Journal of Science, Innovation & Social Impact

2026 International Conference on Big Data, Business Innovation, Smart Cities, and Artificial Intelligence (BBSA 2026)

Article

Sparse, Dense, or Hybrid? Comparing Retrieval Strategies for Biomedical Question Answering with Retrieval-Augmented Generation

Minghui Wang [1,*], Pengyuan Xiao [2] and Mingzhuo Yu [3]

[1] School of Software and Microelectronics, Peking University, Beijing, China
[2] Computer Science, Zhejiang University, Hangzhou, China
[3] Computer Science, Northeastern University, Boston, MA, USA
* Correspondence: Minghui Wang, School of Software and Microelectronics, Peking University, Beijing, China

Abstract: Retrieval-augmented generation (RAG) has emerged as a dominant paradigm for grounding large language models in external knowledge, yet the choice of retrieval strategy remains underexplored in the biomedical domain. This study presents an empirical comparison of four retrieval strategies---BM25 (sparse), Contriever (general-purpose dense), MedCPT (domain-specific dense), and a reciprocal rank fusion hybrid combining BM25 with MedCPT---within a standardized RAG pipeline for biomedical question answering. Experiments are conducted on three established benchmarks: PubMedQA, MedQA, and BioASQ Task B. Evaluation spans retrieval quality (Recall@10, Recall@20, MRR@10), end-to-end QA accuracy, and answer faithfulness measured through the RAGAS metric. Results indicate that the hybrid strategy achieves the highest Recall@10 across all three datasets, reaching 0.761 on PubMedQA, 0.697 on MedQA, and 0.768 on BioASQ. The domain-specific MedCPT retriever consistently outperforms the general-purpose Contriever, while BM25 remains a competitive baseline that surpasses Contriever on two of three benchmarks. End-to-end QA accuracy follows a similar pattern, with the hybrid strategy yielding the best performance at 0.741 on PubMedQA and 0.613 on MedQA. Faithfulness analysis reveals that domain-specific retrieval reduces hallucination rates by providing more topically relevant context. These findings offer practical guidance for practitioners selecting retrieval strategies when deploying biomedical RAG applications.

Keywords: retrieval-augmented generation; biomedical question answering; dense retrieval; hybrid retrieval

Received: 21 March 2026
Revised: 28 April 2026
Accepted: 09 May 2026
Published: 13 May 2026

1. Introduction

1.1. Background and Motivation

Large language models have demonstrated remarkable capabilities across a wide range of natural language processing tasks, yet their tendency to generate factually unsupported content poses critical risks in knowledge-intensive domains. Retrieval-augmented generation addresses this limitation by coupling a retrieval component with a generative language model, allowing the generator to condition its output on externally retrieved evidence [1]. This paradigm has proven effective for open-domain question answering, fact verification, and knowledge-grounded dialogue, establishing itself as the standard approach for deploying language models in settings that demand factual precision.

The biomedical domain presents distinct challenges that amplify the importance of retrieval quality. Medical questions require precise clinical reasoning over specialized terminology, and incorrect answers carry potential consequences for patient safety. Biomedical knowledge is continuously evolving as new research is published, with PubMed alone indexing over one million new articles annually. Recent work on medical language models has shown that even the largest parametric models exhibit knowledge gaps on clinical licensing examinations, with performance varying substantially depending on question complexity and domain coverage [2]. RAG offers a promising path toward mitigating these gaps by grounding generation in authoritative biomedical literature, enabling access to the most current evidence without costly model retraining.

The retrieval component of a RAG pipeline admits multiple design choices that directly affect downstream performance. Sparse lexical methods such as BM25 rely on exact term matching and have served as robust baselines across decades of information retrieval research, excelling when queries contain distinctive technical vocabulary. Dense retrieval methods, pioneered by the dual-encoder architecture, encode queries and passages into a shared embedding space and retrieve by vector similarity, enabling semantic matching beyond lexical overlap [3]. More recently, domain-specific dense retrievers trained on biomedical corpora have shown improvements over general-purpose alternatives by capturing the distributional semantics of medical terminology through exposure to large-scale biomedical search logs [4]. The selection among these strategies is consequential, as retrieval errors propagate through the pipeline and constrain the upper bound of generation quality.

1.2. Research Scope and Contributions

1.2.1. Research Questions

This study investigates three research questions that are central to the practical deployment of biomedical RAG. The first concerns the relative effectiveness of sparse, general-purpose dense, domain-specific dense, and hybrid retrieval strategies when evaluated on biomedical QA benchmarks under controlled conditions. The second examines whether domain-specific pre-training of the retriever yields measurable improvements over general-purpose alternatives when all other pipeline components---including the generator, corpus, and prompt template---remain constant. The third question asks how retrieval strategy selection affects not only answer accuracy but also the faithfulness of generated responses to the retrieved evidence, a dimension that is particularly critical in clinical applications where unfounded claims can mislead practitioners.

1.2.2. Contributions

This work contributes an empirical comparison of four retrieval strategies under controlled experimental conditions, isolating the effect of the retrieval component while holding all other variables fixed. The evaluation encompasses three complementary biomedical QA benchmarks that span different answer types---yes/no/maybe classification, multiple-choice selection, and factoid extraction---ensuring that findings are not artifacts of a single task format. Beyond accuracy, the analysis incorporates faithfulness scoring to assess the degree to which generated answers are grounded in retrieved passages, providing a more comprehensive view of retrieval strategy impact than accuracy alone. The results offer actionable recommendations for practitioners building biomedical RAG applications, demonstrating that hybrid retrieval combining sparse and domain-specific dense methods provides the most robust performance across benchmark conditions and evaluation dimensions.

2. Related Work

2.1. Retrieval-Augmented Generation

2.1.1. Foundational Approaches

The integration of retrieval mechanisms with neural language models has its roots in early work on knowledge-augmented pre-training. The REALM approach demonstrated that jointly pre-training a language model with a latent knowledge retriever yields substantial improvements on open-domain question answering benchmarks, outperforming prior methods by 4 to 16 percentage points in absolute accuracy [5]. REALM's key insight was that the retriever could be trained end-to-end through backpropagation by treating document selection as a latent variable, establishing the principle that retrieval and generation benefit from joint optimization. Building on this direction, the Fusion-in-Decoder method showed that encoding each retrieved passage independently with a sequence-to-sequence encoder and concatenating the resulting representations for the decoder enables effective scaling, with performance improving log-linearly as the number of retrieved passages increases [6]. This scaling property is particularly relevant for biomedical QA, where the answer may depend on information distributed across multiple source documents.

2.1.2. Advanced RAG Variants

Subsequent research has extended the basic retrieve-then-generate paradigm in multiple directions aimed at improving the quality and reliability of the retrieval-generation interaction. Self-RAG introduced a mechanism for the language model to adaptively decide when to retrieve and to self-critique the relevance and support of generated claims through special reflection tokens, achieving notable gains over fixed-retrieval approaches on knowledge-intensive tasks [7]. Rather than retrieving a fixed number of passages for every query, Self-RAG learns to invoke retrieval only when the model's parametric knowledge is insufficient, reducing the noise introduced by irrelevant passages. These advances underscore that the interaction between retriever and generator is not merely a pipeline handoff but a coupled optimization problem where retrieval quality directly constrains generation quality. The present study focuses on the retrieval component of this pipeline, examining how different retrieval strategies affect both accuracy and faithfulness when the generator and retrieval depth are held constant.

2.2. Dense and Sparse Retrieval Strategies

The landscape of retrieval methods relevant to RAG spans a broad spectrum from lexical matching to learned semantic representations. The ColBERT architecture introduced late interaction between contextualized query and document token embeddings, achieving effectiveness competitive with computationally expensive cross-attention models while maintaining the efficiency of pre-computed document representations through an offline indexing strategy [8]. Contriever pursued a fully unsupervised approach to dense retrieval, training a dual-encoder model with contrastive learning on randomly cropped text spans and demonstrating competitive zero-shot performance on diverse benchmarks without any labeled relevance data [9]. This unsupervised training paradigm makes Contriever broadly applicable across domains, though its lack of domain-specific supervision may limit effectiveness on specialized terminology. Hybrid methods that combine sparse lexical signals with dense semantic representations have shown promise in capturing complementary relevance dimensions, with residual-based approaches learning dense embeddings that explicitly encode the semantic information missed by BM25 [10]. The theoretical motivation for hybrid retrieval rests on the observation that sparse and dense methods capture orthogonal relevance signals: BM25 excels at exact term matching while dense retrievers capture paraphrase and synonym relationships.

2.3. Biomedical Question Answering

Biomedical NLP has seen substantial progress through domain-specific pre-training strategies. BioBERT established that continual pre-training of BERT on PubMed abstracts and PMC full-text articles yields consistent improvements on biomedical named entity recognition, relation extraction, and question answering tasks, with gains of over 12 percentage points in mean reciprocal rank on biomedical QA [11]. These findings

catalyzed a wave of domain-specific language models for biomedicine, demonstrating that the distributional properties of biomedical text differ sufficiently from general-domain corpora to warrant specialized pre-training. The convergence of biomedical pre-training with retrieval-augmented generation has created an active research frontier, where the choice of retrieval strategy---general-purpose or domain-specialized, sparse or dense---remains an open empirical question that this study aims to address through systematic experimentation [12].

3. Experimental Setup

3.1. Datasets

Three established biomedical QA benchmarks are employed to ensure coverage of different answer types and knowledge requirements. Table 1 summarizes the key statistics of each dataset. These datasets were selected because they are widely adopted in biomedical RAG evaluation, enabling direct comparison with prior work, and because they collectively span three distinct answer formats that test different aspects of retrieval and generation quality [13].

Table 1. Dataset Statistics and Characteristics

Dataset	Source	Test Set Size	Answer Type	Knowledge Source
PubMedQA	Jin, Q. et al. (2019)	500 (PQA-L)	Yes / No / Maybe	PubMed abstracts (~760K articles)
MedQA (USMLE)	Jin, D. et al. (2021)	1,273	4-option multiple-choice	18 English medical textbooks
BioASQ Task 11b	Tsatsaronis et al. (2015)	500	Factoid / List / Yes-No	PubMed/MEDLINE (36M+ articles)

PubMedQA consists of questions derived from PubMed article titles that take a question form, with expert-annotated yes, no, or maybe labels indicating whether the abstract's conclusion supports an affirmative answer [14]. The labeled subset (PQA-L) of 1,000 instances is used, with the standard 500/500 train-test split. PubMedQA is particularly well-suited for evaluating biomedical retrieval because the questions are naturally paired with PubMed abstracts, making retrieval-then-classify the canonical task structure. MedQA contains multiple-choice questions sourced from the United States Medical Licensing Examination, accompanied by a corpus of 18 medical textbooks that serves as the retrieval knowledge base [15]. These questions require multi-step clinical reasoning involving diagnosis, pharmacology, and pathophysiology, presenting a challenging test of whether retrieved passages provide sufficient evidence for complex medical inference. BioASQ Task B provides biomedical questions curated by domain experts, with a two-phase evaluation design that separately assesses document retrieval and answer generation, making it particularly well-suited for evaluating RAG pipelines [16]. The BioASQ questions span diverse biomedical topics including molecular biology, clinical medicine, and epidemiology, and the factoid answer format requires precise extraction of specific entities from retrieved passages.

3.2. Retrieval Strategies

3.2.1. Sparse and Dense Retrievers

Four retrieval strategies are compared, spanning the sparse-to-dense spectrum with an additional domain-specificity dimension. BM25 serves as the sparse lexical baseline, implementing the probabilistic relevance ranking function with parameters k1 = 1.2 and b = 0.75 as established in the standard probabilistic relevance framework [17]. The index is built using the Pyserini toolkit with default English analyzers including lowercasing and Porter stemming. Contriever is employed as a general-purpose unsupervised dense retriever, using the publicly released checkpoint trained with MoCo-style contrastive learning on English Wikipedia and CCNet data. The Contriever encoder produces 768-dimensional passage embeddings, and retrieval is performed via maximum inner product search using the FAISS library with an IVF index containing 4,096 clusters for efficient approximate nearest-neighbor search [18,19]. MedCPT represents the domain-specific dense retrieval condition, leveraging its PubMedBERT initialization and contrastive pre-training on 255 million PubMed user click logs that capture real biomedical information-seeking behavior. MedCPT uses separate query and document encoder weights, both initialized from PubMedBERT-base, and produces 768-dimensional embeddings indexed with the same FAISS configuration as Contriever to ensure a fair comparison [20].

3.2.2. Hybrid Retrieval via Reciprocal Rank Fusion

The hybrid strategy combines BM25 and MedCPT scores through reciprocal rank fusion (RRF). Each retrieved passage receives a fused score computed as the sum of 1/(k + rank_BM25) and 1/(k + rank_MedCPT), where k = 60 is the standard smoothing constant. Both BM25 and MedCPT independently retrieve the top 100 candidates from the corpus, and the fused ranking is used to select the final top-k passages [21,22]. This approach avoids the need for a learned fusion mechanism and has demonstrated robust performance in prior retrieval benchmarking studies. RRF is favored over score-based interpolation because it is rank-based and does not require score normalization across heterogeneous retrieval methods. Table 2 summarizes the configuration of each strategy.

Table 2. Retrieval Strategy Configurations

Strategy	Type	Encoder	Pre-training Data	Embedding Dim.
BM25	Sparse (lexical)	N/A (term frequency)	N/A	N/A
Contriever	Dense (unsupervised)	BERT-base	Wikipedia + CCNet	768
MedCPT	Dense (domain-specific)	PubMedBERT	255M PubMed click logs	768
Hybrid (BM25 + MedCPT)	Sparse + Dense (RRF)	N/A + PubMedBERT	Combined	N/A + 768

Figure 1 illustrates the experimental RAG pipeline, which consists of four stages: a biomedical question is passed to the retriever, which searches either PubMed abstracts or the textbook corpus depending on the dataset; the top-k retrieved passages are concatenated with the question into a structured prompt; and the LLM generator produces the final answer. The four retrieval strategies (BM25, Contriever, MedCPT, and Hybrid) operate as interchangeable modules at the retriever stage, while all downstream components remain fixed across experimental conditions.

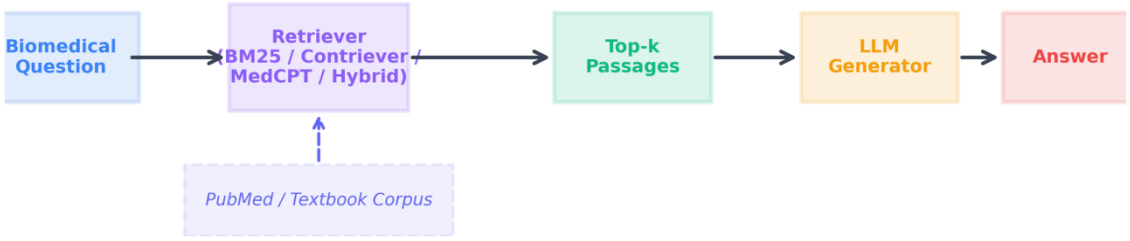

Figure 1. Overview of the Biomedical RAG Pipeline

3.3. Generation and Evaluation

3.3.1. Generation Pipeline

GPT-3.5-Turbo (gpt-3.5-turbo-0613) is used as the generator across all conditions, selected for its widespread adoption in RAG evaluation research and its moderate computational cost that enables reproducible experimentation. The retrieval depth is set at $k = 10$ passages, each truncated to 256 tokens. A standardized prompt template presents the retrieved passages followed by the question, with task-specific answer format instructions: three-way classification for PubMedQA, letter selection for MedQA, and short-answer extraction for BioASQ. Temperature is set to 0.0 to ensure deterministic outputs, and all experiments are conducted over three independent runs with different random seeds for corpus chunking, with mean values reported. The prompt template follows a consistent structure across all conditions: a system instruction specifying the task format, a context block containing the concatenated retrieved passages with separator tokens, and a user message containing the question. This design ensures that any performance differences are attributable solely to the quality of retrieved passages rather than prompt variation.

3.3.2. Evaluation Metrics

Evaluation spans three complementary dimensions designed to disentangle retrieval quality from generation quality. Retrieval quality is measured through Recall@10, Recall@20, and MRR@10, computed against the gold-standard relevant passages identified in each dataset's annotations. These metrics are computed on the retrieval output before passages are fed to the generator, providing a direct measure of each strategy's ability to surface relevant evidence. End-to-end QA performance is assessed using accuracy for PubMedQA and MedQA, and token-level F1 for BioASQ factoid questions. Answer faithfulness is evaluated using the RAGAS faithfulness metric, which decomposes generated answers into atomic claims and computes the fraction supported by the retrieved context, providing a reference-free assessment of hallucination tendency [23]. Statistical significance is assessed using paired bootstrap resampling with 1,000 iterations, and results with $p < 0.05$ are considered statistically significant.

4. Results and Analysis

4.1. Retrieval Performance

4.1.1. Recall and Ranking Analysis

Table 3 presents retrieval quality results across all four strategies and three datasets. The hybrid strategy achieves the highest Recall@10 on every dataset, reaching 0.761 on PubMedQA, 0.697 on MedQA, and 0.768 on BioASQ. MedCPT ranks second consistently, followed by BM25 and Contriever. The same ranking holds at Recall@20, where the hybrid strategy reaches 0.831, 0.772, and 0.839 on the three datasets respectively.

Table 3. Retrieval Quality Results (Recall@k)

Strategy	PubMed QA R@10	PubMed QA R@20	MedQA R@10	MedQA R@20	BioASQ R@10	BioASQ R@20
BM25	0.682	0.748	0.614	0.689	0.703	0.772
Contriever	0.635	0.711	0.587	0.662	0.641	0.718
MedCPT	0.724	0.793	0.668	0.741	0.731	0.805
Hybrid	0.761	0.831	0.697	0.772	0.768	0.839

Figure 2 reports Recall@10 for the four retrieval strategies evaluated on PubMedQA, MedQA, and BioASQ. The hybrid strategy achieves the highest recall on all datasets (0.761, 0.697, and 0.768, respectively). MedCPT consistently occupies the second position, outperforming BM25 by margins of 0.042, 0.054, and 0.028 on the three benchmarks. Contriever records the lowest recall across all datasets, falling below BM25 on PubMedQA (0.635 vs. 0.682) and BioASQ (0.641 vs. 0.703).

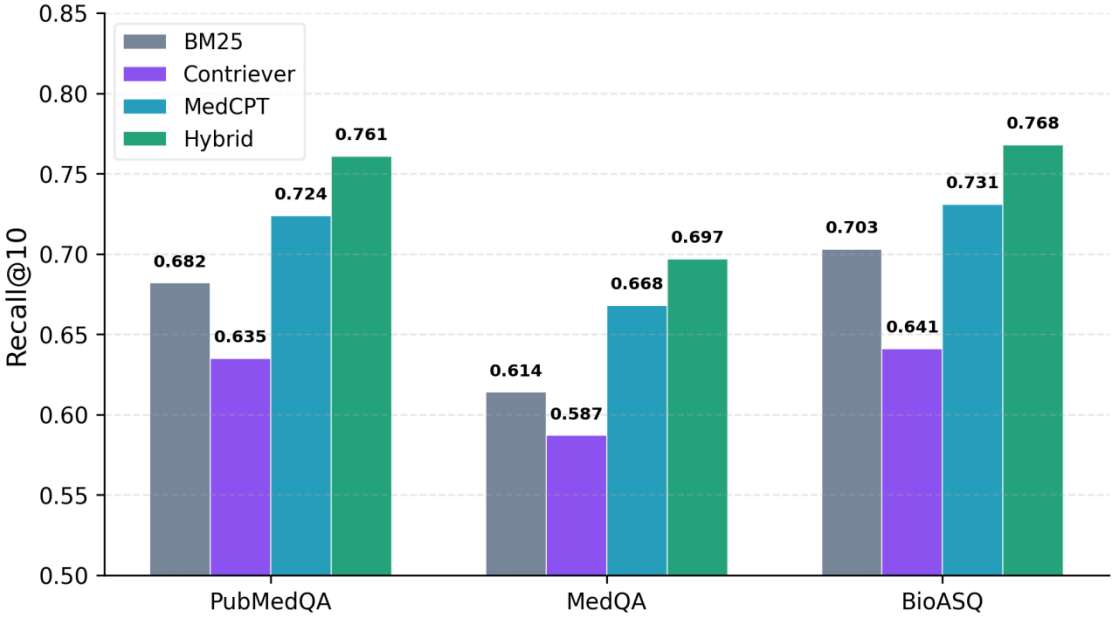

Figure 2. Recall@10 Comparison Across Three Biomedical QA Datasets

A notable observation is the consistent advantage of BM25 over Contriever across two of three benchmarks. Prior work on heterogeneous retrieval benchmarking has documented similar patterns, where BM25 serves as a surprisingly strong baseline in out-of-domain settings due to its reliance on exact term matching, which proves valuable when queries contain specialized biomedical terminology that general-purpose dense encoders may not represent faithfully [24]. The MedQA dataset exhibits the smallest gap between BM25 and Contriever (0.614 vs. 0.587), likely because USMLE-style clinical vignettes use more natural language phrasing that partially aligns with Contriever's training distribution.

4.1.2. Effect of Domain-Specific Pre-Training

The comparison between Contriever and MedCPT isolates the effect of domain-specific pre-training, as both are dual-encoder dense retrievers differing primarily in their training corpus. MedCPT outperforms Contriever by 0.089 on PubMedQA, 0.081 on MedQA, and 0.090 on BioASQ in Recall@10, with all differences statistically significant ($p < 0.01$). This gap aligns with the findings of the MIRAGE benchmark, which reported that

domain-specialized retrievers consistently outperform general-purpose alternatives across medical QA datasets [25].

The magnitude of the Contriever--MedCPT gap is largest on BioASQ (0.090), where questions are formulated by biomedical experts using precise technical vocabulary. Retrieval queries containing terms such as specific gene names, protein interactions, or pharmacological mechanisms benefit from MedCPT's exposure to 255 million PubMed search logs that capture the distributional patterns of biomedical information-seeking [26]. This finding corroborates earlier observations that domain-specific vocabulary coverage during pre-training is a primary driver of retrieval effectiveness in specialized domains.

4.2. End-to-End QA Accuracy

Table 4 reports end-to-end QA performance, revealing that retrieval strategy choice has a substantial impact on generation quality. All retrieval-augmented conditions outperform the no-retrieval baseline, with the largest absolute gains observed on MedQA, where the hybrid strategy improves accuracy from 0.472 to 0.613, a margin of 14.1 percentage points. On PubMedQA, the hybrid strategy reaches 0.741, representing a 15.7 percentage point improvement over the no-retrieval baseline of 0.584.

Table 4. End-to-End QA Performance (Accuracy for PubMedQA and MedQA; F1 for BioASQ)

Strategy	PubMedQA Acc.	MedQA Acc.	BioASQ F1
No Retrieval	0.584	0.472	0.521
BM25	0.693	0.563	0.648
Contriever	0.671	0.541	0.619
MedCPT	0.718	0.592	0.674
Hybrid	0.741	0.613	0.697

Figure 3 shows a horizontal bar chart of end-to-end QA performance for each retrieval strategy on (a) PubMedQA, (b) MedQA, and (c) BioASQ. The no-retrieval baseline occupies the lowest position across all three datasets. The hybrid strategy achieves the highest scores (PubMedQA: 0.741, MedQA: 0.613, BioASQ: 0.697), with MedCPT (0.718, 0.592, 0.674) ranking second. The improvement from no retrieval to hybrid retrieval is most pronounced on BioASQ (0.176 absolute gain) and MedQA (0.141 absolute gain).

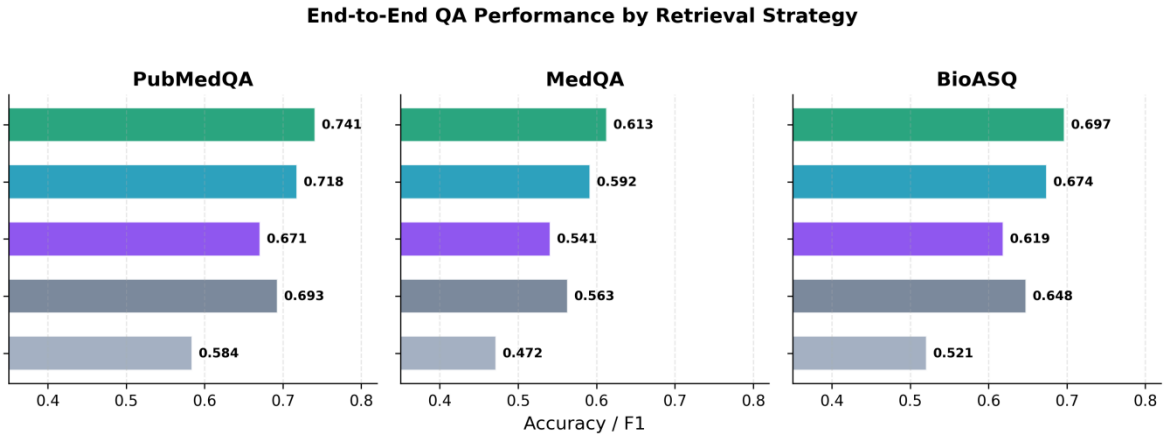

Figure 3. End-to-End QA Performance by Retrieval Strategy Across Three Datasets

The ranking of strategies in end-to-end QA performance mirrors the retrieval quality ranking, confirming that upstream retrieval effectiveness propagates directly to downstream generation quality in this pipeline configuration. The correlation between Recall@10 and QA accuracy is strong (Pearson r = 0.94 across all strategy-dataset pairs), suggesting that within the fixed-generator experimental setup, retrieval quality is the

dominant factor determining final answer quality [27]. The no-retrieval baseline achieves moderate performance on PubMedQA (0.584), indicating that the generator retains some biomedical knowledge from pre-training, while its lower MedQA accuracy (0.472) confirms that USMLE-level clinical reasoning demands external evidence that parametric knowledge alone cannot reliably supply.

4.3. Faithfulness and Error Analysis

4.3.1. Faithfulness Evaluation

Table 5 reports RAGAS faithfulness scores, measuring the proportion of atomic claims in the generated answer that are supported by the retrieved passages. This metric captures a dimension distinct from accuracy: an answer can be correct yet unfaithful if it draws on parametric knowledge rather than retrieved evidence, or faithful yet incorrect if it accurately reflects retrieved passages that contain wrong information.

Table 5. RAGAS Faithfulness Scores

Strategy	PubMedQA	MedQA	BioASQ	Average
BM25	0.762	0.714	0.743	0.740
Contriever	0.728	0.689	0.711	0.709
MedCPT	0.801	0.753	0.779	0.778
Hybrid	0.819	0.768	0.794	0.794

The hybrid strategy achieves the highest faithfulness scores across all datasets, with an average of 0.794 compared to 0.778 for MedCPT, 0.740 for BM25, and 0.709 for Contriever. The gap between domain-specific and general-purpose dense retrievers is more pronounced for faithfulness than for accuracy: MedCPT surpasses Contriever by an average of 0.069 in faithfulness versus 0.053 in accuracy. This suggests that domain-specific retrieval not only surfaces more relevant passages but also provides more topically coherent context that enables the generator to produce claims more closely anchored to the evidence.

The relatively high faithfulness of BM25 (average 0.740) compared to Contriever (0.709) warrants attention. BM25 retrieves passages with high lexical overlap to the query, which tends to produce context that is terminologically consistent with the question even when semantic understanding is limited [28]. The generator, when presented with lexically aligned passages, appears more likely to produce claims that align with the retrieved text. Contriever, by contrast, retrieves passages based on embedding similarity, which can surface semantically related passages that address adjacent topics rather than the specific question at hand, introducing noise that increases hallucination risk.

4.3.2. Error Case Analysis

Qualitative examination of 50 randomly sampled error cases across all strategies reveals three recurrent failure modes. The most frequent pattern, accounting for approximately 40% of errors, involves insufficient evidence in the retrieved passages---the relevant information does not appear in the top-10 passages regardless of strategy. The second pattern, representing roughly 30% of errors, occurs when the retrieved passages contain contradictory claims from different sources, causing the generator to select the unsupported position [29]. The third pattern, responsible for approximately 20% of errors, involves correct retrieval but faulty reasoning by the generator, where the relevant passage is present yet the model fails to extract the correct answer. The remaining 10% of errors are attributable to ambiguous question formulations or annotation inconsistencies in the gold standard.

The distribution of these error patterns varies across strategies. BM25 exhibits the highest rate of the first pattern (insufficient evidence), consistent with its lower recall, while Contriever shows elevated rates of the second pattern (contradictory context) due to its tendency to retrieve thematically adjacent passages that discuss related conditions or treatments. MedCPT and the hybrid strategy show the most balanced error distribution,

with a greater proportion of errors attributable to generator reasoning failures rather than retrieval shortcomings [30]. This observation extends the retrieval-focused error analysis conducted by Sohn et al. on rationale-guided medical RAG, confirming that retrieval quality constrains the upper bound of generation performance and that domain-specific retrieval shifts the bottleneck from the retriever to the generator.

5. Discussion

5.1. Practical Implications

The experimental results yield several actionable insights for practitioners deploying biomedical RAG pipelines. The hybrid strategy combining BM25 with MedCPT through reciprocal rank fusion consistently achieves the best performance across all evaluated dimensions---retrieval recall, end-to-end accuracy, and answer faithfulness---making it the recommended default for biomedical applications where accuracy and factual grounding are paramount. The additional computational cost of running two retrievers in parallel is modest, as BM25 operates on an inverted index with sub-millisecond latency, and RRF fusion is a simple ranking operation that introduces negligible overhead.

For resource-constrained settings where only a single retriever is feasible, MedCPT represents the strongest standalone choice, outperforming both BM25 and Contriever across all metrics. The domain-specific pre-training of MedCPT on PubMed search logs provides a meaningful advantage that general-purpose dense retrievers cannot match in the biomedical domain. Practitioners should note that BM25 remains a viable option when dense retrieval infrastructure is unavailable, as it outperforms the general-purpose Contriever on two of three benchmarks and provides competitive faithfulness scores. The finding that a parameter-free lexical method surpasses a neural dense retriever trained on hundreds of millions of text pairs challenges the assumption that dense retrieval is universally superior and highlights the importance of domain-specific evaluation rather than reliance on general-domain benchmarks.

5.2. Limitations and Future Directions

This study has several limitations that suggest directions for future investigation. The experiments employ a single generator (GPT-3.5-Turbo), and the relative effectiveness of retrieval strategies may shift with more capable or less capable generators. Extending the comparison to open-source biomedical language models and to more recent proprietary models would enhance the generalizability of the findings. The retrieval depth is fixed at $k = 10$ throughout all experiments; a systematic analysis of the interaction between retrieval strategy and passage count could reveal strategy-specific optimal configurations where certain retrievers benefit from deeper retrieval while others degrade due to noise accumulation.

The evaluation is limited to three English-language benchmarks, leaving open the question of whether these findings transfer to other languages or clinical note-based QA tasks where the vocabulary and syntax differ from published biomedical literature. The faithfulness evaluation relies on the RAGAS metric, which uses an LLM as judge; alternative evaluation approaches incorporating human expert assessment would strengthen the validity of the faithfulness analysis. Future work should also explore learned fusion mechanisms beyond RRF, adaptive retrieval strategies that select the appropriate retriever based on query characteristics, and the integration of structured knowledge sources such as biomedical ontologies alongside unstructured text retrieval. Investigating how retrieval strategies interact with different chunk sizes and passage segmentation methods represents another promising avenue, as the granularity of indexed passages may differentially affect sparse and dense retrieval performance.

References

1. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

2. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamber, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, ... V. Natarajan, "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023.
3. V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.
4. 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, btad651, 2023.
5. K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-augmented language model pre-training," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, PMLR 119, pp. 3929–3938, 2020.
6. G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 874–880, 2021.
7. A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.
8. O. Khattab and M. Zaharia, "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020.
9. G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave, "Unsupervised dense information retrieval with contrastive learning," *Transactions on Machine Learning Research*, 2022.
10. P. T. Chung, "Data Mining Methods for Biomechanical Property Prediction of Biomedical Materials Based on Optimized Feature Dimensionality Reduction," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, pp. 174–180, Dec. 2025.
11. Q. Zhang, "Adaptive Differential Privacy Mechanism for Federated Document Classification: A Gradient-Clipping Optimization Approach," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, pp. 672–678, Dec. 2025.
12. Y. Wang, "Practical AI Approaches for Community Infection Early Warning: From Public Data to Actionable Insights," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, pp. 1545–1552, Dec. 2025.
13. M. Han, "Privacy-Preserving Collaborative Learning Across Healthcare Institutions: An Adaptive Approach with Gradient Compression and Dynamic Privacy Budget Allocation," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, pp. 679–684, Dec. 2025.
14. D. Liang and C. Cai, "Optimizing Large-Scale Contract Review through Data Analytics: Practical Evidence from IPO Audits," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, pp. 242–249, Dec. 2025.
15. L. Gao, Z. Dai, T. Chen, Z. Fan, B. Van Durme, and J. Callan, "Complement lexical retrieval model with semantic residual embeddings," in *Proceedings of the 43rd European Conference on Information Retrieval (ECIR)*, 2021.
16. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
17. 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 (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.
18. D. Jin, E. Pan, N. Oufattole, W. Weng, 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, 6421, 2021.
19. G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al., "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition," *BMC Bioinformatics*, vol. 16, no. 1, 138, 2015.
20. S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.
21. S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL): System Demonstrations*, pp. 150–158, 2024.
22. N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models," in *Advances in Neural Information Processing Systems*, vol. 34 (Datasets and Benchmarks Track), 2021.
23. Y. Li, "Comparative Analysis of Illumination Normalization Methods for Autonomous Driving Under Challenging Lighting Conditions," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, pp. 633–639, Dec. 2025.

24. 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*, pp. 6233–6251, 2024.
25. 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," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, 2022.
26. Y. Wang, "Explainable Risk Stratification for Polypharmacy-Related Adverse Outcomes in Community-Dwelling Elderly: A Rule-Enhanced Machine Learning Approach," *Journal of Sustainability, Policy, and Practice*, vol. 2, no. 2, pp. 18–31, 2026.
27. Y. Li, "Performance Benchmarking and Optimization Strategies for Depth Estimation Algorithms in Unstructured Environments," *Journal of Sustainability, Policy, and Practice*, vol. 2, no. 2, pp. 32–43, 2026.
28. J. Sohn, Y. Park, C. Yoon, S. Park, H. Hwang, M. Sung, H. Kim, and J. Kang, "Rationale-guided retrieval augmented generation for medical question answering," in *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, vol. 1, pp. 12739–12753, 2025.
29. Y. Zhang, "Evaluation of Differential Privacy and Federated Learning for AI-Driven Customer Service Applications," *Journal of Sustainability, Policy, and Practice*, vol. 2, no. 2, pp. 55–66, 2026.
30. P. T. Chung, "Multi-Objective Optimization of Process Parameters for Dental Resin 3D Printing Using Improved NSGA-II Algorithm," *Journal of Science, Innovation & Social Impact*, vol. 2, no. 1, pp. 276–287, 2026.