A Hybrid Approach for Biomedical Question Answering: Combining Sparse and Dense Retrieval with LLM-based Reranking

Pawan Makhija[1*] and Sanjay Tanwani[2]

[*1]Research Scholar, Department of Computer Engineering, Institute of Engineering and Technology, DAVV, Indore, India
[2]Professor, School of Computer Science and IT, DAVV, Indore, India
*[1]pawanmakhija@acropolis.in, [2]sanjay_tanwani@hotmail.com

Abstract. Biomedical question-answering systems typically involve retrieving relevant documents and then reranking them based on their relevance to the query. Traditional sparse retrievers, like BM25 often fail to capture semantic relationships. While the dense retrievers address these limitations, they can also miss relevant documents due to short queries, vocabulary mismatch, and document specificity issues stemming from embeddings. To address these types of challenges, we propose a hybrid approach that consolidates the strengths of both sparse and dense retrieval methods, resulting in better performance. This ensemble approach generates a comprehensive list of candidate documents, which is then passed through a LLM-based reranking model named ColBert, a fine-tuned late interaction mechanism that works on document relevance to refine the ranking of the documents. We use the Flan-T5 answer generation model to produce a final answer to the query. The experiments were performed on the BioASQ dataset, which remarkably demonstrated the effectiveness of our approach and showcased its ability to improve retrieval performance.

Keywords: Biomedical Question Answering, Large Language Models (LLMs), Natural Language Processing, ColBERT Reranking.

1 Introduction

Question answering system often referred to as chatbots is designed to understand and answer questions in a natural language automatically. It holds immense use-cases for various applications specially in information retrieval, customer support and education. In the biomedical domain, it can be used to assist healthcare professionals, researchers and patients in accessing information in a much faster way compared to manually searching in documents. It possesses a unique challenge as it requires highly accurate and precise information. [1, 2].

To address the limitations, we propose a hybrid retrieval approach that leverages the strengths of both and dense retrieval methods to maximize the recall of relevant documents. We then employ an LLM-based re-ranker, specifically ColBERT fine-tuned on document relevance, to refine the initial ranking and prioritize the most relevant documents for answer extraction. Our comparative experiments on the

S. Bhalerao et al. (eds.), *Proceedings of the International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2025),*Advances in Intelligent Systems Research 192,
https://doi.org/10.2991/978-94-6463-716-8_82

BioASQ dataset demonstrate that the proposed BM25-LLMs system achieves superior performance over several existing retrieval models, indicating significant improvements in retrieval effectiveness.

We used Flan-T5 as our answer generation model which is the last part of our Q&A system. We pass the query and the top 2 retrieved contexts as the input to Flan-T5 to get the answer. The quality of the generated answers is evaluated using a unigram f1 score against the ideal answer.

2 Literature Survey

In this section, we are doing a literature review of the existing systems that are useful for understanding the Biomedical question answering system.

2.1 Biomedical Question Answering Systems

Several QA systems have been developed and tailored to fit the requirements of the biomedical domain.

BioASQ [14] offers the benchmarking dataset and sets challenges on different types of biomedical systems thus improving the field. ClinicalBERT [15] uses a transformer model that has been fine-tuned for clinical text to enhance its performance in clinical question answering tasks. Other notable systems include Ebm-NLP [16] which deals with the evidence-based medicine questions and OAQA [17], which employs semantic parsing and knowledge base reasoning to open-domain question answering in the biomedical domain. Recent efforts also utilize KGs for biomedical QA to increase answer accuracy and explainability [18].

2.2 Retrieval Models

Effective retrieval of relevant information is fundamental to biomedical QA. Traditional methods utilize sparse retrieval techniques like TF-IDF [6] and BM25 [7], which rely on term matching and have been widely used in the field. However, these methods may struggle with semantic understanding and complex queries. Dense retrieval methods, such as Sentence-BERT [8] and BioBERT [9], leverage semantic embeddings to capture meaning and relationships between queries and documents, improving performance on complex biomedical questions. Hybrid approaches combining sparse and dense retrieval methods have shown promise in leveraging the strengths of both techniques [32].

2.3 Re-ranking, Answer Generation, and Large Language Models

Re-ranking offers the opportunity to improve search results by optimizing the candidates for answer passage selection. ColBERT [12] works by using a late interaction mechanism between query embedding and document embedding which helped it achieve a State of the art in multiple re-ranking tasks. In the context of answer generation, the Flan-T5 model [13] produced fine answers that were accurate and to the point based on the contexts they collected. Other Large Language Models (LLMs) such as those by GPT-3 [19] and LLaMA [20] have performed reasonably well in complex NLP tasks, including question answering and are currently being examined for more biomedical uses as well. More investigations deal with the issue of further adjustments and training of these LLMs models on the appropriate biomedical tasks to increase performance and reliability.

3 Methodology

This section outlines the methodology and techniques utilized in developing the proposed system. We divided our methodology into three primary parts: document Retrieval, ranking, and answer generation with Flan T5.

3.1 Document Retrieval

The document retrieval stage is crucial as it is the first step in the Question-answering system as it is the first step for the Question-Answering System. We detail the two retrieval methods used in our system in this section.

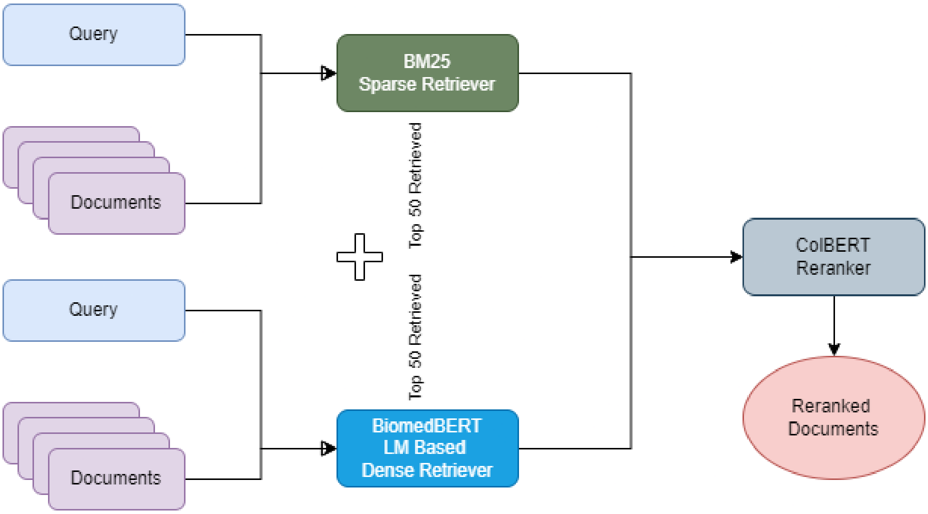


Fig. 1. Document Retrieval System

BM25 (Sparse Retrieval)

In this study, BM25 is employed due to its robustness in exact term matching which is required due to biomedical terminologies (e.g., gene names, drug interactions). It can capture those documents where rare or domain-specific terms appear which may be inadequately captured by dense retrieval models due to insufficient embedding coverage or semantic drift.The BM25 algorithm [7] is based on the probabilistic retrieval framework which ranks documents based on the presence of query terms, with additional consideration for term frequency (TF) and inverse document frequency (IDF). To fairly compare longer and shorter documents it incorporates a document length normalization.

The PubMed abstract corpus is indexed using BM25, with each query yielding a ranked list of documents based on lexical term matches. This approach fetches documents relevant to individual terms.

BioMedBERT (Dense Retrieval)

Dense retrievers capture the semantic meaning of queries, which helps them retrieve documents that are contextually relevant even when synonyms or paraphrases are used. For example, in a query like "effects of BRCA1 mutation on breast cancer" it recognizes the synonymous relationship between "breast cancer" and "carcinoma of the breast."
We utilize BioMedBERT [9] as our dense-retriever as it is a domain-specific variant of BERT pre-trained on PubMed abstracts. Given the pubmed's abstract used in the dataset as the documents[14], this is particularly suitable as it aligns closely with the data being retrieved. This ensures that the model is highly attuned to the semantic nuances of biomedical language, allowing it to perform better than general-purpose language models.
This can handle complex queries involving multiple interacting terms (e.g., age, gender, mutation effects) making it particularly effective for biomedical information retrieval use cases. However, its performance can degrade with very short queries, where sparse models like BM25 may perform better. Thus, combining both retrieval approaches offers complementary strengths.

3.2 Reranking

This section outlines the reranking process implemented after the initial document retrieval. The retrieved document combination is combined for reranking.

ColBERT Reranker

We employ ColBERT [12] at our re-ranker stage because unlike traditional BERT-based re-rankers that independently encode the query and document, ColBERT utilizes a late interaction mechanism where the query and document representations interact at a finer granularity during the scoring process. This allows ColBERT to capture more nuanced relationships between query and document terms, leading to improved ranking accuracy. This design splits the embedding process (done once for the documents) and the interaction (done later at query time). The key advantage of this approach is that it enables deep contextual comparison between the query and documents at the token level, while still being scalable for large corpus.

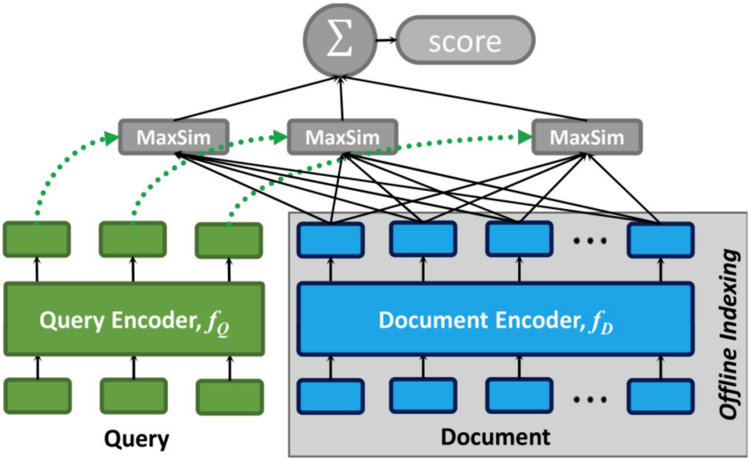


Fig. 2. Late Interaction Mechanism

ColBERT's late interaction architecture is illustrated in Figure 2 and explained in the below steps.

Step1. Document Encoding:

Each document in the corpus is encoded into token-level embeddings using BERT-based architecture. This step happens offline, meaning documents are pre-encoded and stored as token embeddings, reducing computation at query time.

Step2. Query Encoding:

At query time, the query is encoded into token embeddings using the same BERT-based model. This produces embeddings for each token in the query.

Step3. Late Interaction:

ColBERT does not compute a single dense embedding for the document as a whole. Instead, it compares individual token embeddings between the query and the document in a late interaction step. Specifically, for each query token, ColBERT finds the most relevant document token based on a similarity function (e.g., dot product or cosine similarity). This allows ColBERT to focus on highly contextualized matches between the query and the document tokens, emphasizing local matches that matter for ranking.

Step4. Aggregation and Scoring:

After computing the interactions between query tokens and document tokens, the scores are aggregated (typically by max-pooling across tokens) to generate a final ranking score for each document. This token-wise comparison allows ColBERT to identify precise semantic alignments between query and document content, even when they are not explicitly similar on a surface level.

3.3 Answer Generation with Flan-T5

After retrieving and reranking the most relevant documents, the final stage of our pipeline involves generating answers to the query using the Flan-T5 [13] model. This sequence-to-sequence model is well-suited for generating natural language answers, given its strong performance on a wide range of language understanding and generation tasks[13]. Below, we describe the steps involved in the answer generation process and provide a rationale for choosing Flan-T5 for this task.

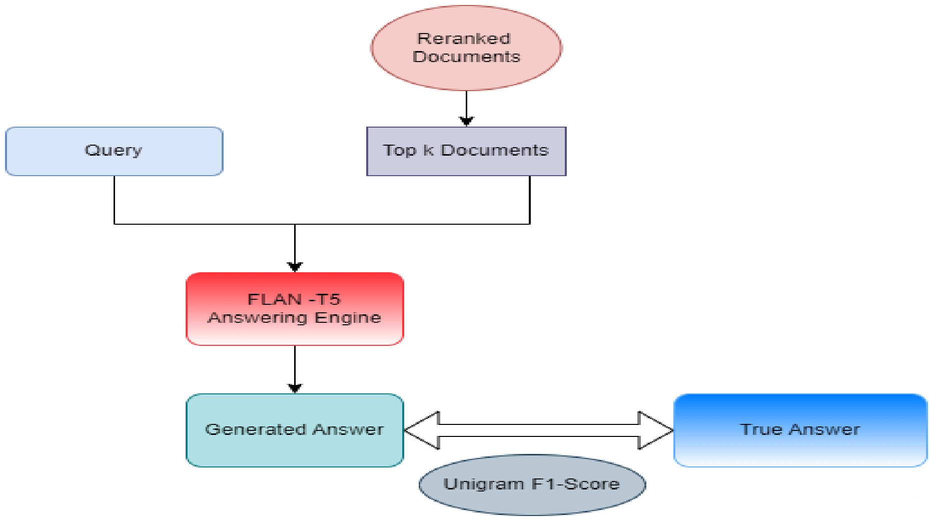


Fig. 3. Answer Generating Mechanism

Answering Mechanism

Once the top-ranked documents are identified from the reranking stage (handled by ColBERT), they are passed, along with the query, to the **Flan-T5** model. The model is tasked with generating a coherent and concise answer based on the provided information.

1. **Input Preparation**:

   The top retrieved documents (usually the top 5 or 10) are concatenated with the original query. This provides the model with the relevant context necessary for generating an accurate answer.

2. **Answer Generation**:

   Flan-T5, a transformer-based model, processes the concatenated query-document input and generates a natural language answer. Leveraging its pre-training on various natural language understanding and generation tasks, Flan-T5 is able to produce answers that are both contextually relevant and linguistically coherent.

4 Dataset & Ground Truth Table

4.1 Dataset used

We have used the BioASQ Task B dataset for assessing the quality of the model. Quality dataset was used for biomedical question answering systems [21]. This dataset consisted of a collection of biomedical questions, PubMed articles, and corresponding ideal and exact answers. There are 4 types of questions namely factoid, summary, yes-no and list was presented in the dataset.

4.2 Creation of Corpus

We have used the Task 11 B from the BioASQ dataset in order to create a corpus of PubMed abstracts by taking the abstracts corresponding to the PubMed IDs provided in the dataset using the Entrez EUtils API [22].

Table 1. Dataset description of the Pubmed abstracts

Dataset Division	YesNo	Factoid	List	Summary	Total
Training	1271	1417	901	1130	4719
Testset 1	8	6	4	7	25
Testset 2	16	12	8	14	50
Testset 3	32	24	16	28	100

4.3 Query Length Analysis

Previous studies [23, 24] have highlighted the challenge of handling short queries in dense retrievers due to limited contextual information. In the context of biomedical retrieval, where queries often vary in length, it becomes critical to understand how query length influences retrieval performance. Based on this, we conducted a query length analysis for the BioASQ dataset (Figure 3) to better inform our hybrid retrieval strategy.

To understand the characteristics of the queries in the BioASQ dataset and their potential impact on retrieval performance, we analyzed the distribution of query lengths. We plotted the frequency of queries against their lengths (measured in the number of words) and observed that the majority of queries fall within a range of 5 to 10 words in length.

This analysis highlights the prevalence of relatively short queries in the dataset. As mentioned earlier, dense retrieval models can be sensitive to short queries due to limited contextual information and potential vocabulary mismatch issues.

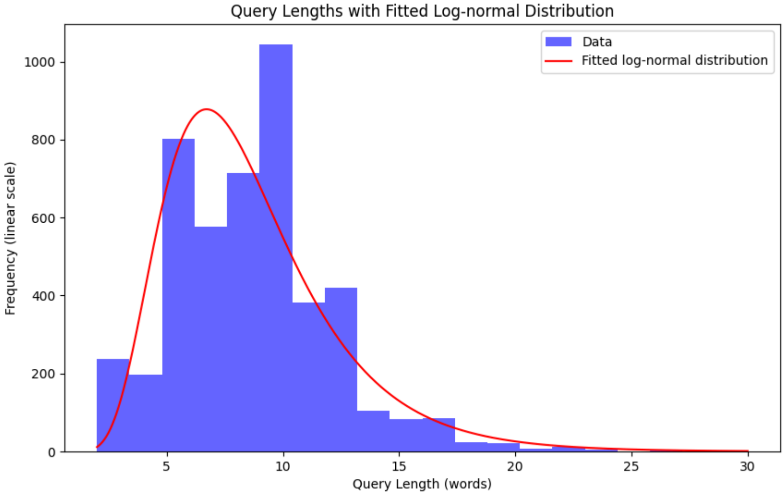


Fig. 4. Query Length Analysis

As observed in the query length frequency analysis (Figure 3), a significant portion of queries in the BioASQ dataset are shorter than 10 words, which may affect PubMedBERT's [9] ability to generate meaningful embeddings. Short queries often lack sufficient context, leading to suboptimal retrieval performance in dense retrievers.

4.4 Ground Truth Table:

To establish a ground truth for evaluation, we created a table mapping each question in the BioASQ dataset to the relevant PubMed abstract IDs and their corresponding exact and ideal answers within our corpus. Some of the examples are shown in Table 2.

Table 2. Question Answer Entry Set from the corpus

Type	Question	Ideal Answer
Factoid	Name synonym of Acrokeratosisparaneoplastica?	Bazex syndrome
Summary	Is Hirschsprung disease a mendelian or a multifactorial disorder?	Coding sequence mutations in RET, GDNF, EDNRB, EDN3, and SOX10 are involved in the development . . .
List	List signaling molecules (ligands) that interact with the receptor EGFR.	epidermal growth factor, betacellulin, epiregulin, heparin-binding epidermal growth factor

Yes/ No	Is the protein Papilin secreted?	Yes

5 Experiments and Results

5.1 Experimental Setup:

Our experiments were conducted on a system with an Intel i7 processor and NVIDIA P100 GPU. We used Python as the programming language and leveraged the Hugging Face Transformers library [25] for seamless model downloading and utilization. We have divided our evaluation into two parts: retrieval and answer generation, which are explained below.

5.2 Evaluating Retrieval Performance:

We focused on evaluating the retrieval performance of the underlying information retrieval component. The retrieval process is crucial as it directly impacts the accuracy and relevance of the final answers provided by the system.

Evaluation Metrics & Baseline:.
We evaluate the performance of our retrieval system using standard information retrieval metrics, including:

a) **Precision:** The proportion of retrieved documents that are relevant.

Precision = (Number of relevant documents retrieved) / (Total number of documents retrieved)

b) **Recall:** The proportion of relevant documents that are retrieved.

Recall = (Number of relevant documents retrieved) / (Total number of relevant documents)

c) **F1 Score:** The harmonic means of precision and recall.

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

d) **Mean Average Precision (MAP):** The average precision across all queries.

$$MAP = \left(\frac{1}{Q}\right) * \sum_{1}^{Q} \frac{1}{m_q} \sum_{k=1}^{m_q} Precision \quad (1)$$

where Q is the number of queries,mq is the documents retrieved
The performance measure is calculated and compared with the state-of-the-art methods used in the challenge and we achieved a remarkable improvement in the evaluation metrics. The table below contains all the evaluation scores of different retrieval methods compared with our proposed approach.

Table 3. Performance comparison of different retrieval methods with the proposed approach

System	Mean Precision	Recall	F-Measure	MAP
TestSet 1				
A&Q2	0.1747	0.5378	0.2069	0.3995
MindLab QA System	0.2820	0.3369	0.2692	0.2631
Bioinfo-0	0.3052	0.6100	0.3381	0.5053
Proposed System	0.3209	0.6870	0.3610	0.6497
TestSet2				
A&Q2	0.1747	0.3728	0.1761	0.2339
MindLab QA System	0.2283	0.2835	0.1876	0.1661
Bioinfo-0	0.2841	0.4993	0.2913	0.4244
Proposed System	0.3477	0.5234	0.4156	0.4956
Testset3				
A&Q2	0.2289	0.4574	0.2236	0.3775
MindLab QA System	0.1600	0.2100	0.1440	0.1312
Bioinfo-0	0.2823	0.4794	0.2808	0.3568
Proposed System	0.3935	0.5134	0.3732	0.4192

Graphical Visualization

To visualize the performance of the systems across all three test sets, Fig. 4 shows a combined bar chart comparing precision, recall, F1-score, and MAP across all systems and test sets.

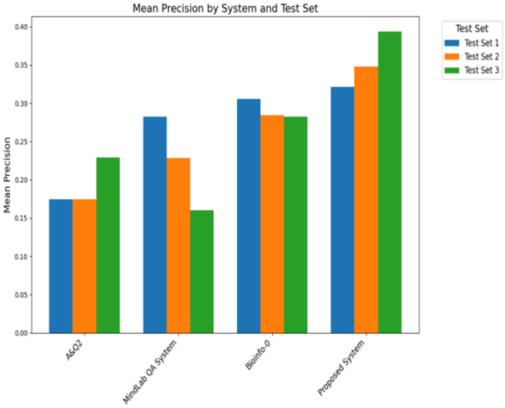


Fig. 5. Comparison of Mean Precision

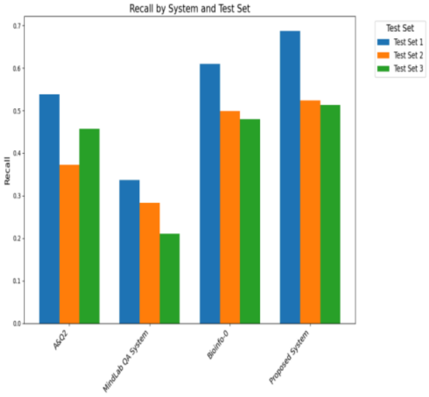
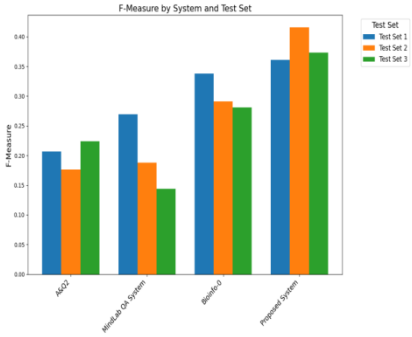


Fig.6. Comparison for Recall



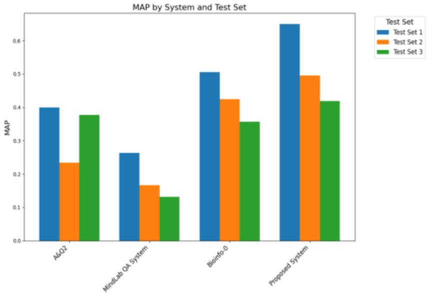


Fig. 7. Comparison for F-Measure

Fig. 8 Comparison for Mean Average Precision

Analysis of the Retrieval system

Our proposed system consistently outperformed baseline systems across all test sets, particularly in terms of Mean Average Precision (MAP) and F1-Score, indicating strong performance in balancing precision and recall.The proposed System achieved the highest F1-Score and MAP across all test sets, demonstrating its superior ability to retrieve relevant documents while minimizing irrelevant ones. It showed significant improvements in Test Set 1 and Test Set 2, with an increase in MAP by over 10% compared to Bioinfo-0, the best-performing baseline. The results align with findings in [28, 29, 30] demonstrating the effectiveness of multi-stage and hybrid retrieval approaches in biomedical QA.

5.3 Answer Generation Performance Evaluation

Experimental Metrics & Baseline

We compare the performance of our hybrid retrieval system against two baselines :

a) BM25 + FLAN-T5[31]: This system uses BM25 [7] to retrieve documents and passes the top k documents to FLAN-T5 for answer generation.

b) ColBERT + FLAN-T5[31]: This system uses ColBERT [12] to retrieve documents, and the top k documents are passed to FLAN-T5.

For both setups, the performance of the generated answers was evaluated using Unigram F1-score, which measures the overlap between the generated answer and the golden (ground-truth) answer.

Results

We found that in our proposed system, performance improved with up to five documents answer generation. After five documents, the F1-score slightly decreased, similar to the findings of the cited research [31].
Table 4 and Fig.9 compares the F1-scores of BM25, ColBERT, and our hybrid retrieval method, showing the superiority of our approach. We write approx as the author[31] showed the graph but not the table.

Table 3. performance of our hybrid retrieval system against two baselines

Retrieval Method	Generating Engine	k	Unigram F1 Score (%)
BM25[31]	FLAN-T5	5	31.0 (approx.)
ColBERT[31]	FLAN-T5	5	35.0 (approx.)
(BM25 + BiomedBERT) &ColBERT(Reranker)	FLAN-T5	5	40.35

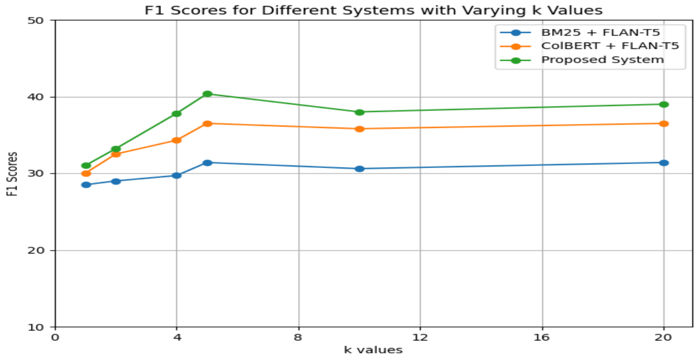


Fig. 9. Performance comparison of various systems with varying k values

Discussion
The results in Table 3 illustrate a clear trend where using ColBERT improves answer generation quality over BM25, though the improvements are modest (2–5 points) as reported in the paper [31]. Specifically, ColBERT slightly outperforms BM25 by a

2–5 point margin in the biomedical domain when used in conjunction with FLAN models.

However, our hybrid retrieval system followed by reranking further enhances performance. For *k=5*, our system achieved a 5-point improvement in Unigram F1-score compared to the ColBERT + FLAN-T5 baseline, reaching an F1-score of 40.35%. This can be attributed to two factors:

a) The combination of BM25 and BioMedBERT helps retrieve more contextually relevant documents.

b) The reranking step using ColBERT ensures that the most semantically appropriate documents are selected, which is critical in specialized domains like biomedicine.

Hsia et al. [31] also suggested that the biomedical domain's specificity can amplify the differences in downstream tasks such as answer generation. Our findings confirm this, as the additional use of PubMedBERT in document retrieval likely contributed to the enhanced relevance of retrieved documents, which in turn positively impacted the generated answers.

Thus, while both BM25 and ColBERT perform similarly in biomedical retrieval tasks, the integration of our proposed hybrid retrieval system followed by ColBERT reranking and passed to FLAN-T5 demonstrates a marked improvement in generating accurate answers.

6 Conclusion

Biomedical question-answering systems are an active area of interest due to their huge potential for revolutionizing the medical system. For this, we proposed a hybrid retrieval system that integrates the strengths of both sparse and dense retrieval methods, combined with a reranking and answer generation pipeline. This effectively balances the need for accurate term retrieval with the ability to capture nuanced semantic relationships.

Our experimental results, evaluated on the BioASQ dataset, demonstrated that this hybrid retrieval method showed better results than traditional methods across standard metrics, including Precision, Recall, F1-Score, and Mean Average Precision (MAP). For answer generation, our system showed significant improvements, particularly when compared to other document retrieval approaches such as BM25 or ColBERT alone.

Our approach underscores the importance of combining multiple retrieval techniques in biomedical information retrieval tasks, where the complexity and specificity of language demand both lexical and contextual understanding. Our findings suggest that hybrid retrieval systems, enhanced by advanced reranking models like ColBERT and powerful generative models like FLAN-T5, can provide a more accurate and reliable solution for biomedical question-answering, ultimately aiding researchers and clinicians in efficiently accessing and synthesizing critical information from vast biomedical corpora.

References

1. Gudivada, V. N., et al. "Information retrieval on the world wide web." IEEE internet computing 1.5 (1997): 58-68.
2. Hersh, W. "Information retrieval: A health and biomedical perspective." Springer Science & Business Media, 2009.
3. Nenkova, A., and McKeown, K. "Automatic summarization." Foundations and Trends® in Information Retrieval 5.2–3 (2012): 103-233.
4. Liu, S., et al. "A survey of deep learning methods for relation extraction." arXiv preprint arXiv:2012.00938 (2020).
5. Lin, J., et al. "A survey of retrieval models in document-grounded conversation systems." arXiv preprint arXiv:2207.06160 (2022).
6. Salton, G., and Buckley, C. "Term-weighting approaches in automatic text retrieval." Information processing & management 24.5 (1988): 513-523.
7. Robertson, S. E., and Zaragoza, H. "The probabilistic relevance framework: BM25 and beyond." Foundations and Trends® in Information Retrieval 3.4 (2009): 333-389.
8. Reimers, N., and Gurevych, I. "Sentence-BERT: Sentence embeddings using Siamese BERT-networks." EMNLP-IJCNLP 2019.
9. Lee, J., et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.
10. Teflioudi, C., et al. "Joint Learning of Semantic Representations and Keyword Weights for Search." IEEE Transactions on Knowledge and Data Engineering (2023).
11. Cormack, G. V., Clarke, C. L. A., and Buettcher, S. "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods." Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 2009.
12. Khattab, O., and Zaharia, M. "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020.
13. Chung, H. W., et al. "Scaling instruction-fine tuned language models." arXiv preprint arXiv:2210.11416 (2022).
14. Tsatsaronis, G., et al. "BioASQ: A challenge on large-scale biomedical semantic indexing and question answering." AAAI Workshop on Semantic Search. 2012.
15. Huang, K., et al. "ClinicalBERT: Modeling clinical notes and predicting hospital readmission." ACL 2019.
16. Nye, B., et al. "A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for evidence-based medicine." LREC. 2018.
17. Abacha, A. B., et al. "Overview of the OAQA question answering system." Proceedings of the Text Analysis Conference. NIST, 2015.
18. Fecho, et al. Biomedical Data Translator Consortium. (2023). An approach for collaborative development of a federated biomedical knowledge graph-based question-answering system: Question-of-the-Month challenges. Journal of Clinical and Translational Science, 7(1), e214.
19. Brown, T. B., et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
20. Touvron, H., et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).

21. BioASQ official website. (https://www.bioasq.org/)
22. NCBI Entrez Programming Utilities Help. (https://www.ncbi.nlm.nih.gov/books/NBK25501/)
23. Izacard, G., & Grave, E. (2020). Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint arXiv:2007.01282.
24. Sharma, S., & Panda, S. P. (2023). Efficient information retrieval model: overcoming challenges in search engines-an overview. Indonesian Journal of Electrical Engineering and Computer Science, 32(2), 925-932.
25. Wolf, T., et al. "Huggingface's transformers: State-of-the-art natural language processing." arXiv preprint arXiv:1910.03771 (2019).
26. Manning, C. D., Raghavan, P., & Schütze, H. Introduction to information retrieval. Cambridge university press, 2008.
27. Baeza-Yates, R., & Ribeiro-Neto, B. Modern information retrieval. ACM press New York, 1999.
28. Shin, A.; Jin, Q.; Lu, Z. Multi-stage literature retrieval system trained by PubMed search logs for biomedical question answering. In Proceedings of the Conference and Labs of the Evaluation Forum (CLEF), Thessaloniki, Greece, 18–21 September 2023.
29. Rosso-Mateus, A.; Muñoz-Serna, L.A.; Montes-y Gómez, M.; González, F.A. Deep Metric Learning for Effective Passage Retrieval in the BioASQ Challenge. In Proceedings of the CLEF 2023: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 18–21 September 2023.
30. Almeida, T.; Jonker, R.A.A.; Poudel, R.; Silva, J.M.; Matos, S. BIT.UA at BioASQ 11B: Two-Stage IR with Synthetic Training and Zero-Shot Answer Generation. In Proceedings of the Conference and Labs of the Evaluation Forum, Bologna, Italy, 5–8 September 2022.
31. Hsia, J., Shaikh, A., Wang, Z., & Neubig, G. (2024). RAGGED: Towards Informed Design of Retrieval Augmented Generation Systems. arXiv preprint arXiv:2403.09040.
32. Chen, S., et al. "Hybrid Retrieval-Augmented Generation for Biomedical Question Answering with PubMed Evidence." Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023.