retrieval. Additionally, NCBI provides pre-computed MedCPT embeddings for most PubMed abstracts, eliminating the need to download, embed, or store the corpus. These embeddings enable efficient FAISS (Facebook AI Similarity Search) index construction. Furthermore, MedCPT can return the unique PubMed identifier (PMID), which is used to download the corresponding articles.

The HSRDR is a dual-path retrieval framework (Figure 2), integrating semantic search (via MedCPT) and enhanced term-based search (via E-Utilities with LLM-rewritten queries) across three data sources: PubMed Abstracts (D1), PMC Reviews (D2), and other PMC documents (D3). After getting PMIDs of related documents through both search approaches, we use E-Utilities for downloading and temporal filtering to address temporal irrelevance problem, then parse E-Utilities results to identify PMC reviews or exclude documents without abstracts. Besides, evidence is categorized based on how it retrieved: 1) by semantic search (E1), 2) by enhanced term-based search (E2), or 3) by both methods (E3).

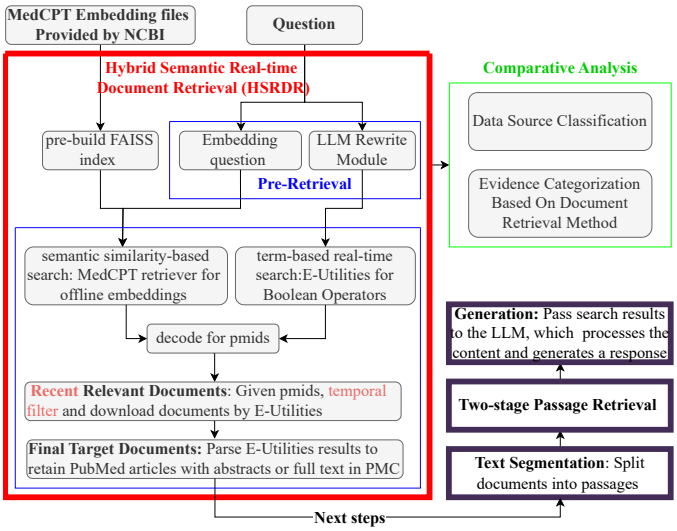

Fig. 2. The HSRDR employs dual retrieval strategies, then downloads and filters candidate documents. After document Retrieval, next steps and comparative analyses are conducted

C. Two-Stage Passage Retrieval

While MedCPT excels in document retrieval tasks, we need embedding models (dense retrievers) that excel in generating sentence-level representations to handle shorter, more specific text spans for matching query-passage pairs in the initial passage retrieval. Sentence-transformer-based models, employing pooling mechanisms to aggregate word vectors into sentence vectors for direct computation of semantic similarity between sentences, are better suited for this task because they are designed to work with smaller textual units [44]. Therefore, it is crucial to explore sentence transformer models as the dense retriever (embedding model) for passage retrieval. Additionally, the embedding model selection requires careful consideration of domain specificity, computational efficiency (model size) and retrieval accuracy. Therefore, we investigated various top-performing sentence-transformer-based embedding models (Table I) from the Massive Text Embedding Benchmark (MTEB) Leaderboard, a comprehensive benchmark that evaluates models using eight embedding tasks encompassing 58 datasets [45]. In the subsequent reranking step, we consider cross-encode models due to their ability to learn non-linear interaction patterns between two sequences (the query and the passage) via cross-entropy loss and capture complex semantic matches via jointly encoding the query and the passage with full attention mechanism (token-level cross-sequence alignment) [46]. Specifically, we evaluated two models:1) bge-reranker-v2-m3 [47] with superior performance on general reranking tasks and 2) MedCPT-reranker [20], specialized training on PubMed query-article pairs.

Embeddings	Selection Reasons	Domain
pubmedbert-base-embeddings-matryoshka (109M)	* Using sentence-transformers to fine tune BiomedBERT [39] that pretrained from scratch with PubMed and PMC literature * Achieved State-of-the-Art (SOTA) performance on many biomedical NLP tasks. * Applied Matryoshka Representation Learning [40] for dynamic embeddings sizes to save space.	Biomedical
UAE-Large-V1 (335M) [41]	* Used to be the best free model in MTEB to offer SOTA performance. * Prominent embedding models have been multi-task instruct tuned, perform comparably to commercial models like OpenAI Embedding.	General
bge-large-en-v1.5 (335M) [42]		
SFR-Embedding-Mistral (7B) [43]		

TABLE I
COMPARISON OF SELECTED EMBEDDING MODELS

Using real query-answer pairs in CMMQA for evaluation has limited coverage of edge cases and rare cancer diseases, but choosing the best embedding-reranker pair in CPQA needs comprehensive coverage for robustness. Thus, we rely on synthetic data generation to create 1000 query-evidence pairs for the sampled cancier corpus. While efficient, this approach is limited in fully representing real-world query. Consequently, we restricted its use to the experiment group in this section.

D. Precise Semantic Representation

To address the limitations of existing text splitters that ignore semantic information and embedding model requirements, we propose the **S**emantic **E**nhanced **O**verlap **S**egmentation (SEOS) algorithm, and the pseudo-code is:

Algorithm 1 SemanticNodeParser

```
1: procedure SemanticNodeParser(Document, k, embed_model)
2:     Initialize chunk_size according the embed_model
3:     Split Document into sentences (sentence lists) using sentence_splitter
4:     for i ← 0 to len(sentences) - 1 do
5:         sentence_groups[i] ← combine sentences[i] with k sentences on both sides
6:     end for
7:     For each sentence group (context information of single sentence), compute combined embeddings using embed_model
8:     distances ← Compute similarity between adjacent sentence groups
9:     cnt_breakpoint ← (len(Document.tokens) // chunk_size)
10:    potential_breakpoints ← find_SimilarityInflectionPoints(distances)
11:    final_points ← select_top_n(potential_breakpoints, cnt_breakpoint, distances)
12:    start_indices ← [bp + 1 for bp in final_points]
13:    end_indices ← [bp for bp in final_points]
14:    chunks ← [sentences[start:end] for start, end in zip(start_indices, end_indices)]
15:    chunks ← add_overlap(chunks, first_sentence_of_next_chunk)
16:    return chunks
17: end procedure
```

The key features and innovations of this method are: 1) Overlap Integration: SEOS also incorporates chunk overlap to keep chunk order information, enabling LLMs to access the context of a retrieved chunk when necessary. 2) Semantic Enhancement: SEOS improves upon the Text Tiling algorithm's boundary detection by replacing the original Bag-of-Words approach with a domain-specific transformer-based embedding model to better identify semantic relationships, effectively handling discourse relations (e.g., causal, anaphora), domain specificity, and polysemy (like matching "myocardial infarction" with "heart attack"). 3) Sentence Integrity: SEOS ensures that each chunk contains complete sentences (lines 14 and 15 in pseudo-code), which is important because sentence-transformers achieve better results with complete sentences [25]. 4) Adaptive Chunk Sizing: The algorithm adjusts chunk sizes based on the preferred chunk size of different embedding models (e.g., 512-token chunks for OpenAI models [25] and 128-word chunks with 32-word overlaps for BERT-based architectures [48]), thus overcoming limitations of fixed parameter settings in existing methods.

E. Analysis Method

In this study, we evaluate retrieval using various metrics based on the evaluation dataset type: 1) Query-Answer Paired Datasets: The **accuracy of answers generated** serves as a proxy for retrieval quality. This approach is implemented by maintaining a fixed configuration of the reader (generation) module across all experiments, ensuring that variations in answer quality can be attributed to differences in retrieval performance; 2) Query-Evidence Paired Datasets (synthetic data only used in choosing embedding-reranker combinations): The Hit Rate (Hits) measures retrieval success by quantifying the presence of relevant evidence within the top-k retrieved passages, regardless of their specific ranking. Meanwhile, Mean Reciprocal Rank (MRR), a rank-based metric, is calculated by averaging the reciprocal of the rank positions of the first relevant chunk of queries, rewarding systems that position relevant passages higher in the retrieval results.

Enriching chunk information through metadata annotation can enhance retrieval [26]. This includes annotating chunks with basic metadata (such as title or author), professional metadata (such as descriptions or keywords) or social metadata (such as rating and citation) to provide additional context [49], enabling filtered retrieval or weighted focus on crucial information [26]. Most current research leverages pre-existing data, which often falls short when experiments demand specific categorization based on comparative methodologies or data sources, necessitating the manual creation of metadata for classification and analysis. In this study, we will manually create metadata for Data Source Categorization and Evidence Categorization Based On the Document Retrieval Method. Reciprocal Rank Fusion (RRF [50]), Information Entropy [51] and **Proportions of each category in top 5 evidence** will be used for comparative analysis of the importance of each category: 1) RRF compares or aggregates results from different retrieval methods, and a higher RRF Score indicates better performance in retrieving relevant documents. 2) Information Entropy measures the diversity of retrieved documents, with lower values indicating a higher concentration of relevance. 3) Proportions of each category in the top 5 evidence reveals the importance of each category.

IV. Experimental Results and Discussion

We compared the proposed method (RAG with query pipeline optimization) against two baselines while keeping the generation module (Claude-3-haiku) fixed. **Baseline 1:** Chain-of-Thought (CoT) [52] encourages the LLM to perform step-by-step reasoning to improve the quality of the generated answers. Baseline 1 uses the standalone LLM under the COT setting. **Baseline 2:** Baseline 2 uses the whole text of the Top 5 relevant documents retrieved by the MedCPT to directly guide LLM in generating an answer with COT prompting.

To optimize experimental time and resources, we employed the concept of hard negatives [53], constructing a test dataset comprised of questions that both Baseline 1 and Baseline 2 methods incorrectly answered, to observe the impact of different methods and corpora on the query pipeline.

A. Data Source and Methods of Document Retrieval

1) Hybrid Semantic Real-time Document Retrieval (HSRDR): Hybrid Semantic Real-time Document Retrieval (HSRDR) combines real-time search (via E-Utilities with rewritten queries) with semantic search (via MedCPT). The following HSRDR result analyses are shown in Table II:

	Metrics	E1	E2	E3
HealthSearchQA	RRF Score	20.848	6.202	1.358
	Information Entropy	0.716	0.539	0.469
	Proportion in Top 5	0.736	0.218	0.045
Negative Cancer QA	RRF Score	19.473	11.319	1.490
	Information Entropy	1.488	1.162	0.440
	Proportion in Top 5	0.62	0.336	0.044

TABLE II
COMPARATIVE ANALYSIS BASED ON EVIDENCE CATEGORIZATION: RETRIEVED BY MEDCPT (E1), RETRIEVED BY E-UTILITIES WITH REWRITTEN QUERIES (E2), OVERLAP EVIDENCE (E3).

MedCPT Dominance and Query Length Sensitivity: Semantic search based on MedCPT (dense retriever) consistently outperforms E-Utilities-based search, with a more pronounced advantage in HealthSearchQA (RRF Score: 20.85 vs. 6.20, Proportion in top 5: 73.6% vs. 21.8%) than in Negative Cancer QA (RRF Score: 19.47 vs. 11.32, Proportion in top 5: 62% vs 33.6%). This disparity is attributed to the impact of average query length (HealthSearchQA: 6.9 tokens, Negative Cancer QA: 38 tokens) on MedCPT: 1) Information Density Variability: Longer queries, constrained by the same vector dimensionality, risk losing critical details, diluting the information density. 2) Normalization Impact: The cosine similarity includes a normalization step, and longer queries might lead to vectors with higher denominators. This highlights the importance of text-length robustness for retrievers [54] and considering text length with similarity score at the same time.

Complementary Value of E2: The non-negligible RRF scores and proportions of E2 show the important complementary value of E-Utilities-based online search in document retrieval. This is partly because: 1) While MedCPT relies on static embedding files, E-Utilities-based search can access real-time data, thereby covering newer publications. 2) E-Utilities-based search provides fine-tuned control through Boolean operators, suitable in contexts requiring exact matches.

The Overlap Analysis : The low entropy values of E3 for overlapping documents suggest that when both methods agree on a document's relevance, it is highly pertinent to the query. However, the small overall overlap, reflected by low RRF scores and proportion (4.5% and 4.4%), indicates that the two methods focus on different subsets of relevant documents.

In conclusion, the E-Utilities-based and the MedCPT-based method serve as complementary document retrievers. Based on experimental results we have seen, using HSRDR to search PubMed abstracts and PMC reviews is the best method, balancing information breadth, depth, and real-time data access.

2) Comparative Analysis of Online NCBI Sources: Figure 3 reveals the comparative retrieval value of each NCBI source.

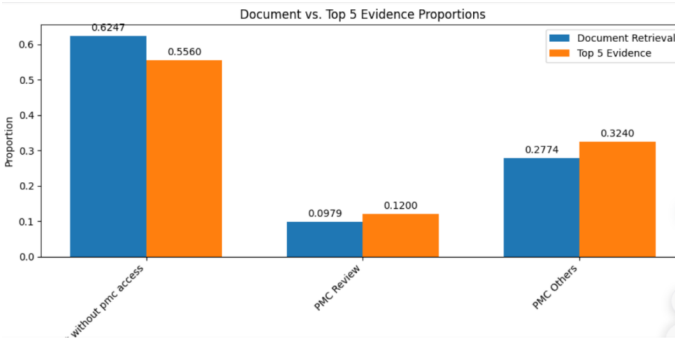

Fig. 3. Distribution comparison between Initial Document Pool and Top-5 Retrieved Evidence when HSRDR's Retrieval Source involving PubMed Abstract, PMC Reviews and PMC Others

PubMed Abstracts Dominance and Decline in Evidence: PubMed abstracts comprise a substantial portion of the top-5 evidence, likely due to their wider coverage than PMC (23.9M citations with valid abstract vs. 8M free full-text articles), suggesting that PubMed abstracts are the most dominant retrieval source for CPQA. However, their representation decreases in final evidence selection while PMC sources show increased representation, potentially because abstracts lack details and PMC literature full-text availability provides richer information, suggesting the value of complementary data sources.

PMC reviews versus PMC non-reviews: PMC reviews relative representation to PMC others increases, indicating that review articles may have higher retrieval value for question answering than standard scientific articles. Table III shows

Data source	Negative Cancer QA Dataset			
	Accuracy	Precision	Recall	F1-score
D1	44.00	41.12	40.12	39.14
D1+D2	46.00	43.78	38.13	38.77
D1+D2+D3	46.00	40.12	33.86	35.33

TABLE III
DATA SOURCES' IMPACT ON RETRIEVAL PERFORMANCE. D1: PUBMED ABSTRACTS; D2: PMC REVIEWS; D3: OTHER PMC ARTICLES.

that question answering based only on PubMed abstracts and PMC reviews outperforms question answering where all PMC artciles are included. The results may be because that PMC reviews, which integrate findings across multiple studies in their full text to offer comprehensive analyses and cross-study insights, better match broad queries than PMC others, which are more context-specific and can be represented by abstracts.

B. Two-Stage Passage Retrieval

We implemented retrieval evaluation based on LlamaIndex [26], a framework for building RAG systems. Specially, we used E-Utilities to query "cancer" to retrieve PubMed abstracts and PMC full-text articles as the text corpus for building an evaluation dataset, then used LLMs to generate pairs (query, context) from each chunk of the prepared text corpus, ensuring this evaluation was suitable for all data sources. In the experiment, we evaluated retrieval performance using Hits@5 and MRR@5, which aligns with the practical constraints of RAG systems, where the limited context window of the LLM generator requires a focus on retrieving the most relevant chunks [13]. Results are shown in Figure 4.

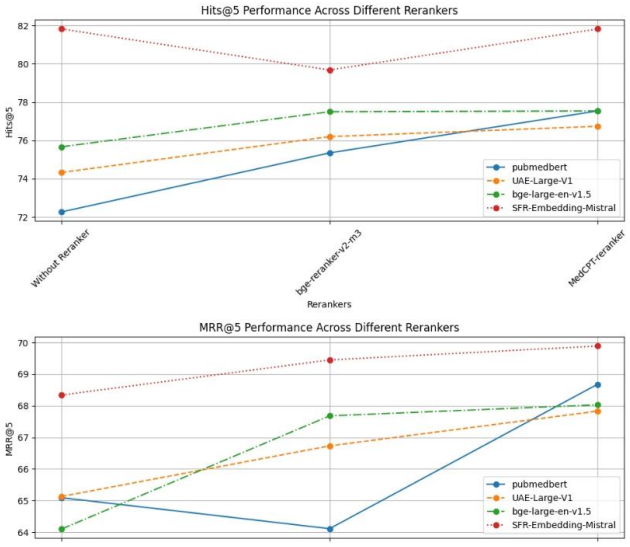

Fig. 4. Performance of Embedding Models with rerankers

Domain-Specific Feature is Crucial: Pubmedbert-matryoshka , despite its smaller size and absence from the MTEB leaderboard, achieved the second-best performance when paired with the MedCPT-reranker. This suggests that the size of the embedding model is not the only determinant of effectiveness and that domain-specific fine-tuning or training can significantly improve performance by leveraging domain-specific understanding [2]. The importance of domain-specific features is also demonstrated by the MedCPT-reranker outperforming the general domain reranker in enhancing retrieval relevance across all embedding models.

Reranker Impact on Embedding Models: PubMedBERT-Matryoshka performed poorly without a reranker but benefited substantially from the MedCPT Reranker. This synergy likely stems from the MedCPT Reranker being trained on negative samples from the MedCPT retriever (derived from Pubmedbert), aligning the reranker more effectively with PubMedBERT's embedding space and enabling it to capture complementary information. Meanwhile, the BGE Reranker enhances the performance of BGE Embedding in terms of hits, also suggesting the importance of compatibility and complementarity between the embedding model and the reranker. However, the observed performance decline when BGE Reranker is paired with incompatible embeddings highlights the risks of mismatched reranker-embedding combinations. If a reranker cannot align with the embedding space or provide complementary semantic insights, it can fail to capture semantic nuances or introduce noise, leading to performance degradation. In conclusion, while rerankers can enhance retrieval, selecting compatible reranker-embedding combinations is crucial.

C. Precise Semantic Representation

Text Splitter	Retriever		
	Pubmedbert-matryoshka	BM25	MedCPT
512Overlap0	46	20	22
512Overlap32	52	18	24
512Overlap128	42	16	22
SEOS method	**54**	**36**	**38**

TABLE IV
ACCURACY (%) ACROSS DIFFERENT TEXT SPLITTER AND RETRIEVER COMBINATIONS ON NEGATIVE CANCER QA DATASET

This study focused on three single retrievers (the BM25 retriever, the MedCPT retriever, and the Pubmedbert-matryoshka in the previous section) paired with the MedCPT reranker.

Table IV demonstrates the SEOS method's superiority in text segmentation, outperforming fixed-parameters strategies. The SEOS method excels by preserving natural and meaningful text boundaries based on semantic similarity and its variations, which enhances the retrievers' ability to locate relevant information. Its advantages also include sentence overlap and automatic chunk-size adjustment tailored to the embedding model. Both Pubmedbert-matryoshka and MedCPT benefited from automatic chunk-size adjustment. The corresponding chunk-size adjustment rule obtained from research indicates that 128-word chunks with 32-word overlaps optimize BERT-based models for QA tasks [48]. This finding can also be shown by the 512Overlap32's top performance among fixed-length strategies. Despite a 512 chunk size, the actual retrieval text space is roughly 128, with metadata integrated consuming the remainder of the chunk, typically around 384 tokens.

Notably, MedCPT, which does not utilize a sentence-transformer structure and is trained on PubMed query-article pairs, is better suited for query-document similarity and document retrieval. For finer-grained passage retrieval, the optimal combination of dense retriever and reranker is recommended to achieve the nuanced understanding required.

D. Comparison

To assess the effectiveness of our query pipeline optimization to enhance the zero-shot capability of LLMs in answering cancer questions, we conducted experiments with Claude-3-Haiku using our method (RAG with query pipeline optimization) and two baselines, as shown in Table V.

Both Baseline 2 and our method incorporate external relevant knowledge via retrieval components, which improves answer accuracy across most QA datasets over Baseline 1, a retrieval-free approach that relies solely on the LLM's parametric knowledge. Our method outperforms Baseline 2 by

Method	CMMQA					Avg
	MMLU	MedQA	MedMCQA	PubMedQA	BioASQ	
Baseline 1	78.26	68.60	65.59	45.00	80.49	67.15
Baseline 2	82.61	67.44	65.59	56.67	81.71	69.48
Our method	82.61	**69.77**	68.82	**65.00**	**81.71**	**72.39**

TABLE V
COMPARISON OF PERFORMANCE ON CMMQA DATASETS FOR THREE METHODS: BASELINE 1 (LLM + COT), BASELINE 2 (NAIVE RAG), AND THE PROPOSED METHOD (RAG WITH QUERY PIPELINE OPTIMIZATION - HSRDR FOR DOCUMENT RETRIEVAL, SEOS METHOD AS THE TEXT SPLITTER, PUBMEDBERT-MATRYOSHKA RETRIEVER AND MEDCPT RERANK FOR TWO-STAGE PASSAGE RETRIEVAL)

about 3% in answer accuracy. This performance gap demonstrates the effectiveness of our query pipeline optimization, which includes two key enhancements: First, by integrating the MedCPT retriever and E-Utilities, HSRDR expands the retrieval scope and improves document retrieval efficiency. This method addresses the limitations of Baseline 2, which solely relies on static PubMed abstract data stored in MedCPT article JSON files and lacks real-time, term-based search. Second, SEOS and two-stage passage retrieval ensure that Claude-3-haiku receives only the most relevant, semantically segmented passages from source documents. This enhancement reduces the "distraction phenomenon" observed in Baseline 2, where all search results were passed to the model, diluting its focus.

Our method shows a 5% improvement over Baseline 1. However, this advantage may seem inconsistent when evaluating the negative QA dataset. Two factors contribute to this discrepancy: First, **Dataset Construction Bias**: Questions uniquely answered incorrectly by Baseline 1 or Baseline 2 were excluded, limiting its ability to reflect broader performance differences. Second, **Inherent Limitations of RAG-Based Approaches**: RAG does not consistently enhance responses. Sometimes, questions that Baseline 1 answered correctly are answered incorrectly under RAG, as observed prominently in the MedMCQA dataset. Prior studies [55], [56] have reported that RAG can negatively impact the original outcome, particularly when LLM's inherent parametric knowledge suffices for the query [55]. This occurs because LLMs tend to rely on retrieval results, impairing creativity and versatility, even accuracy [56].

V. CONCLUSION

This study conducted query pipeline optimization for CPQA by developing a novel document retrieval method (HSRDR), exploring optimal pairings of dense retrievers and rerankers for passage retrieval, and integrating the proposed chunk technique (SEOS). SEOS can be used in more domains beyond healthcare, while HSRDR is tailored for broad biomedical applications using PubMed and PMC. Document retrieval evaluations show the varying effectiveness of distinct data sources for CPQA and the complementary value of semantic search and term-based online retrieval. Document retrieval evaluations show the varying retrieval effectiveness of distinct data sources for CPQA and the complementary value of semantic search and term-based online retrieval. Passage retrieval experiments emphasize the importance of domain-special models and the reranker-embedding alignment. Limitations include: 1) We utilize Boolean operator conversion to enhance term-based online search, whether other query refinement methods (like Query2Doc [57] and RAG-Fusion [58]) can further improve performance needs to be explored. 2) We use multiple-choice-style evaluation, which may overlook nuanced details and is susceptible to guessing bias, where random guesses will boost accuracy. 3) We find that RAG does not always improve performance, thus adaptive retrieval mechanisms should be implemented to select retrieval strategies based on the need for external information for each query.

REFERENCES

[1] E. Landhuis, "Scientific literature: Information overload," *Nature*, vol. 535, no. 7612, pp. 457–458, 2016.

[2] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 6233–6251. [Online]. Available: https://aclanthology.org/2024.findings-acl.372

[3] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[4] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards mitigating llm hallucination via self reflection," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 1827–1843.

[5] L. K. Umapathi, A. Pal, and M. Sankarasubbu, "Med-halt: Medical domain hallucination test for large language models," *arXiv preprint arXiv:2307.15343*, 2023.

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[7] M. Kang, S. Lee, J. Baek, K. Kawaguchi, and S. J. Hwang, "Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[8] D. Truhn, J. S. Reis-Filho, and J. N. Kather, "Large language models should be used as scientific reasoning engines, not knowledge databases," *Nature medicine*, vol. 29, no. 12, pp. 2983–2984, 2023.

[9] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling large language models to generate text with citations," *arXiv preprint arXiv:2305.14627*, 2023.

[10] T. Merth, Q. Fu, M. Rastegari, and M. Najibi, "Superposition prompting: Improving and accelerating retrieval-augmented generation," in *Forty-first International Conference on Machine Learning*.

[11] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, and C. D. Manning, "Raptor: Recursive abstractive processing for tree-organized retrieval," in *The Twelfth International Conference on Learning Representations*, 2023.

[12] D. Arora, A. Kini, S. R. Chowdhury, N. Natarajan, G. Sinha, and A. Sharma, "Gar-meets-rag paradigm for zero-shot information retrieval," *arXiv preprint arXiv:2310.20158*, 2023.

[13] Y. Tang and Y. Yang, "Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries," *arXiv e-prints*, pp. arXiv–2401, 2024.

[14] O. Yoran, T. Wolfson, O. Ram, and J. Berant, "Making retrieval-augmented language models robust to irrelevant context," in *The Twelfth International Conference on Learning Representations*, 2023.

[15] J. Kans, "Entrez direct: E-utilities on the unix command line," in *Entrez programming utilities help [Internet]*. National Center for Biotechnology Information (US), 2024.

[16] Q. Jin, R. Leaman, and Z. Lu, "Pubmed and beyond: biomedical literature search in the age of artificial intelligence," *Ebiomedicine*, vol. 100, 2024.

[17] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, "Pmc-llama: toward building open-source language models for medicine," *Journal of the American Medical Informatics Association*, p. ocae045, 2024.

[18] G. Frisoni, M. Mizutani, G. Moro, and L. Valgimigli, "Bioreader: a retrieval-enhanced text-to-text transformer for biomedical literature," in *Proceedings of the 2022 conference on empirical methods in natural language processing*, 2022, pp. 5770–5793.

[19] J. Vladika and F. Matthes, "Improving health question answering with reliable and time-aware evidence retrieval," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 4752–4763.

[20] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, and Z. Lu, "Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval," *Bioinformatics*, vol. 39, no. 11, p. btad651, 2023.

[21] J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, and X. Cheng, "Semantic models for the first-stage retrieval: A comprehensive review," *ACM Transactions on Information Systems (TOIS)*, vol. 40, no. 4, pp. 1–42, 2022.

[22] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[23] Y. Zhou, T. Shen, X. Geng, C. Tao, C. Xu, G. Long, B. Jiao, and D. Jiang, "Towards robust ranker for text retrieval," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 5387–5401.

[24] R. Nogueira, W. Yang, K. Cho, and J. Lin, "Multi-stage document ranking with bert," *arXiv preprint arXiv:1910.14424*, 2019.

[25] "Retrieval-augmented generation for large language models: A survey."

[26] J. Liu, "Llamaindex," *Acceso el*, vol. 6, 2022.

[27] M. A. Hearst, "Text tiling: Segmenting text into multi-paragraph subtopic passages," *Computational linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[28] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu *et al.*, "Can generalist foundation models outcompete special-purpose tuning? case study in medicine," *Medicine*, vol. 84, no. 88.3, pp. 77–3, 2023.

[29] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[30] V. Liévin, C. E. Hother, A. G. Motzfeldt, and O. Winther, "Can large language models reason about medical questions?" *Patterns*, vol. 5, no. 3, 2024.

[31] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 2567–2577.

[32] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," in *International Conference on Learning Representations*, 2020.

[33] D. Jin, E. Pan, N. Oufattole, W. Wei-Hung, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.

[34] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering," in *Conference on health, inference, and learning*. PMLR, 2022, pp. 248–260.

[35] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal *et al.*, "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, 2023.

[36] C. J. Lu, A. R. Aronson, S. E. Shooshan, and D. Demner-Fushman, "Spell checker for consumer language (cspell)," *Journal of the American Medical Informatics Association*, vol. 26, no. 3, pp. 211–218, 2019.

[37] A. B. Abacha, Y. Mrabet, M. Sharp, T. R. Goodwin, S. E. Shooshan, and D. Demner-Fushman, "Bridging the gap between consumers' medication questions and trusted answers," in *MEDINFO 2019: Health and Wellbeing e-Networks for All*. IOS Press, 2019, pp. 25–29.

[38] D. MANNING, "Introduction to information retrieval," *Journal of the American Statistical Association*, vol. 15, 2008.

[39] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," 2020.

[40] A. Kusupati, G. Bhatt, A. Rege, M. Wallingford, A. Sinha, V. Ramanujan, W. Howard-Snyder, K. Chen, S. Kakade, P. Jain *et al.*, "Matryoshka representation learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 30 233–30 249, 2022.

[41] X. Li and J. Li, "Angle-optimized text embeddings," *arXiv preprint arXiv:2309.12871*, 2023.

[42] S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff, "C-pack: Packaged resources to advance general chinese embedding," 2023.

[43] S. R. J. C. X. Y. Z. S. Y. Rui Meng, Ye Liu, "Sfr-embedding-mistral:enhance text retrieval with transfer learning," Salesforce AI Research Blog, 2024. [Online]. Available: https://blog.salesforceairesearch.com/sfr-embedded-mistral/

[44] X. Yang, X. He, H. Zhang, Y. Ma, J. Bian, Y. Wu *et al.*, "Measurement of semantic textual similarity in clinical texts: comparison of transformer-based models," *JMIR medical informatics*, vol. 8, no. 11, p. e19735, 2020.

[45] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "Mteb: Massive text embedding benchmark," *arXiv preprint arXiv:2210.07316*, 2022. [Online]. Available: https://arxiv.org/abs/2210.07316

[46] F. Jiang, Q. Xu, T. Drummond, and T. Cohn, "Boot and switch: Alternating distillation for zero-shot dense retrieval," in *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[47] C. Li, Z. Liu, S. Xiao, and Y. Shao, "Making large language models a better foundation for dense retrieval," 2023.

[48] Z. Wang, P. Ng, X. Ma, R. Nallapati, and B. Xiang, "Multi-passage bert: A globally normalized bert model for open-domain question answering," *arXiv preprint arXiv:1908.08167*, 2019.

[49] I. Ullah, S. Khusro, and I. Ahmad, "Improving social book search using structure semantics, bibliographic descriptions and social metadata," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5131–5172, 2021.

[50] G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 758–759.

[51] J. Liang, Z. Shi, D. Li, and M. J. Wierman, "Information entropy, rough entropy and knowledge granulation in incomplete information systems," *International Journal of general systems*, vol. 35, no. 6, pp. 641–654, 2006.

[52] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[53] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, and A. Overwijk, "Approximate nearest neighbor negative contrastive learning for dense text retrieval," *arXiv preprint arXiv:2007.00808*, 2020.

[54] W. L. Tam, X. Liu, K. Ji, L. Xue, X. Zhang, Y. Dong, J. Liu, M. Hu, and J. Tang, "Parameter-efficient prompt tuning makes generalized and calibrated neural text retrievers," *arXiv preprint arXiv:2207.07087*, 2022.

[55] Y. Wang, P. Li, M. Sun, and Y. Liu, "Self-knowledge guided retrieval augmentation for large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 10 303–10 315.

[56] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-rag: Learning to retrieve, generate, and critique through self-reflection," in *The Twelfth International Conference on Learning Representations*, 2023.

[57] L. Wang, N. Yang, and F. Wei, "Query2doc: Query expansion with large language models," *arXiv preprint arXiv:2303.07678*, 2023.

[58] Z. Rackauckas, "Rag-fusion: a new take on retrieval-augmented generation," *arXiv preprint arXiv:2402.03367*, 2024.