iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.83121

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Hybrid Adaptive Retrieval-Augmented Generation for Context Relevance in Biomedical Question Answering

N. Devi[1], P. LeelaRani[2], V. Vidhya[3], D. Jayanthi[4], AR. GuruGokul[5]
Department of Information Technology, Sri Venkateswara College of Engineering, Sriperumbudur-602117, Tamil Nadu, India

Abstract: ***Retrieval-Augmented Generation (RAG) has emerged as an effective paradigm for improving the factual accuracy of large language models by grounding responses in external knowledge. This paper proposes a Hybrid Adaptive RAG framework that combines sparse (BM25) and dense retrieval techniques to enhance context relevance in biomedical question answering using the PubMedQA. The proposed approach dynamically integrates lexical and semantic retrieval strategies to improve document ranking and retrieval quality. Experimental results demonstrate that dense retrieval achieves the highest performance, while the hybrid model significantly outperforms BM25. Although Exact Match scores remain low due to the abstractive nature of answers, the system produces contextually grounded responses with high faithfulness. The findings highlight the effectiveness of hybrid retrieval in improving RAG performance while identifying the need for advanced adaptive mechanisms and improved generation models.***

Keywords: Retrieval-Augmented Generation (RAG), Large Language model (LLM), Information retrieval, biomedical question answering.

I. INTRODUCTION

Recent works in biomedical and digital health records have curated a crucial necessity for smart systems that have the ability to retrieve and orchestrate the appropriate information in an efficient manner. Commonly used information retrieval methods, such as keyword-based search, often stumble to capture the semantic complexity of medical searches, producing results that are either irrelevant or insufficient. Simultaneously, large language models (LLMs) have exhibited extraordinary abilities in both natural language comprehension and generation; however, they are susceptible to hallucinations and do not have access to current or specialised knowledge. This crafts a precarious challenge in creating systems that are both precise and contextually relevant, especially in critical areas such as healthcare.

Retrieval-Augmented Generation (RAG) has emerged as a favourable paradigm to cater for the above challenge by amalgamating exterior knowledge retrieval with generative models. Within a RAG framework, a retriever initially locates pertinent documents from an extensive corpus, after which a generator formulates responses based on the retrieved context. This amalgamation facilitates models to provide more realistic, understandable, and context-aware outputs analogous to standalone LLMs. Nevertheless, despite these benefits, the efficacy of RAG systems is significantly reliant on the calibre of the retrieval component. Current approaches normally depend on either sparse retrieval methods, such as BM25, which outshine at keyword matching, or dense retrieval methods, which capture semantic resemblance. Even though both approaches have intrinsic hassles when used self-sufficiently, especially in addressing intricate biomedical inquiries that necessitate both lexical accuracy and semantic comprehension.

To eradicate these hussles, this paper emphasises incorporating the hybrid adaptive retrieval mechanism within the RAG architecture. The proposed approach seeks to dynamically amalgamate both sparse and dense retrieval strategies in order to enhance the relevance and ranking of the documents retrieved. By leveraging the complementary strengths of both retrieval paradigms, the hybrid model seeks to enhance the overall performance of RAG systems in terms of retrieval accuracy and downstream answer generation quality. This is particularly crucial in tasks comprising biomedical question answering, where accurate information that fits the context is vital. In summary, this paper emphasises the curbs of the available retrieval methods in RAG systems by proposing an adaptive hybrid approach that enhances the context pertinence. The amalgamation of such a mechanism has the capacity to considerably augment the consistency and effectiveness of RAG-based solutions in practical, knowledge-driven applications.

©IJRASET: All Rights are Reserved

II. LITERATURE REVIEW

Recent progress in Retrieval-Augmented Generation (RAG) has greatly enhanced the ability of LLMs to generate responses that are both contextually pertinent and empirically precise. RAG amalgamates exterior knowledge retrieval with generative models, thus reducing false information and enhancing trustworthiness. Most of the recent studies have investigated the various facets of RAG, encompassing architectural design, adaptive retrieval methods, evaluation frameworks, and applications tailored to specific domains.

Comprehensive surveys have provided foundational insights into RAG systems. Zhao et al. [1] present a comprehensive synopsis of RAG techniques, emphasising the significance of retrieval superiority in augmenting cohort performance. Similarly, Sharma [2] categorises RAG architectures and talks about the difficulties of robustness, highlighting the importance of flexible and mixed retrieval methods.

New research emerged involving dynamic and adaptive retrieval. Su et al. [3] presented DRAGIN, a framework that actively assesses information retrieval necessities while generating content, presenting enriched efficacy in various tasks that necessitate extensive knowledge. GARAG [7] and RAFT [8] are the additional frameworks that present adaptive retrieval and tuning specific to domains, demonstrating that fixed retrieval processes are inadequate for intricate queries.

Another important area that is enormously discussed in the literature is retrieval structure and relevance. Wang et al. [6] proposed topology-aware retrieval mechanisms to boost contextual rationality, while FeB4RAG [5] investigates federated retrieval methods for decentralised knowledge repositories. This literature points out the significance of structured and multi-source retrieval strategies in augmenting the performance of RAG.

Another curb of the RAG system that needs to be discussed extensively is how to evaluate the systems. Common metrics that are used in LLMS when used to assess the RAG systems often miss to analyse the interaction between retrieval and generation. RAGAS, proposed by Es et al. [9], presents an automated evaluation framework that emphasises fidelity and the relevance of answers. Yu et al. [4] further examine the limitations of evaluation metrics and suggest a new benchmarking strategy for RAG systems.

Many domain-specific applications adopting RAG have demonstrated the efficacy of RAG in real-world scenarios. PaperQA [11] employ RAG for scientific literature analysis, accomplishing resilient performance in multifaceted question-answering tasks. Enterprise and biomedical applications [14] have been developed to demonstrate the application of RAG in decision support systems, where accuracy and explainability are critical.

Furthermore, improvements in LLMs and prompting techniques have indirectly aided in enhancing RAG. Models such as LLaMA [12] and reasoning techniques like Chain-of-Thought prompting [13] boost the generative element of RAG systems. Lack of adaptive fusion, and evaluation limitations are some of the open challenges ascertained by consolidating current trends in RAG research [10][15].

Despite these progressions, present RAG systems continue to hurdle snags when trying to amalgamate sparse and dense retrieval techniques commendably. Most of the approaches presented in the literature rely on static or heuristic-based fusion strategies, which limit their capacity to acclimate to fluctuating query categories. This breach persuades the requisite for a hybrid adaptive retrieval mechanism that dynamically amalgamates numerous retrieval strategies to enhance context relevance and overall system performance.

III. PROPOSED HARAG FRAMEWORK

This paper proposes a Hybrid Adaptive Retrieval-Augmented Generation (HARAG) framework intended to enhance context pertinence and answer generation features in the domain of biomedicine. The framework amalgamates both sparse and dense retrieval techniques with a generative language model, as depicted in figure 1, empowering the system to simultaneously influence lexical precision and semantic understanding.

The proposed HARAG system constitutes six major modules, viz., Data Preparation,Sparse Retrieval,Dense Retrieval, Hybrid Adaptive Fusion, Answer Generation, and Evaluation. The workflow initiates with the preprocessing of the dataset and the formation of a document corpus, which is subsequently followed by parallel retrieval employing BM25 and dense embeddings. The retrieved results are then fused together through an adaptive weighting mechanism given by lin to generate the final context, which is then passed to a language model for answer generation. The assessment of the final retrieved documents is done through metrics for retrieval and generation

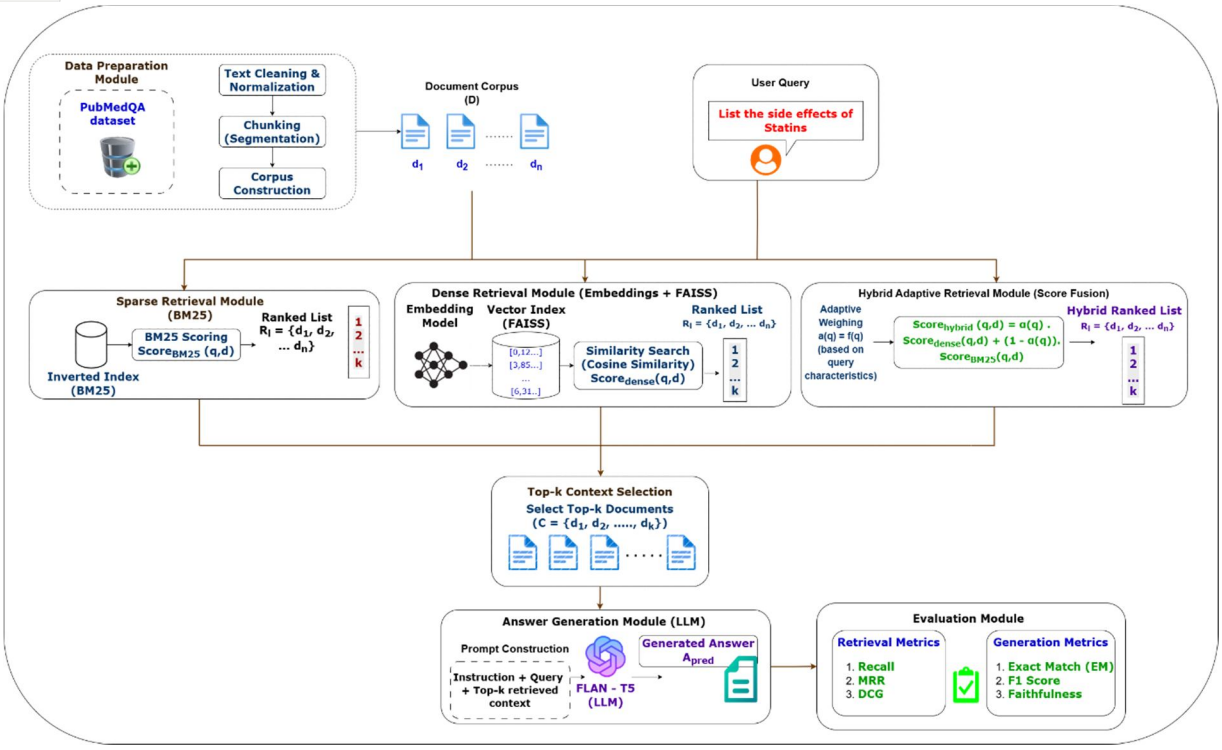


Figure 1 Hybrid Adptive Retrieval-Augmented Architecture

1) Data Preparation

The data preparation module hypothesises a retrieval-ready corpus from the PubMedQA by fetching questions, contexts, and their corresponding answers. Every context document undergoes a data preparation pipeline that comprises normalisation, breaking into tokens, and dividing into smaller parts to enhance retrieval detail. Formally, the corpus is characterised as set D and Q as defined in equations 1 and 2.

$$D=\{d_1,d_2,d_3,\ldots d_n\} \quad (1)$$

$$Q=\{q_1,q_2,q_3,\ldots q_m\} \quad (2)$$

where (D) symbolizes the set of document chunks and (Q) signifies the set of queries. Each document (d_i) is further converted into a sequence of tokens as shown in equation 3:

$$di=\{t_1,t_2,t_3,\ldots t_k\} \quad (3)$$

This organized format consents for effective indexing and retrieval in later modules.

2) Sparse Retrieval

The sparse retrieval module applies the BM25 scoring method to assess how pertinent a document is, considering the frequency of terms and the inverse frequency of documents. Given a query (q) and document (d), the BM25 score is calculated using the formula given in equation 4:

$$Score_{BM25(q,d)} = \sum_{t\in q} IDF(t).\frac{f(t,d)(k_1+1)}{f(t,d)+k_1(1-b+b.\frac{|d|}{avgdl}} \quad (4)$$

where (f(t,d)) denotes the frequency of term (t) in document (d), (|d|) is the document length, and (avgdl) is the average document length in the corpus. The variables k_1 and b regulate term frequency saturation and extent normalization. This module is principally effective in apprehending precise keyword matches and lexical patterns.

3) Dense Retrieval

The dense retrieval component translates both queries and documents into incessant vector representations employing a transformer-based embedding model. Each document (d) and query (q) are mapped into a vector space as given in equation 5:

$$v_d = f(d)\ ,\ v_q=f(q) \quad (5)$$

where (f(.)) is the embedding function. The similarity between query and document is computed using cosine similarity as given in equation 6.

$$Score_{dense(q,d)} = \frac{v_q . v_d}{\|v_q\| \|v_d\|} \quad (6)$$

This approach captures intrinsic semantic interactions beyond precise word matching, enhancing recall for contextually pertinent documents.

4) Proposed Hybrid Adaptive Retrieval

The proposed module amalgamates the scores computed from sparse and dense retrieval modules utilising an adaptive weighting mechanism. Instead of trusting on a static combination, the model dynamically regulates the influence of every retrieval method based on query features. The hybrid score is calculated as given in the equation 7:

$$Score_{hybrid(q,d)} = \propto (q) . Score_{dense(q,d)} + (1 - \propto (q)) . Score_BM25(q, d) \quad (7)$$

where $\propto (q) \in [0,1]$ is a query-dependent weighting function. In this work, $\propto (q)$ is heuristically defined based on query length as shown in equation 8:

$$\propto (q) = \begin{cases} 0.7, & if\ |q| < \tau \\ 0.4, & otherwise \end{cases} \quad (8)$$

where (|q|) is the number of tokens in the query and τ is a predefined threshold. This adaptive fusion leverages both exact word choice and meaning comprehension, enhancing the total effectiveness of retrieval.

5) Answer Generation Using LLM

The answer generation module functions based on LLM model that has been adjusted for instructions, like Flan-T5, to craft replies that are aware of the context. The model crafts an answer C accustomed on the query (q) and the top-(k) retrieved documents as given in the equation 9:

$$C=\{d_1,d_2,d_3,..d_k\} \quad (9)$$

The cohort process is established using the probability formula as given in the equation 10:

$$A_{pred} = \arg\max_A P\langle A|q.C\rangle \quad (10)$$

where A_{pred} is the predicted answer. An organized prompt is employed to coerce the model to craft answers chastised in the retrieved context, thereby decreasing hallucinations.

6) Evaluation Module

The evaluation module inspects both retrieval and generation efficacy by utilizing standard measures. Retrieval effectiveness is measured as Recall@k using the formula given in equation 11.

$$Recall@k = \frac{|Relevant\ Documents \cap Retrieved\ Documents|}{Relevant\ Documents} \quad (11)$$

Mean Reciprocal Rank (MRR) is mathematically calculated using the equation 12.

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank_i} \quad (12)$$

where $rank_i$ is the position of the first relevant document. The normalized Discounted Cumulative Gain (nDCG) evaluates ranking quality which is computed using the equation 13.

$$nDCG = \frac{DCG}{IDCG} \quad (13)$$

For generation, Exact Match (EM) is computed using the equation 14:

$$EM = \begin{cases} 1, & if\ A_{pred} = A_{true} \\ 0, & otherwise \end{cases} \quad (14)$$

The F1 score measures token overlap is computed using the equation 15.

$$F1 = \frac{2.Precision.Recall}{Precision+Recall} \quad (15)$$

Finally, faithfulness is estimated as shown in equation 16.

$$Faithfulness = \frac{|Tokens(A_{pred}) \cap Tokens(C)|}{|Tokens(A_{pred})|} \quad (16)$$

which measures the extent to which the generated answer is supported by retrieved context.

The proposed methodology amalgamates adaptive hybrid retrieval with LLM-based generation, empowering enhanced context relevance and answer eminence. The mathematical formulation guarantees intelligibility and reproducibility, while the modular design facilitates extensibility for future enhancements.

IV. DATASET DESCRIPTION

The PubMedQA[16] is a benchmark dataset intended for evaluating question answering systems in the biomedical domain. The dataset is curated based on real queries acquired from biomedical research and clinical scenarios, intended at assessing systems that reply to questions within the biomedicine domain, with substantiating evidence sourced from PubMed abstracts. Each and every instance instituted the dataset consists of a question, a set of relevant context documents, and a corresponding answer annotated by domain experts. The dataset is particularly challenging because it requires models to perform multi-document reasoning over complex scientific texts, rather than relying on simple keyword matching.

In addition to short categorical answers (Yes/No/Maybe), PubMedQA also offers long-form explanatory answers that recapitulate the rationale inferred from the supporting documents. Due to this analogy, this dataset makes it extremely qualified for assessing Retrieval-Augmented Generation (RAG) systems, where both response generation and its quality are critical. Semantic conception and contextual amalgamation are essential as each question's context usually consists of several abstracts with experimental findings, clinical observations, and scientific conclusions.. As the available documents are lengthy and noisy, the dataset curbs certain significant limitations for retrieval models, especially in ascertaining the most pertinent facts. Moreover, the concise answer to the queries makes the assessment process more critical as exact answers are very infrequent and fractional overlays may be deliberated. Table 1 provides the detailed statistical description about the data available in PubMedQA.

Table 1: PubMedQA Dataset Description

Statistics	Value
Total Sample (Labeled)	1000
Total Sample (UnLabeled)	61000
Train Set Size	1000
Avg. Question Length	10-15 words
Avg. Context Length(per Document)	150-250 words
Avg. No. of Context per Question	5-10 documents
Avg Answer Length	50-150 words
Answer Type	Yes/No/Maybe
Domain	Biomedical

V. EXPERIMENTAL RESULTS

In order to assess the efficacy of the proposed model, a document corpus is augmented using the PubmedQA dataset. The information pertaining to the dataset is listed in the table. Input queries are framed using the questions available in the dataset, the set of answers with associated context forms the response space and the detailed answers are served as ground truth for evaluating cohort performance using Recall@K, MRR, nDCG@K, EM, F1 score and faithfulness. The table 2 presents the performance comparison of the proposed Hybrid Adaptive Retrieval-Augmented Generation (RAG) model with other models viz., BM25 and Dense model.

Table 2: Performance Comparison of RAG Models

Model	Recall@k	MRR	nDCG@k	EM	F1	Faithfulness
BM25	0.533571	0.846667	0.846667	0.0	0.201699	0.983122
Dense	0.779405	0.98	0.961297	0.0	0.223115	0.981212
Hybrid(proposed)	0.699405	0.93	0.923127	0.0	0.213373	0.981143

From the table2, we can infer that response generation for the PubMedQA dataset through a dense model shows a comparatively higher performance than the other two. However, the proposed hybrid approach spectacles a remarkable enhancement over BM25, ratifying that combining lexical and semantic retrieval augments context significance. Conversely, it does not outperform dense retrieval, comprehending that the existing adaptive weighting strategy necessitates further improvement. In terms of generation performance, all models exhibit low Exact Match (EM) scores due to the abstractive nature of answers in the dataset, while moderate F1 scores indicate partial overlap between generated and ground truth responses. Faithfulness scores remain consistently high across all models, indicating that the generated answers are well grounded in the retrieved context.
In order to appraise the efficacy of the proposed system, a sample input–output instance is explored and presented in the figure 2.

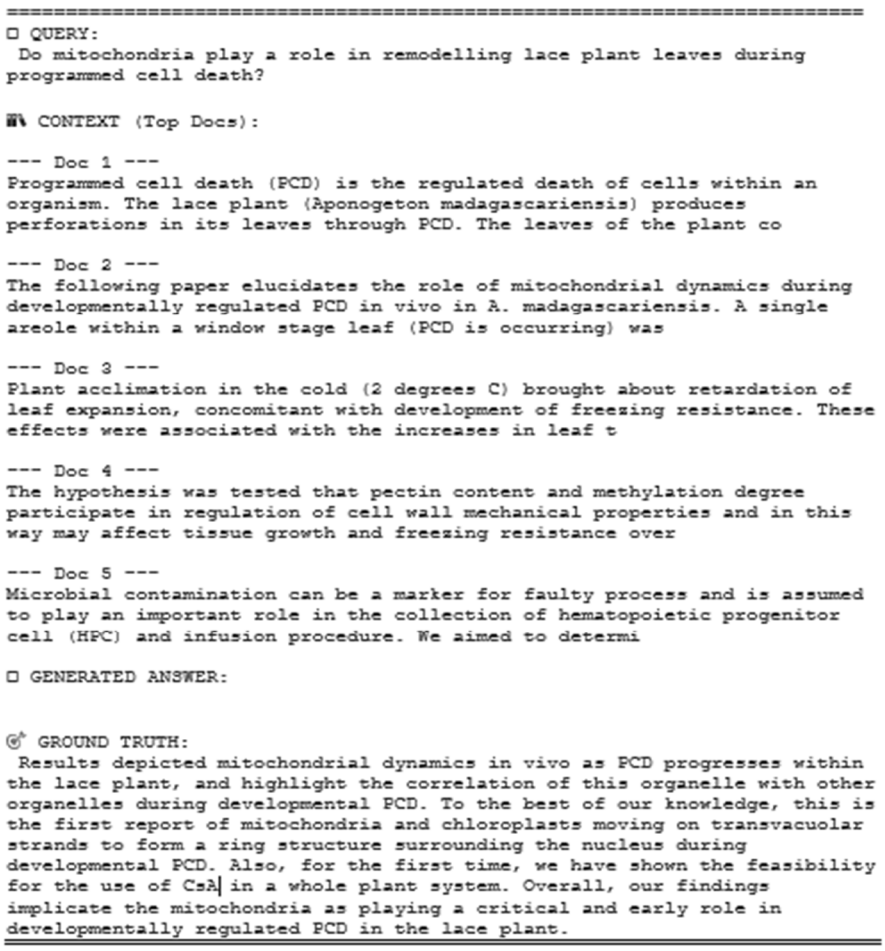


Figure 2 Sample Response generated by our proposed RAG model

From the figure 2, it is evident that the documents with key biomedical information associated with the requested topic are retrieved. This illustrates that the proposed method commendably acquires the pertinent data. Thus, the generated answer incarcerates the principal theory of the ground truth, illustrating how clear and relevant answers are curated by the proposed RAG system. However, variances in comprehensiveness and specificity are still observed, as the curated answer may be skewed to only certain specifics and offer a shortened elucidation of the evidence. This curb highlights the craving for the progression in both the accuracy of fetching evidence and the ability of the language model to conglomerate comprehensive biomedical details.

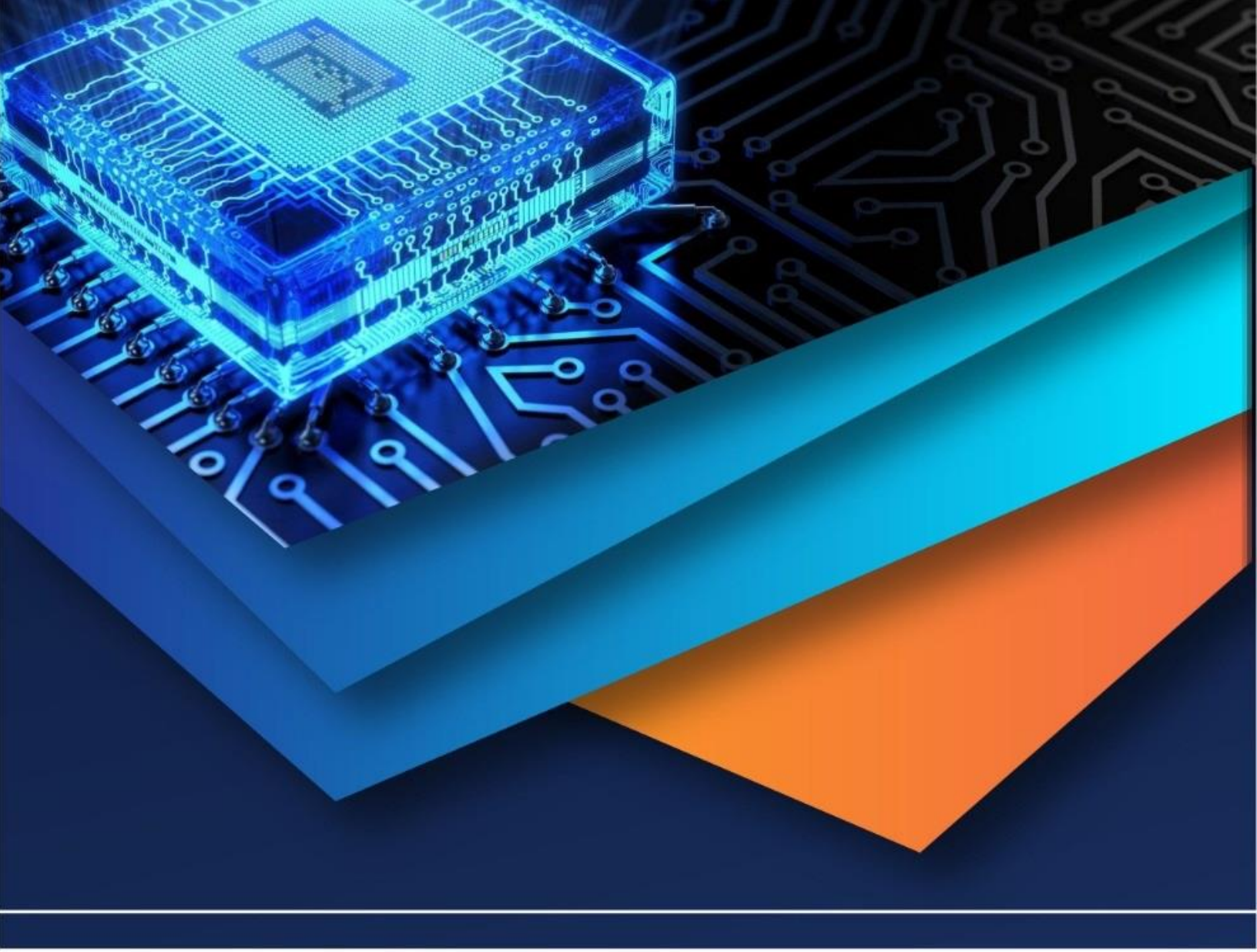
VI. CONCLUSION

In this paper, a Hybrid Adaptive Retrieval-Augmented Generation (RAG) framework is proposed to progress context significance in biomedical question answering using the PubMedQA. Experimental results depict that dense retrieval attained the superlative performance, while the proposed hybrid method enhanced over BM25. Though Exact Match scores endured very low due to the abstractive nature of answers, the system curate contextually grounded responses with high faithfulness. Overall, the proposed work illustrates that hybrid retrieval augments performance, but further optimisation is also required to outperform dense retrieval and enhance answer completeness. Future research will focus on adaptive learning methods for regulating weights and cross-encoder reordering to enhance the efficacy of curating the evidence. Moreover, more powerful LLMs are incorporated with domain-specific fine-tuning to enrich the accuracy and completeness of the curated answers.

REFERENCES

[1] P. Zhao, H. Liu, Y. Chen, Z. Wang, and X. Zhang, “Retrieval-Augmented Generation for AI-Generated Content: A Survey,” Data Science and Engineering, 2026.
[2] C. Sharma, “Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers,” arXiv preprint arXiv:2506.00054, 2025.
[3] W. Su, X. Han, Z. Lin, P. Yu, and Z. Sun, “DRAGIN: Dynamic Retrieval Augmented Generation based on the Real-time Information Needs of Large Language Models,” in Proc. ACL, 2024.
[4] Y. Yu, H. Wang, X. Liu, and J. Li, “Evaluation of Retrieval-Augmented Generation: A Survey,” in Proc. ICCBD, 2024.
[5] S. Wang, R. Zhang, L. Liu, and K. Chen, “FeB4RAG: Evaluating Federated Search in Retrieval-Augmented Generation,” in Proc. SIGIR, 2024.
[6] Y. Wang, Z. Li, H. Zhang, and Q. Liu, “Topology-aware Retrieval Augmentation for Text Generation,” in Proc. CIKM, 2024.
[7] Z. Wei, Y. Chen, H. Liu, and X. Zhang, “GARAG: Adaptive Question Answering using Retrieval-Augmented Generation,” in Proc. ICCBD, 2024.
[8] T. Zhang, X. Liu, J. Wang, and Y. Li, “RAFT: Adapting Language Models to Domain-Specific Retrieval-Augmented Generation,” arXiv preprint arXiv:2403.10131, 2024.
[9] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAS: Automated Evaluation of Retrieval Augmented Generation,” arXiv preprint arXiv:2309.15217, 2023.
[10] A. Brown, D. Green, M. Taylor, and S. Wilson, “A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges,” arXiv, 2025.
[11] J. Lála, A. Mallen, S. Asai, and H. Hajishirzi, “PaperQA: Retrieval-Augmented Generative Agent for Scientific Research,” arXiv preprint arXiv:2312.07559, 2023.
[12] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, and others, “LLaMA: Open and Efficient Foundation Language Models,” arXiv preprint arXiv:2302.13971, 2023.
[13] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” in Proc. NeurIPS, 2022.
[14] R. Kumar, S. Patel, A. Singh, and V. Sharma, “Retrieval-Augmented Generation and LLMs for Enterprise Knowledge Management,” Applied Sciences, 2026.
[15] 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,” arXiv preprint arXiv:2005.11401.
[16] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, “PubMedQA: A Dataset for Biomedical Research Question Answering,” in Proc. EMNLP-IJCNLP, 2019, pp. 2567–2577.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)