Towards semantic reliable clinical QA: Query pipeline optimization for cancer patient question answering systems

Anonymous ACL submission

Abstract

Large Language Models (LLMs) show promise in medical Question-Answering (QA) but suffer from hallucinations that jeopardize patient safety. While Retrieval-Augmented Generation (RAG) mitigates this by grounding outputs in external evidence, existing pipelines struggle with the complex, rapidly evolving nature of oncology. We present **CoMeta**, a three-level controllable metadata-aware framework optimized for Cancer Patient QA (CPQA). We introduce Clinical Hybrid Semantic-Symbolic Document Retrieval (CHSDR), which synergizes real-time Boolean search via NCBI E-Utilities with semantic retrieval to overcome metadata blindness. Additionally, we propose **S**emantic **E**nhanced **O**verlap **S**egmentation (SEOS) to prevent contextual fragmentation. Our results demonstrate that CHSDR significantly improves retrieval performance, CoMeta 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.

1 Introduction

Large Language Models (LLMs) have shifted the paradigm of online information seeking from traditional search engines to conversational agents. While promising for Question-Answering (QA), their reliance on internal parameters renders them prone to hallucination—generating fluent but factually incorrect outputs (Ji et al., 2023). This issue is acute in the **clinical domain**, where accuracy directly impacts patient safety (Umapathi et al., 2023). To mitigate this, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) integrates external evidence to ground LLM responses. Especially in medical QA tasks where queries are knowledge intensive, LLMs excel as generators rather than knowledge databases (Truhn et al., 2023).

To enhance medical QA, prior studies (Jeong et al., 2024; Liang et al., 2025) broadly use RAG with domain-specific adaptation of retrieval strategies. However, they overlook the query pipeline, the foundational mechanism for query-to-evidence mapping that underpins all retrieval strategies. In Cancer Patient QA (CPQA), these pipelines cause systematic retrieval failures. First, **retrieval reliability** face the Staleness-Semantic Dilemma: the rapid evolution of oncology (Landhuis, 2016) means that standard query pipelines (e.g., Dense or BM25), built upon static and metadata-blind indexes, risk surfacing outdated and thus clinically unreliable evidence, while real-time, metadata-aware interfaces like E-Utilities (Kans, 2024) are semantically brittle to informal patient queries. Second, a **retrieval-depth paradox** stems from the asymmetric utility of publication types. We argue that reviews require full-text retrieval to capture high-level therapeutic syntheses (Wang et al., 2025), while primary research articles often benefit from abstract-only retrieval to avoid methodological noise. Most pipelines apply uniform retrieval depth across article types, either starving the model of synthesis or overwhelming it with clutter. Third, **semantic representation** is undermined by contextual fragmentation: prior encoder-agnostic segmentation (pre-defined length (Liu, 2022) or lexical (Hearst, 1997)) severs clinical qualifiers (e.g., specific mutation criteria) from therapeutic statements, yielding recommendations that appear evidence-based yet lack essential medical constraints.

We address these failures with a specialized framework that enhance controllability across three pillars: 1) Retrieval Robustness, ensuring stable performance across expert queries and informal patient narratives; 2) Evidence Metadata-awareness, enabling time filtering and adaptive retrieval depth based on publication types; 3) Relational Integrity,

utilizing encoder-aware segmentation to safeguard clinical logic. To support this architecture, we curate the Cancer-related Medical QA (CMMQA) dataset and its "Clinical Narrative Variant" to simulate real-world patient interactions [1]. Our specific contributions are:

- We propose the Clinical Hybrid Semantic-Symbolic Document Retrieval (CHSDR) method. To our knowledge, this is the first pipeline in the clinical RAG domain to combine real-time Boolean search (via E-Utilities with LLM-rewritten queries) with MedCPT (Jin et al., 2023) semantic search.
- We introduce **S**emantic **E**nhanced **O**verlap **S**egmentation (SEOS), a novel text segmentation method that integrates sentence semantics and embedding model characteristics while utilizing chunk overlap to preserve critical context.
- We provide the first comparative analysis of NCBI sources for CPQA, demonstrating that PMC review articles possess significantly higher retrieval value than non-review PMC papers among clinical cancer QA datasets.

2 Related Work

2.1 Retrieval in Medical RAG

Medical RAG systems rely on their query pipelines, which map user queries to evidence. The predominant paradigm is dense retrieval, leveraging domain-specific embedding models for vector-similarity search in offline index (Miao et al., 2025). Recent advances have shifted from general-purpose embeddings to domain-specific retrievers like MedCPT (Jin et al., 2023). Advanced retrieval strategies, including hybrid search that integrates sparse methods like BM25 (Robertson et al., 2009) for lexical-semantic complementarity (Xiong et al., 2024), adaptive retrieval that dynamically selects retrieval timing (Jeong et al., 2024) and recursive search that iteratively refine queries through feedback loops (Liang et al., 2025), are essentially optimizations built upon static-index pipelines.

The reliance on offline indexing restricts metadata filtering (e.g., by publication date or study type), incurs substantial preprocessing and storage overhead (e.g., >400GB for PubMed indices (Jeong et al., 2024)), and introduces timeliness gaps where newly published studies may remain inaccessible in critical decision windows. These issues are acute in oncology, where clinical standards evolve rapidly. NCBI's E-Utilities offers a potential remedy through real-time access and metadata filtering, but its utility is hindered by a term-centric matching paradigm. Unlike BM25, E-Utilities lacks mechanisms like length normalization or term weighting, making retrieval highly sensitive to query formulation and ill-suited for natural language inputs without significant adaptation. Thus, existing systems are implicitly forced to trade off semantic robustness against retrieval controllability.

2.2 Semantic Representation of Clinical Knowledge

Dense retrieval efficacy hinges on precise semantic representation, which is governed by two coupled factors: document segmentation and text encoding (gao). While recent medical RAG systems extensively optimize retriever configuration (Xiong et al., 2024; Tang and Yang, 2024), document segmentation, the foundational step determining input granularity, remains critically under-explored.

Current approaches largely rely on heuristic strategies that compromise clinical context. Naive fixed-length chunking (Jeong et al., 2024) frequently truncates sentences mid-thought, severing syntactic dependencies. Even sentence-aware methods like LlamaIndex's SentenceSplitter (Liu, 2022), while preserving sentence integrity, impose rigid window sizes that fail to adapt to the variable information density of medical narratives. Similarly, TextTiling (Hearst, 1997) infers topic boundaries via lexical overlap, but its surface-level matching struggles with the high synonymy and complex semantic shifts inherent in biomedical literature.

The fundamental limitation of these approaches is twofold. First, they function as "semantic-blind" preprocessing steps, risking the separation of a therapeutic recommendation from its prerequisite genomic markers or contraindications. Second, they neglect the interaction between chunk size and encoder performance, ignoring evidence that optimal segmentation is highly dependent on the specific embedding model used (gao).

[1] All codes and datsets can be found via: anonymous.4open.science/r/COMP90005-E51E/README.md.

3 Methods

3.1 Dataset and Evaluation

We applied a MeSH-based filter to HealthSearchQA (Singhal et al., 2023b) and the MIRAGE benchmark (Xiong et al., 2024) that includes PubMedQA (Jin et al., 2019), BioASQ (Tsatsaronis et al., 2015), MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), the medical subsets in MMLU (Hendrycks et al., 2020). Specifically, we use all terms and synonyms under the 'neoplasm' MeSH subtree to identify questions strictly pertaining to cancer as a disease entity. While not exhaustive, it ensures a focus on core disease-related queries. We refer to this filtered collection as the Cancer-related Medical QA Dataset (CMMQA) (Figure 1). Then, we use LLama-3-70b to rewrite the CMMQA queries into non-expert clinical narratives (mainly simulating patients), creating a derived dataset we term the "Clinical Narrative Variant". This variant serves as a challenging testbed to evaluate both retrieval stability and QA accuracy across distinct query modalities (standard vs. narrative).

Datasets	Original Datasets	Questions count after filter:520	Answer Type
CMMQA-Literature	PubMedQA	60	Yes / No / Maybe
	BioASQ	82	Yes / No
CMMQA-Exam	MMLU-Med	23	A/B/C/D
	MedMCQA	93	A/B/C/D
	MedQA-US	86	A/B/C/D
Length-Impact-QA	HealthSearchQA	176	None

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

Our *document* retrieval evaluation focuses on PubMedQA and BioASQ because these two datasets provide gold-standard citations (PMIDs) but others lack source annotations. For *passage* retrieval evaluation, we utilized synthetic Query-Answer pairs generated from PubMed Abstracts, PMC Full Texts and medical textbooks (Jin et al., 2021) as a validation set. Retrieval Performance is measured at a cutoff of K=10 using three standard metrics: Hit Rate (Hit@K), representing the percentage of queries with at least one relevant document retrieved; Mean Reciprocal Rank (MRR@K), assessing the ranking quality of the first relevant document; and Recall (Recall@K), quantifying the coverage of gold-standard evidence. Besides, we report Hit0, the count of queries yielding zero results. This metric serves as a critical indicator of retrieval robustness, highlighting system failures in handling complex or restrictive queries.

3.2 Controllable Query Pipeline Design

To ensure clinical rigor across the retrieval lifecycle, we design a hierarchical query pipeline (Figure 2) that prioritizes controllability at both the document and passage levels. Specific configurations of the underlying Corpora, Retrievers, and LLMs are detailed in Appendix.

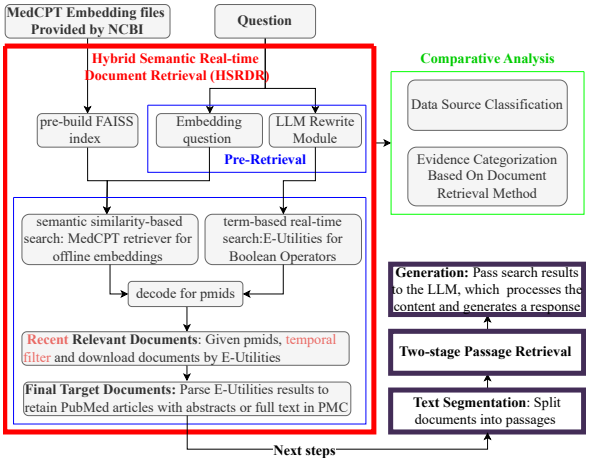

Figure 2: The CHSDR employs dual retrieval strategies, then downloads and filters candidate documents. After document Retrieval, next steps and comparative analyses are conducted

3.2.1 Clinical Hybrid Semantic-Symbolic Document Retrieval (CHSDR)

CHSDR (stage-1 in figure 2) aims to preserve metadata control while ensuring robust document-level retrieval across various queries (standard vs. clinical narrative).

Boolean-constrained retrieval with adaptive query execution Raw consumer questions contain orthographic and grammatical errors and vary in length (Lu et al., 2019; Abacha et al., 2019), which are incompatible with E-Utilities. CHSDR thus use a LLM-based rewriter that performs not only query error correction and text normalization (Lu et al., 2019), but also patient intent analysis, clinical abstraction (i.e.,mapping narrative descriptions into standardized medical concepts and PICO elements), boolean expression and temporal constraint generation. To ensure robustness, we design a *adaptive query execution*: generated outputs are executed progressively from strict boolean conjunctions to clinical abstraction of raw queries to

relaxed boolean formulations, until sufficient documents are retrieved.

Hybrid semantic–symbolic retrieval with metadata-aware document control We integrate symbolic search with semantic retrieval via Reciprocal Rank Fusion (RRF) (Cormack et al., 2009), further overcoming the term-centric limitations of E-Utilities. We adopt MedCPT (Jin et al., 2023) for semantic document retrieval, because it is trained for query-article retrieval and achieves a balance between performance and efficiency. Both retrieval streams return PubMed identifiers (PMIDs), which serve as a unified document key. All candidate documents and their associated metadata are subsequently accessed via E-Utilities. In this study, we use metadata to support further temporal filtering, conflict resolution based on recency, and filter records with valid abstract. We additionally tag each document by its retrieval provenance and publication types. The first captures how a document is retrieved: semantic retrieval only (E1), symbolic retrieval only (E2), or both (E3), enable comparative analysis of retrieval methods. The second captures what the document is: PubMed abstracts (D1), PMC review articles (D2), and other PMC full-text articles (D3), enable empirical analysis of retrieval depth across publication types.

.

3.2.2 Passage-Level Retrieval and Semantic Representation

Semantic Enhanced Overlap Segmentation (SEOS) To address the twofold limitation in section 2.2, we propose SEOS (pseudo-code in Appendix). SEOS is inspired by TextTiling (Hearst, 1997), which identifies segmentation boundaries via local cohesion shifts, but extends it through following key innovations. First, We replace bag-of-words with domain-specific dense embeddings in text representation, enabling robust handling of medical terminology and discourse relations in clinical studies. Second, instead of brittle similarity thresholds, SEOS determines the optimal partition cardinality (N) from a target token budget, selecting the top-N semantic minima as breakpoints.The budget is adaptively configured to align with the effective input granularity observed for different embedding models, as prior work has shown retrieval performance to be sensitive to chunk length (gao; Wang et al., 2019). Third, to mitigate contextual fragmentation, SEOS incorporates adaptive sentence overlap determined based on semantic continuity around each breakpoint, preserving unresolved semantic dependencies. Furthermore, neighboring chunk identifiers are explicitly stored, enabling downstream module to recover local context when evidence spans multiple segments.

Dense retrieval and reranking for passage matching The training objective of MedCPT is optimized for matching queries to PubMed abstract, but our passage retrieval corpus mainly is PMC full text that exhibit different discourse structures and information densities. This distributional shift from abstracts to full-text passages necessitates encoders with broader generalization capabilities. Therefore, we select top-performing embedding models from the MTEB benchmark (Muennighoff et al., 2022). For the subsequent reranking stage, we employ cross-encoder architectures to capture the fine-grained, non-linear semantic interactions that bi-encoders often miss (Jiang et al.). To evaluate the trade-off between domain specialization and general semantic reasoning, we only select representative ones: 1) bge-reranker-v2-m3 (Li et al., 2023), a general-purpose model with superior performance; and 2) MedCPT-reranker (Jin et al., 2023), a domain-specific reranker pre-trained on large-scale biomedical literature.

4 Results

Claude-3-haiku is selected as the LLM backbone for CPQA. We compare our work with two settings: 1) LLM with COT. Chain-of-Thought (CoT) (Wei et al., 2022) encourages the LLM to perform step-by-step reasoning to improve the quality of the generated answers. 2) LLM with naive RAG, using the whole text of the Top 5 relevant documents retrieved by the MedCPT to directly guide LLM to generate an answer.

4.1 Comparison in Document retrieval

In Table 1, we conduct the ablation study of Clinical Hybrid Semantic-Symbolic Document Retrieval (CHSDR) on retrieval performance. Results indicate: 1. the "Zero-Hit" Barrier. Across most datasets, the limitations of sparse search are evident in the **E-utils** baseline. In contrast, **Adapt** method (LLM Rewriter + Adaptive Fallback) significantly improve retrieval performance across all datasets, particularly in the *Clinical Narrative* setting where it help the system to recover from the baseline's

total failure (high Hit0). This demonstrates it effectively acting as a crucial bridge between patient queries and biomedical literature. 2. the Hybrid Dilemma. The **Hybrid** model achieves best performance on BioASQ, outperforming **Adapt** method. This indicates that semantic retrieval (MedCPT) successfully recalls relevant documents that are not retrieved by sparse retriever. However, on PubMedQA, the Hybrid approach slightly underperforms the symbolic-only Adapt method in the *Standard Narrative*. We attribute this anomaly to query specificity: PubMedQA questions are often derived directly from abstract sentences, creating a bias toward exact lexical matching. 3. Impacts of Clinical Narrative. While shifting to *Clinical Narrative* inputs universally degrades performance, the hybrid model maintains robustness. This suggests that when the input text is noisy, the semantic understanding provided by the Hybrid system becomes the critical mechanism for retrieval resilience.

4.1.1 Ablation Study on Evidence Sources and Depth

Figure 3 reveals the comparative retrieval value of each NCBI source.

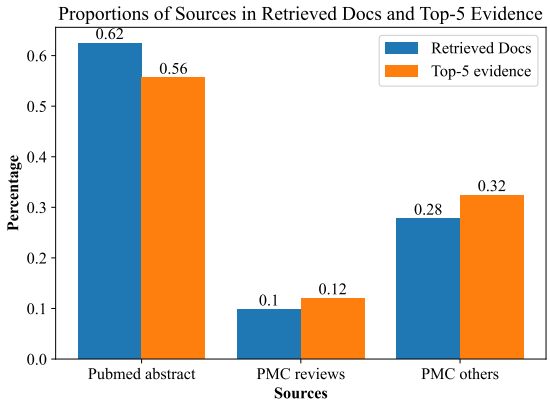

Figure 3: Distribution comparison between Initial Document Pool and Top-5 Retrieved Evidence when CHSDR's Retrieval Source involving PubMed Abstract, PMC Reviews and PMC Others

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 2 shows that question answering based only on PubMed abstracts and PMC reviews outperforms question answering where all PMC articles 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.

4.2 Analysis in Passage Retrieval

4.2.1 Comparison of Text chunking Strategies

Table 3 evaluates the robustness of SEOS across passage retriever, including lexical retrieval (BM25), domain-specific dense retrieval (MedCPT), and sentence-level embedding models (PubMedBERT-matryoshka). Results 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 (Wang et al., 2019). 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

Dataset	Method	Standard Narrative				Clinical Narrative			
		Hit@10	MRR@10	Recall@10	Hit0↓	Hit@10	MRR@10	Recall@10	Hit0↓
PubMedQA	E-utils	41.67	38.06	41.67	22	0.00	0.00	0.00	55
	Adapt	**48.33**	**44.72**	**48.33**	**0**	8.33	6.04	8.33	**1**
	MedCPT	10.00	9.17	10.00	0	3.33	2.50	3.33	0
	Hybrid	46.67	33.25	46.67	0	**10.00**	**5.28**	**10.00**	0
BioASQ	E-utils	52.44	35.49	28.87	18	1.22	1.22	1.22	76
	Adapt	65.85	41.94	41.61	**0**	50.00	28.93	33.32	**1**
	MedCPT	63.41	42.39	36.50	0	41.46	21.43	21.70	0
	Hybrid	**80.49**	**49.74**	**51.21**	0	**60.98**	**35.36**	**40.07**	0

Table 1: Ablation study on retrieval performance. **Adapt** refers to our Boolean-constrained retrieval with adaptive query execution. **E-utils** refer to E-utilizes.

Source	Acc	Prec	Rec	F1
D1 (Abstracts)	44.00	41.12	40.12	39.14
D1 + D2 (Reviews)	**46.00**	**43.78**	**38.13**	**38.77**
D1 + D2 + D3 (Other)	46.00	40.12	33.86	35.33

Table 2: Impact of data sources on retrieval performance (Negative Cancer QA Dataset). D1: PubMed Abstracts; D2: PMC Reviews; D3: Other PMC articles.

	Retriever Accuracy (%)		
Splitter Configuration	PubMedBERT	BM25	MedCPT
512 (Overlap 0)	46	20	22
512 (Overlap 32)	52	18	24
512 (Overlap 128)	42	16	22
SEOS (Ours)	**54**	**36**	**38**

Table 3: Retrieval accuracy across different text segmentation strategies on the Passage Retrieval Evaluation Dataset. SEOS consistently outperforms fixed-size chunking.

combination of dense retriever and reranker is recommended to achieve the nuanced understanding required.

4.2.2 Comparison of Retrievers and Rerankers

We implemented retrieval evaluation based on LlamaIndex (Liu, 2022), 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 (Tang and Yang, 2024). Results are shown in Figure 4.

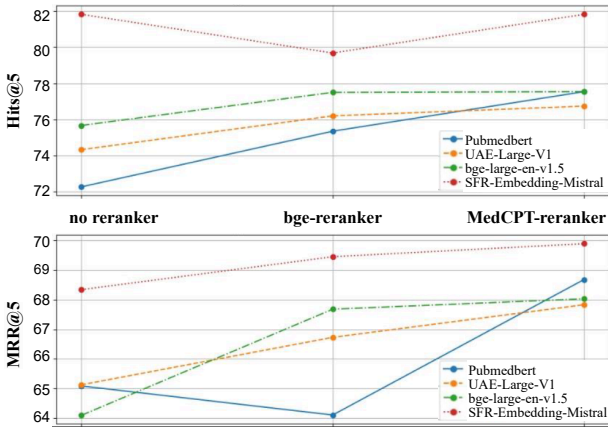

Figure 4: Performance of different reranker-retriever (reranker-embedding) pairs measured by Hit@5 (top) and MRR@5 (bottom).

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 (Xiong et al., 2024). 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

Method	MMLU	MedQA	MedMC	PMQA	BioASQ	Avg
LLM + CoT	78.26	68.60	65.59	45.00	80.49	67.15
Naive RAG	**82.61**	67.44	65.59	56.67	81.71	69.48
Ours	**82.61**	**69.77**	**68.82**	**65.00**	**81.71**	**72.39**

Table 4: Performance comparison on CMMQA datasets. **Ours** combines CHSDR , SEOS with matryoshka embeddings and medcpt reranker, significantly outperforming baselines.

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.

4.3 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 4.

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 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 (Wang et al., 2023b; Asai et al., 2023) have reported that RAG can negatively impact the original outcome, particularly when LLM's inherent parametric knowledge suffices for the query (Wang et al., 2023b). This occurs because LLMs tend to rely on retrieval results, impairing creativity and versatility, even accuracy (Asai et al., 2023). Our framework helps bridge the gap between patients and a rapidly expanding biomedical literature by leveraging both real-time and semantic retrieval pathways, ensuring timely and trustworthy cancer-related information. Clinically, this supports more informed shared decision-making based on accurate medical knowledge.

5 Conclusion

In this study, we optimized the query pipeline for Cancer Patient QA (CPQA) by introducing **HSRDR**, a novel hybrid document retrieval method, and **SEOS**, a semantic-aware segmentation technique. Our evaluations yield two critical insights. First, document retrieval benefits significantly from the complementarity of real-time Boolean search and semantic retrieval, with PMC reviews showing distinct utility over other sources. Second, passage retrieval performance relies heavily on the alignment between domain-specific embedding models and rerankers. Beyond immediate performance gains in oncology, our contributions have broader applicability: while HSRDR is tailored for the high-stakes, knowledge-intensive biomedical domain, the SEOS segmentation strat-

egy is domain-agnostic and offers potential improvements for any RAG system dealing with complex, semantic-rich text. Future work will explore extending SEOS to multi-modal data and further refining real-time filtering mechanisms for clinical safety.

Limitations

We utilize Boolean operator conversion to enhance term-based online search, whether other query refinement methods (like Query2Doc (Wang et al., 2023a) and RAG-Fusion (Rackauckas, 2024)) can further improve performance needs to be explored. We use multiple-choice-style evaluation, which may overlook nuanced details and is susceptible to guessing bias, where random guesses will boost accuracy. 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

Retrieval-augmented generation for large language models: A survey.

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. In MEDINFO 2019: Health and Wellbeing e-Networks for All, pages 25–29. IOS Press.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In The Twelfth International Conference on Learning Representations.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. 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, pages 758–759.

Marti A Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. Computational linguistics, 23(1):33–64.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In International Conference on Learning Representations.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. Bioinformatics, 40(Supplement_1):i119–i129.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1–38.

Fan Jiang, Qiongkai Xu, Tom Drummond, and Trevor Cohn. Boot and switch: Alternating distillation for zero-shot dense retrieval. In The 2023 Conference on Empirical Methods in Natural Language Processing.

Di Jin, Eileen Pan, Nassim Oufattole, Weng Wei-Hung, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. 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), pages 2567–2577.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. Bioinformatics, 39(11):btad651.

Jonathan Kans. 2024. Entrez direct: E-utilities on the unix command line. In Entrez programming utilities help [Internet]. National Center for Biotechnology Information (US).

Esther Landhuis. 2016. Scientific literature: Information overload. Nature, 535(7612):457–458.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33:9459–9474.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models a better foundation for dense retrieval. Preprint, arXiv:2312.15503.

Sichu Liang, Linhai Zhang, Hongyu Zhu, Wenwen Wang, Yulan He, and Deyu Zhou. 2025. Rgar: Recurrence generation-augmented retrieval for factual-aware medical question answering. arXiv preprint arXiv:2502.13361.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? Patterns, 5(3).

Jerry Liu. 2022. Llamaindex. Acceso el, 6.

Chris J Lu, Alan R Aronson, Sonya E Shooshan, and Dina Demner-Fushman. 2019. Spell checker for consumer language (cspell). Journal of the American Medical Informatics Association, 26(3):211–218.

Yiqun Miao, Yuhan Zhao, Yuan Luo, Huiying Wang, and Ying Wu. 2025. Improving large language model applications in the medical and nursing domains with retrieval-augmented generation: Scoping review. Journal of Medical Internet Research, 27:e80557.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. Medicine, 84(88.3):77–3.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Conference on health, inference, and learning, pages 248–260. PMLR.

Zackary Rackauckas. 2024. Rag-fusion: a new take on retrieval-augmented generation. arXiv preprint arXiv:2402.03367.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval, 3(4):333–389.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023a. Large language models encode clinical knowledge. Nature, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, and 1 others. 2023b. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. arXiv e-prints, pages arXiv–2401.

Daniel Truhn, Jorge S Reis-Filho, and Jakob Nikolas Kather. 2023. Large language models should be used as scientific reasoning engines, not knowledge databases. Nature medicine, 29(12):2983–2984.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics, 16(1):138.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. arXiv preprint arXiv:2307.15343.

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

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023b. Self-knowledge guided retrieval augmentation for large language models. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 10303–10315.

Yinuo Wang, Baiyang Wang, Robert Mercer, Frank Rudzicz, Sudipta Singha Roy, Pengjie Ren, Zhumin Chen, and Xindi Wang. 2025. Trustworthy medical question answering: An evaluation-centric survey. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27477–27490.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering. arXiv preprint arXiv:1908.08167.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In Findings of the Association for Computational Linguistics: ACL 2024, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

A Example Appendix

A.1 Test Data Details

We construct evaluation dataset by appling a MeSH-based filter on six widely used medical datasets:

- PubMedQA (Jin et al., 2019) and BioASQ (Xiong et al., 2024) are QA datasets with answers from biomedical literature. PubMedQA (500 questions) is characterized by **single-document grounding**, where each manually annotated question is paired with a single corresponding PMID. In contrast, **BioASQ** (Xiong et al., 2024) (618 questions) features **multi-document grounding**, providing a list of multiple relevant PMIDs for each question. They are suitable for document retrieval evaluation due to varying evidence densities and the explicit annotation of gold-truth PMIDs.

- MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and the medical subsets in MMLU (Hendrycks et al., 2020) collected questions from medical exams. Specifically, MedQA (Jin et al., 2021) contains 1273 questions from the US Medical Licensing Examination (USMLE), focusing on complex clinical decision-making scenarios faced by professionals. MedMCQA is a large-scale dataset covering 21 medical subjects with 4183-question development set, focusing on evaluating across diverse healthcare topics. MMLU-med is a subset of the MMLU benchmark comprising 1,089 questions across six subtasks (e.g., professional medicine, human genetics), serving as a testbed for general biomedical reasoning.

- HealthSearchQA (Singhal et al., 2023b), comprising general consumer search queries, lacks definitive answers and is therefore unsuitable for accuracy assessment. However, its question lengths make it suitable for investigating the impact of question length on document retrieval methods. (Nori et al., 2023; Singhal et al., 2023a; Liévin et al., 2024).

First five datasets are multiple-choice QA datasets, because using multiple-choice questions can simplify evaluation, eliminate biases from text similarity computations or human annotation, and align with large-scale medical QA systems evaluations.

A.2 The pseudo-code of SEOS

Algorithm 1 Semantic **E**nhanced **O**verlap **S**egmentation

```
procedure SEOS(document, k, embed_model)
    Split document into sentences using sentence_splitter
    Merge each sentence with its k neighboring sentences
    Get each sentence group's embedding by embed_model
    Use cosine dissimilarity between adjacent groups to identify similarity inflection (semantic shift) points
    Adjust chunk_size based on embed_model
    n ← len(document)/chunk_size
    Breakpoints:
top n points with largest semantic shift
    Split document into chunks at breakpoints
    Overlap is added to per chunk by including the first sentence of the next chunk to preserve sequence order
    return Chunks
end procedure
```

A.3 Retrieval Examples

Retrieval Examples (Output)

"step_by_step_thinking":
"The question asks whether pneumologists' skepticism about chemotherapy and survival in advanced non-smallcell lung carcinoma is justified. The context information suggests that chemotherapy can prolong survival and improve quality of life for patients with advanced non-small cell lung cancer, especially those with good performance status. Several studies have shown that platinum-based chemotherapy regimens are the standard of care and can lead to improved survival compared to supportive care alone.

However, the context also suggests that there may be some skepticism among pneumologists about the benefits of chemotherapy.
The context indicates that while chemotherapy can prolong survival, the benefits are moderate, with median and 1-year survival of 8--10 months and 30--40\%, respectively. Additionally, the survival benefits vary depending on factors like disease stage, patient characteristics, and chemotherapy regimen.

Overall, the context information suggests that chemotherapy can provide meaningful survival and quality of life benefits for many patients with advanced non-small cell lung cancer, but the benefits are not dramatic and must be weighed against the potential for toxicity. This may explain some of the skepticism among pneumologists.

Based on the information provided, I believe the answer is B: no, the pneumologists' skepticism is not justified. While the benefits of chemotherapy are not as large as some might hope, the context information indicates that it can still provide important survival and quality of life benefits for many patients, and the risks are generally manageable. The skepticism may be understandable, but the evidence does not seem to fully support it."

"answer_choice": "B"

A.4 LLM Prompts

Prompt: Rewrite to Clinical Narrative Queries

```
prompt_searchapi = """
### Role
You are a simulator for a Consumer Cancer QA System. Real users have no medical background,
including cancer patients and their family members.

### Input Data
RAW QUESTION: {question}

### Task
Identify the medical scenario, You need to transform RAW QUESTION into a a first-person query from
a real user.

### Requirements
1. Use layperson language (e.g., "spread" instead of "metastasize").
2. Include 1-2 minor typos (e.g., "cancr" for "cancer") or grammar slips.
3. Considering the real situation of non-expert users, rewrite from patient perspective.
4. Keep the core medical doubt unchanged so it can still be answered.

### Output Format
Return ONLY the rewritten string. No explanations.

"""
```

LLM with CoT + RAG

```
prompt_template = """
You are a helpful medical expert.

QUESTION: {question}
POTENTIAL CHOICES: {options}
EVIDENCE: {context}

Task: answer a multi-choice medical question with provided evidence.

Instructions:
1.think step-by-step then choose the answer for QUESTION from the POTENTIAL CHOICES, using the
EVIDENCE.
2.Organize your output in a json formatted as Dict{{"step_by_step_thinking":Str(explanation),
"answer_choice": Str{{A/B/C/...}}}}.

Note:
1.Only Dict in output is needed. No need for other text.
2.Your responses will be used for research purposes only, so please have a definite answer.
3.Keep your answer ground in the facts of the EVIDENCE.
"""
```

Prompt: Rewrite for boolean queries and set time constraints

```
prompt_searchapi = """
You are a helpful expert in information extraction.

QUERY: {query}

Task: generate boolean search queries for QUERY

Instructions:
1. Preprocessing Stage:
- **Correct** any orthographic or grammatical errors in the QUERY.
- **Analyze** the Question Type and User intent to identify core concepts and **SEPARATE** the
"Noise" from the "Core Dilemma".
- **Extract** the core terms. Especially identify **PICO** (Population, Intervention, Outcome)
- **Abstraction** specific scenario details into general medical concepts (e,g. "Don't tell" ->
"Truth-telling").
- **Time**: If time constraints (like "recent") appears, set reasonable year_range (e.g., 5).
Otherwise None.
2. Boolean Queries Generation Stage:
- Provide 5 possible boolean search queries using core terms.
- Start with the most specific query (Strict AND of all terms).
- Progressively relax the strict query based on your Intent Analysis in stage 1:
[Strategy A (Synonyms): If a term is a CORE concept, relax it by adding synonyms];
[Strategy B (Pruning): If a term is a SPECIFIC but NOt CORE detail (e.g., "in Italy"), remove]
[Strategy C (Logical Relaxation)]: Switch AND to OR between distinct core concepts


Examples:
# Example 1:
Input: "prevalence of concurrent diabetes and cardiovascular disease in women in Beijing"
Output: (
  [
    '"prevalence" AND "diabetes" AND "cardiovascular disease" AND "women" AND "Beijing"',
    '"prevalence" AND "diabetes" AND "heart disease" AND "women"',
    '"epidemiology" AND "concurrent" AND "diabetes" AND "cardiovascular disease"',
    '"diabetes" AND "cardiovascular disease" AND "comorbidity"',
    '"prevalence" AND ("diabetes" OR "cardiovascular disease")'
  ],
None
)
# Example 2:
# Abstractions: Stroke->End of life; Disconnect->Withdrawal; Son/Daughter->Family Conflict
Input: "75-year-old man with severe stroke and ventilator. Daughter wants to disconnect the
machine but son disagrees."
Output: (
  [
      '"withdrawal of life support" AND "family conflict" AND "surrogate decision making"',
      '"withdrawal of life support" AND ("family conflict" OR "disagreement") AND "decision
      making"',
      '"withdrawal of life support" AND "family conflict"',
      '"surrogate decision making" AND "family conflict"',
      '"withdrawal of life support" OR "end of life care"'
  ],
None
)
# Example 3: Time Sensitive
Input: "latest treatments for glioblastoma in the last 8 years"
Output: (
    [
        '"Glioblastoma"[MeSH] AND "Therapeutics"[MeSH]',
        '"Glioblastoma" AND "Treatment" AND "Novel"',
        '"Brain Neoplasms" AND "Therapy"',
        '"Glioblastoma" AND "Immunotherapy"',
        '"Glioblastoma" OR "Brain Tumor"'
    ],
    8
)

Note: Return ONLY the Python tuple (list_of_strings, year_range_int). No other text.

"""
```