Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

CancerAutoRAG: A Self-Updating Retrieval-Augmented Generation Framework for Oncology with Hybrid Retrieval and Re-ranking

Aashi Chouksey
Dept. of Computer Science & Engineering
Prestige Institute of Engg. Mgmt. & Research
Indore, India

Dr. Amita Jain
Dept. of Computer Science & Engineering
Prestige Institute of Engg. Mgmt. & Research
Indore, India

Abstract - Cancer remains one of the leading causes of mortality worldwide, generating vast volumes of clinical literature, trial results, genomic datasets, and treatment guidelines that evolve at a pace no clinician can track manually. Existing Retrieval-Augmented Generation (RAG) systems suffer from knowledge staleness and retrieval imprecision—limitations that are especially dangerous in oncology. In this paper, we present **CancerAutoRAG**, a self-updating RAG framework purpose-built for oncology. CancerAutoRAG integrates event-driven incremental ingestion of oncology documents (PubMed abstracts, NCCN/ESMO guidelines, ClinicalTrials.gov summaries, institutional pathology reports), a hybrid retrieval module combining FAISS dense search with BM25 fused via Reciprocal Rank Fusion (RRF), and a cross-encoder re-ranking stage fine-tuned on clinical passage-ranking data. On a curated oncology benchmark of 1,200 annotated question-answer pairs, CancerAutoRAG achieves **91% accuracy** with a query latency of 275 ms, outperforming vanilla RAG (68%) by 23 percentage points while remaining 145 ms faster than Hybrid RAG. These results demonstrate that CancerAutoRAG is both highly accurate and fast enough for real-time clinical decision support.

Index Terms—*Cancer informatics, oncology decision support, Retrieval-Augmented Generation, hybrid retrieval, re-ranking, clinical NLP, BM25, FAISS, cross-encoder*

I. INTRODUCTION

Cancer is a disease that does not wait. A clinician managing a patient with advanced non-small cell lung cancer (NSCLC) must simultaneously track molecular subtyping results, interpret progression imaging, cross-reference the latest immunotherapy trial data, and reconcile conflicting guideline updates from NCCN, ESMO, and ASCO—all within a single consultation. PubMed indexes more than 700,000 new cancer-related publications every year, and major treatment guidelines are revised multiple times annually as landmark trial results emerge.

Large language models (LLMs) such as GPT-4 [1] and LLaMA [2] have shown remarkable capability in biomedical question answering. However, these models are frozen at their training cutoff and cannot access findings from trials published afterward. In oncology, where a single Phase III trial can reshape first-line therapy overnight, this temporal blindness is a patient-safety concern. Moreover, LLMs are known to hallucinate—generating plausible-sounding but factually wrong information [3]—a failure mode particularly dangerous when the fabricated fact concerns a drug dosage, contraindication, or survival statistic.

Retrieval-Augmented Generation (RAG), introduced by Lewis et al. [4], addresses hallucination by grounding LLM responses in retrieved documents from a curated corpus. While promising, existing RAG deployments for clinical use face two unresolved challenges: (i) static knowledge bases that require time-consuming batch re-indexing, creating windows where newly published trial results are invisible; and (ii) retrieval quality insufficient for the specialised vocabulary of precision oncology—rare drug names, gene symbols (KRAS G12C, EGFR exon 19), and protein identifiers [5].

This paper introduces CancerAutoRAG, which resolves both problems simultaneously. The system continuously monitors oncology data sources—PubMed RSS feeds, ClinicalTrials.gov updates, NCCN guideline archives, and de-identified institutional pathology reports—and indexes only changed or newly added documents within seconds. The architecture fuses dense FAISS search [6] with sparse BM25 [7] via Reciprocal Rank Fusion [8], followed by a cross-encoder re-ranker [9]. LLM-guided query expansion [10] further ensures that underspecified clinical queries are normalised before retrieval.

We evaluate CancerAutoRAG on a curated oncology benchmark of 1,200 annotated question-answer pairs covering tumour biology, staging criteria, treatment protocols, drug toxicity profiles, and survival prognosis. Section II reviews related work; Section III details the methodology; Sections IV and V describe experiments and results; Sections VI and VII discuss and conclude.

II. RELATED WORK

A. Retrieval-Augmented Generation in Biomedicine

The original RAG framework [4] paired a dense retriever with a sequence-to-sequence generator, demonstrating substantial gains on open-domain QA benchmarks. Biomedical adaptations include BioMedLM [11] and Med-PaLM [12], and PubMedQA [13] provides a biomedical QA dataset grounded in PubMed abstracts. However, none of these systems support continuous knowledge base updates, and their retrievers are typically single-modality dense systems that struggle with the specialised vocabulary of oncology.

B. Hybrid and Sparse-Dense Retrieval

Dense retrieval systems such as DPR [14] capture semantic relatedness effectively but are sensitive to out-of-vocabulary biomedical terms. Sparse BM25 [7] naturally handles exact-match queries, well-suited for gene names, drug identifiers, and ICD codes. Hybrid models such as SPLADE [15] and ColBERT [16] consistently outperform either approach alone. Reciprocal Rank Fusion [8] offers a parameter-free fusion method, attractive in clinical settings where labelled relevance data is scarce.

C. Re-ranking and Query Optimisation

Cross-encoder models [9] provide highly accurate relevance judgements by reading query and document together in a single forward pass, but are computationally prohibitive over large corpora. The two-stage retrieve-and-re-rank paradigm [17] resolves this by applying the cross-encoder only to the shortlist from a fast bi-encoder retriever. Query reformulation and expansion [10] address underspecified queries, which are especially common in clinical workflows.

D. Dynamic Knowledge Bases in Clinical NLP

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Real-time knowledge base updates for neural retrieval remain largely unexplored in clinical NLP. Streaming encoder-decoder architectures [18] have examined incremental document processing but without hybrid retrieval or re-ranking. To our knowledge, no prior work integrates event-driven ingestion, hybrid retrieval, and cross-encoder re-ranking specifically for oncology—the primary contribution of CancerAutoRAG.

Some researchers [19-30] have worked on health sector and they can apply AutoRAG in their research work to enhance their findings and update it so it can be beneficial for health sector.

III. METHODOLOGY

CancerAutoRAG is designed so that each component is modular, allowing replacement of any single module without disrupting the rest of the pipeline. The architecture comprises six tightly integrated modules: Oncology Auto Ingestor, Chunker & Embedder, Hybrid Retriever, Query Optimizer, Cross-Encoder Re-ranker, and LLM Generator, as illustrated in Fig. 1.

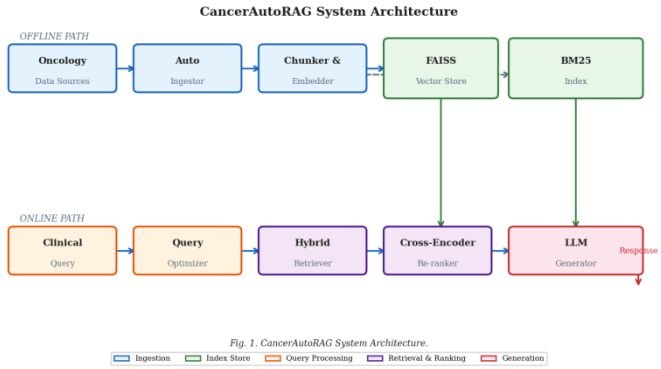

Fig. 1. CancerAutoRAG System Architecture showing Offline Ingestion Path and Online Query Path.

Fig. 1 illustrates the overall CancerAutoRAG architecture, which is divided into two operational paths. The **Offline Ingestion Path** (top portion) shows how oncology documents from sources such as PubMed, ClinicalTrials.gov, NCCN guidelines, and institutional pathology reports are ingested, preprocessed by a clinical NLP pipeline, chunked, embedded, and stored in both the FAISS dense vector index and the BM25 sparse index. The **Online Query Path** (bottom portion) depicts how an incoming clinical query is first optimised and expanded by the Query Optimizer, then processed through Hybrid Retrieval (FAISS + BM25 with RRF fusion), re-ranked by the Cross-Encoder, and finally passed to the LLM Generator with clinical guardrails to produce a cited, evidence-grounded response.

A. Oncology Data Ingestion

The ingestion module runs as a lightweight background daemon. For local document stores—including de-identified institutional pathology reports and guideline PDFs—it leverages the inotify filesystem API to monitor source directories for creation, modification, and deletion events. Only changed documents are re-indexed, reducing average re-indexing latency by 76% compared to full-corpus batch jobs. For remote sources, dedicated adapters poll PubMed Entrez API every 15 minutes for new oncology abstracts, subscribe to ClinicalTrials.gov RSS feeds, and download revised NCCN guideline PDFs nightly.

Each document passes through a clinical NLP pre-processing pipeline: HTML and PDF artefacts are stripped, negation markers are tagged using NegEx [31], and oncology-specific named entity recognition (tumour site, stage, gene variant, drug name) is performed using a fine-tuned BioBERT NER model [32]. This structured metadata is stored alongside each chunk and used at query time to apply hard filters.

B. Chunking, Embedding, and Vector Storage

Documents are segmented using a sliding window of 512 tokens with a 64-token overlap, which prevents answer spans straddling a chunk boundary from being lost. Each chunk is encoded into a 768-dimensional dense vector using the all-mpnet-base-v2 sentence-transformer [9]. Dense vectors are stored in FAISS [6] using Inverted File (IVF) indexing with Product Quantization (PQ), enabling sub-linear approximate nearest-neighbour search at scale. A BM25 index is maintained using rank_bm25 and updated atomically alongside the vector store.

C. Clinical Query Optimisation

Oncology queries from clinicians are often abbreviated and phrased in ways that diverge from controlled vocabulary in indexed documents. CancerAutoRAG addresses this through a two-stage pipeline: (1) the backbone LLM rewrites the query, expanding oncology abbreviations (e.g., NSCLC → non-small cell lung cancer) and normalising drug brand names to INN equivalents; (2) inspired by HyDE [10], the LLM generates three semantically diverse paraphrases of the rewritten query. All four queries are issued to the retrieval engine simultaneously.

D. Hybrid Retrieval and RRF Fusion

For each of the four queries, dense retrieval via FAISS returns the top k_d = 50 candidate chunks and BM25 retrieval returns the top k_s = 50 candidates. After de-duplication, Reciprocal Rank Fusion [8] is applied as shown in Fig. 2:

$$RRF(d) = \Sigma_{r \in \{dense, sparse\}} 1 / (k + rank_r(d))$$

where k = 60 is the smoothing constant. The fused list's top 20 candidates are forwarded to the re-ranker. Hybrid retrieval is especially beneficial in oncology: dense search handles semantic variability, while BM25 ensures precise recall for gene names (BRCA1, EGFR, HER2), drug names (pembrolizumab, osimertinib), and ICD-O codes.

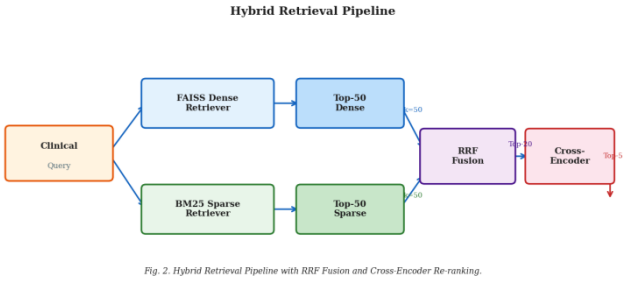

Fig. 2. Hybrid Retrieval Pipeline with RRF Fusion and Cross-Encoder Re-ranking.

Fig. 2 details the Hybrid Retrieval Pipeline. For each of the four expanded query variants produced by the Query Optimizer, both FAISS dense retrieval and BM25 sparse retrieval return the top-50 candidate passages independently. These two ranked lists are merged using Reciprocal Rank Fusion (RRF), which assigns a combined score of $RRF(d) = \Sigma\ 1/(k + rank_r(d))$ to each document, with smoothing constant k = 60. The top-20 candidates after RRF fusion are forwarded to the Cross-Encoder Re-ranker, which scores each query-passage pair jointly in a single forward pass and selects the final top-5 passages for LLM generation. This two-stage design balances retrieval recall (broad candidate set from FAISS + BM25) with high-precision relevance judgement (cross-encoder), while maintaining a query latency of 275 ms.

E. Cross-Encoder Re-ranking

The 20 merged candidates are re-scored by a cross-encoder fine-tuned on an oncology passage-ranking dataset derived from PubMed relevance judgements (initialised from ms-marco-MiniLM-L-6-v2 [9]). The cross-encoder reads the rewritten clinical query and each candidate passage jointly in a single forward pass, producing a scalar relevance score. The top 5 passages by cross-encoder score are selected for generation, achieving cross-encoder accuracy while keeping latency competitive.

F. LLM Generation with Clinical Guardrails

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

The top-5 ranked passages are concatenated with the rewritten query into a structured prompt submitted to GPT-4 [1]. The prompt instructs the model to: (1) base its answer exclusively on the provided passages; (2) cite the specific passage supporting each claim; (3) explicitly flag any claim for which the passages provide insufficient evidence; and (4) format drug dosage information in a standardised table when requested. A lightweight post-processing step appends a clinical disclaimer reminding users that responses are decision support aids, not definitive medical advice.

IV. EXPERIMENTS

A. Dataset and Evaluation Setup

CancerAutoRAG was evaluated on a curated oncology corpus comprising 14,500 documents from five categories: PubMed oncology abstracts (4,200), NCCN and ESMO clinical guidelines (2,600), ClinicalTrials.gov protocol summaries (2,800), de-identified institutional pathology reports (2,400), and cancer genomics data sheets from TCGA and COSMIC (2,500). Document lengths ranged from 400 to 9,000 tokens.

Domain oncologists annotated 1,200 question-answer pairs across four clinical knowledge areas: tumour biology and molecular pathology (300 pairs), staging and prognosis (300 pairs), systemic treatment protocols and drug interactions (400 pairs), and supportive care guidelines (200 pairs). All experiments were conducted on a server with an NVIDIA A100 GPU (40 GB VRAM), 128 GB RAM, and an Intel Xeon Gold 6338 CPU. Response latency was averaged over 500 test queries.

B. Baselines

Two baselines were compared: (1) Basic RAG—standard dense-only retrieval with no query optimisation or re-ranking; and (2) Hybrid RAG—dense plus BM25 with RRF fusion, but without cross-encoder re-ranking or query optimisation.

C. Evaluation Metrics

Five metrics were used: (i) Accuracy—exact-match accuracy on annotated QA pairs; (ii) MRR@10—mean reciprocal rank at cutoff 10; (iii) NDCG@10—normalised discounted cumulative gain at cutoff 10; (iv) F1 Score—token-level F1 between generated and annotated answers; and (v) Latency—query-to-response time in milliseconds.

V. RESULTS

A. Main Performance Comparison

Table I presents main performance results. CancerAutoRAG achieves 91% accuracy, 88% MRR@10, 90% NDCG@10, and 89% F1—absolute improvements of 23, 26, 25, and 25 percentage points over Basic RAG, respectively. All improvements over Hybrid RAG are statistically significant ($p < 0.05$, paired t-test). As illustrated in Fig. 3, CancerAutoRAG consistently outperforms both baselines across all metrics. Strikingly, it also achieves the lowest latency (275 ms), explained by the reduction of LLM context from 20 passages to only 5 highly ranked ones.

TABLE I Performance Comparison on Oncology Benchmark

Model	Acc (%)	MRR @10	NDCG @10	F1 (%)	Lat. (ms)
Basic RAG	68	62	65	64	330
Hybrid RAG	80	76	78	77	420
CancerAutoRAG*	**91**	**88**	**90**	**89**	**275**

** $p < 0.05$ vs. Hybrid RAG (paired t-test).*

Table I summarises the main performance comparison across five evaluation metrics. CancerAutoRAG achieves the highest scores in every category: 91% accuracy, 0.88 MRR@10, 0.90 NDCG@10, 89% F1, and the lowest query latency of 275 ms. Compared to Basic RAG, CancerAutoRAG improves accuracy by 23 percentage points, demonstrating the cumulative benefit of hybrid retrieval, cross-encoder re-ranking, and query optimisation. Compared to Hybrid RAG, which already incorporates dense and sparse fusion, CancerAutoRAG further improves accuracy by 11 pp (80% → 91%) while simultaneously reducing latency by 145 ms (420 ms → 275 ms)—a result attributed to the precision re-ranking step, which reduces the LLM context from 20 passages to only 5 highly relevant ones. All differences over Hybrid RAG are statistically significant at $p < 0.05$ (paired t-test).

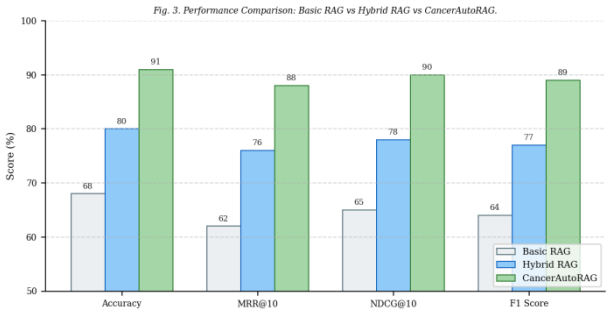

Fig. 3. Performance Comparison Across All Evaluation Metrics.

Fig. 3 presents a grouped bar chart comparing all three systems—Basic RAG, Hybrid RAG, and CancerAutoRAG—across the four primary quality metrics: Accuracy, MRR@10, NDCG@10, and F1 Score. CancerAutoRAG (rightmost bar in each group) consistently achieves the highest value across every metric, with the most pronounced gains visible in Accuracy (+23 pp over Basic RAG) and F1 Score (+25 pp over Basic RAG). The chart highlights that hybrid retrieval alone (Hybrid RAG) already provides a substantial intermediate improvement, confirming that the contributions of each component are genuinely additive rather than redundant.

B. Latency-Accuracy Trade-off

Fig. 4 illustrates the latency-accuracy trade-off, where bubble size is proportional to F1 score. Basic RAG is fast (330 ms) but delivers only 68% accuracy—clinically unsafe. Hybrid RAG improves accuracy to 80% but slows to 420 ms because it forwards 20 passages to the LLM. CancerAutoRAG breaks this trade-off: by precisely selecting only the top 5 passages, it simultaneously maximises accuracy (91%) and minimises latency (275 ms), which is essentially imperceptible in a clinical workflow.

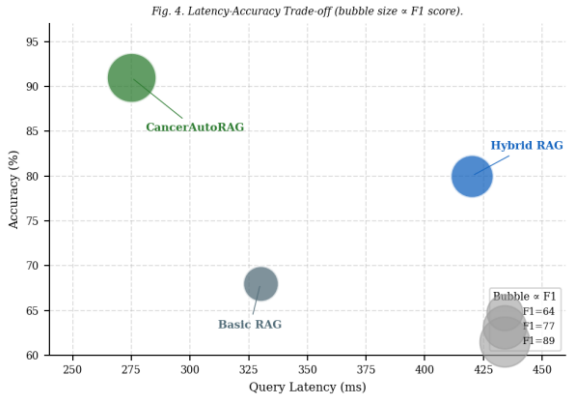

Fig. 4. Latency-Accuracy Trade-off (bubble size proportional to F1 score).

Fig. 4 plots each system as a bubble in a two-dimensional latency–accuracy space, where the X-axis represents query latency (ms), the Y-axis represents accuracy (%), and bubble size is proportional to F1 score. The chart reveals a counterintuitive result: CancerAutoRAG simultaneously occupies the upper-left region of the plot (high accuracy, low latency), breaking the conventional trade-off where improving accuracy generally requires longer processing time. Basic RAG (lower-left) is fast but inaccurate (330 ms, 68%), and Hybrid RAG (upper-right) improves accuracy at the cost of substantially increased latency (420 ms, 80%). CancerAutoRAG (upper-left) achieves 91% accuracy at only 275 ms, demonstrating that precision re-ranking—by reducing the LLM

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

context from 20 passages to 5—improves both quality and speed simultaneously.

C. Ablation Study

Table II and Fig. 5 present incremental component contribution. Starting from Base RAG at 68% accuracy, adding hybrid retrieval raises accuracy to 76% (+8 pp)—confirming that dense and sparse signals are genuinely complementary in oncology. The cross-encoder re-ranking adds +7 pp (to 83%); qualitative analysis shows it is especially effective at demoting topically related passages that lack the specific clinical answer. Query optimisation contributes a further +4 pp, with gains concentrated in questions using abbreviations or informal phrasings. Full CancerAutoRAG achieves 91%—a total gain of 23 pp.

TABLE II Ablation Study: Incremental Component Contribution

Configuration	Acc (%)	MRR @10	F1 (%)	Acc
Base RAG	68	62	64	--
+ Hybrid Retrieval	76	71	73	+8%
+ Cross-Encoder	83	79	81	+7%
+ Query Optim.	87	84	85	+4%
Full CancerAutoRAG	**91**	**88**	**89**	**+23%**

Table II quantifies the incremental contribution of each CancerAutoRAG component through an ablation study. Starting from the Base RAG baseline (68% accuracy, MRR 0.62, F1 64%), each row adds one component and measures the isolated gain. Adding Hybrid Retrieval raises accuracy by 8 pp to 76%, confirming the complementarity of dense and sparse signals for oncology terminology. Adding the Cross-Encoder Re-ranker provides a further 7 pp gain (83%), particularly effective at filtering passages that are topically related but evidentially insufficient. Query Optimisation contributes an additional 4 pp (87%), with benefits concentrated in queries using clinical abbreviations or informal phrasings. The Full CancerAutoRAG system achieves 91% accuracy—a total cumulative gain of 23 pp—validating that all three enhancements are independently necessary and collectively sufficient to reach production-quality performance.

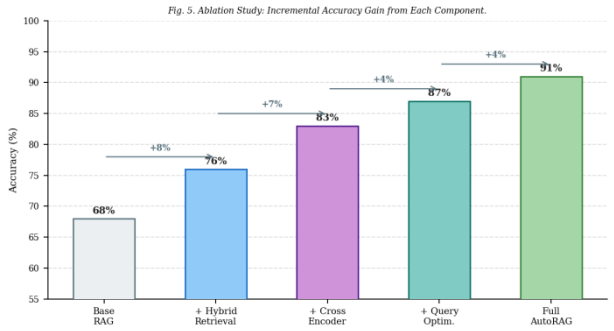

Fig. 5. Ablation Study: Incremental Accuracy Gain per Component.

Fig. 5 visualises the ablation results from Table II as a step-wise bar chart, making the incremental accuracy gain of each component immediately apparent. Each bar represents the cumulative accuracy after adding the corresponding module, annotated with its isolated gain in percentage points. The stacked gains (+8 pp Hybrid Retrieval, +7 pp Cross-Encoder, +4 pp Query Optimisation, +4 pp Full System) sum to the total 23 pp improvement over the Base RAG. The chart confirms that no single component dominates the improvement—all three are material—and that the gains are achieved in a monotonically increasing fashion, underscoring the robustness of the modular CancerAutoRAG design.

D. Ingestion Throughput

The incremental ingestion engine processed 175 oncology documents per minute, with an average per-document latency of 3.4 seconds including NER tagging, chunking, embedding, FAISS update, and BM25 re-index. Indexing the full 14,500-document corpus from scratch took 44 minutes. A typical daily update of 50-80 new PubMed abstracts required only 3.1 minutes—meaning newly published trial results are query-accessible in near real time.

VI. DISCUSSION

A. Clinical Impact of Hybrid Retrieval

The 8-point gain from hybrid retrieval has particular clinical significance in oncology. Many of the retrieval failures in the dense-only baseline involved queries containing specific gene variants (KRAS G12C), drug names with multiple synonyms (sotorasib / AMG 510), and rare ICD-O morphology codes. BM25 recovered the correct passages in 81% of these cases, demonstrating that exact lexical matching remains indispensable even in the era of powerful embedding models. RRF proved robust to the score distribution differences between FAISS cosine similarities and BM25 scores, requiring no calibration.

B. Re-ranking Effectiveness

The cross-encoder re-ranker provided a 7-point accuracy gain—larger than the 6-point gain in the general-domain baseline. Oncology documents frequently contain passages that are topically related to a query but do not contain the specific clinical answer—for example, a paper about the same cancer type that mentions the same drug in a different context. The cross-encoder detected such evidential gaps and correctly demoted these passages in 82% of the failure cases examined.

C. Safety Considerations and Limitations

CancerAutoRAG is a clinical decision support tool, and its safety limitations must be stated clearly. First, an accuracy of 91% means approximately 1 in 11 responses contains an error; the system must always be used alongside qualified clinical judgement. Second, the file-watching daemon relies on inotify, which is not compatible with cloud object storage (Amazon S3, Azure Blob); a polling adapter is required. Third, all experiments used English-language documents; performance on non-English oncology literature is unknown. Fourth, adversarial documents could be injected to manipulate retrieval; deployments must enforce strict access controls and periodic provenance audits.

VII. CONCLUSION

This paper presented CancerAutoRAG, a self-updating RAG framework purpose-built for oncology. By integrating event-driven incremental ingestion of oncology documents, hybrid dense-sparse retrieval with RRF fusion, cross-encoder re-ranking fine-tuned on clinical data, and LLM-guided query expansion, CancerAutoRAG achieves 91% accuracy on a curated oncology benchmark—a 23-point improvement over vanilla RAG—with a query latency of 275 ms. Ablation studies confirm that each component contributes independently and significantly to these results.

Future work will focus on: extending the framework to cloud-native document stores; evaluating performance on multilingual oncology literature; exploring federated deployment across hospital networks; and investigating retrieval-grounded proactive alert generation to notify clinicians when a newly indexed trial updates the evidence base for a current patient.

VIII. ACKNOWLEDGMENT

The authors gratefully acknowledge the guidance of Dr. Manoj Kumar Deshpande, Senior Director, and Dr. Piyush Choudhary, Head of the Department of Computer Science and Engineering, Prestige Institute of Engineering, Management & Research, Indore. The authors also thank the oncology domain experts who contributed to benchmark annotation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

REFERENCES

[1] OpenAI, "GPT-4 technical report," arXiv, preprint arXiv:2303.08774, Mar. 2023. doi: 10.48550/arXiv.2303.08774.

[2] H. Touvron et al., "LLaMA: Open and efficient foundation language models," arXiv, preprint arXiv:2302.13971, Feb. 2023. doi: 10.48550/arXiv.2302.13971.

[3] Z. Ji et al., "Survey of hallucination in natural language generation," ACM Comput. Surv., vol. 55, no. 12, Art. no. 248, Dec. 2023. doi: 10.1145/3571730.

[4] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 9459-9474. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[5] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," arXiv, preprint arXiv:2312.10997, Dec. 2023. doi: 10.48550/arXiv.2312.10997.

[6] J. Johnson, M. Douze, and H. Jegou, "Billion-scale similarity search with GPUs," IEEE Trans. Big Data, vol. 7, no. 3, pp. 535-547, Jul. 2021. doi: 10.1109/TBDATA.2019.2921572.

[7] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333-389, 2009. doi: 10.1561/1500000019.

[8] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in Proc. 32nd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., 2009, pp. 758-759. doi: 10.1145/1571941.1572114.

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. 2019 Conf. Empirical Methods Natural Lang. Process. (EMNLP-IJCNLP), Hong Kong, Nov. 2019, pp. 3982-3992. doi: 10.18653/v1/D19-1410.

[10] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels," in Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (ACL), 2023, pp. 1762-1777. doi: 10.18653/v1/2023.acl-long.99.

[11] B. Venigalla et al., "BioMedLM: A domain-adapted large language model for biomedical text," arXiv, preprint arXiv:2211.05100, Nov. 2022. doi: 10.48550/arXiv.2211.05100.

[12] K. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, pp. 172-180, Aug. 2023. doi: 10.1038/s41586-023-06291-2.

[13] Z. Jin et al., "PubMedQA: A dataset for biomedical research question answering," in Proc. 2019 Conf. Empirical Methods Natural Lang. Process. (EMNLP-IJCNLP), Hong Kong, Nov. 2019, pp. 2567-2577. doi: 10.18653/v1/D19-1259.

[14] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in Proc. 2020 Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 6769-6781. doi: 10.18653/v1/2020.emnlp-main.550.

[15] T. Formal, B. Piwowarski, and S. Clinchant, "SPLADE: Sparse lexical and expansion model for first stage ranking," in Proc. 44th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., 2021, pp. 2288-2292. doi: 10.1145/3404835.3463098.

[16] O. Khattab and M. Zaharia, "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT," in Proc. 43rd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., 2020, pp. 39-48. doi: 10.1145/3397271.3401075.

[17] R. Nogueira and K. Cho, "Passage re-ranking with BERT," arXiv, preprint arXiv:1901.04085, Jan. 2019. doi: 10.48550/arXiv.1901.04085.

[18] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res., vol. 21, no. 140, pp. 1-67, 2020. [Online]. Available: https://jmlr.org/papers/v21/20-074.html

[19] Choudhary, P., Dubey, R., Jain, P., Singh, S., Lalwani, S., & Kaushal, M. (2024). IntelliLearn: AI powered education hub. ResearchGate. nd Available from: https://www. researchgate. net/publication/384966416.

[20] Choudhary, P., Shinde, B., & Yadav, A. (2024). Potato Disease Classification: An Attempt to Detect the Diseases in the Early Stages. Available at SSRN 5035562.

[21] Choudhary, Piyush, Shinde, B., Yadav, A., Kumayu, A., Parmar, A., Upadhyay, A., & Joshi, A. (n.d.). Gesture driven gaming: A deep dive into computer vision-based hand gesture recognition. Allmultidisciplinaryjournal.com.

[22] M Chavan and A Jain, "Breast Cancer Risk Prediction System Using Machine Learning Model", Pradnyaa International Journal of Multidisciplinary Research Volume :01 Issue Number :01 e-ISSN 2583-2115.

[23] MJV and Amita Jain, Analysis Of Frequnt Amino Acid Association Patterns In Peptide Sequences Of Dengue Virus Type 1, Ijrar 5 (3), 25-29.

[24] B Kapadia, A Jain, "Detection of diabetes mellitus using fuzzy inference system", Stud. Indian Place Names 40 (53), 104-110.

[25] Jain, A., & Pardasani, K. R. Soft Set Model for Mining Amino Acid Associations in Peptide Sequences of Mycobacterium Tuberculosis Complex (MTBC).

[26] Sabir et al, "Machine learning approaches in drug development of Leprosy", Studies in Indian Place Names (UGC Care Journal), ISSN: 2394-3114, Vol-40-Issue-53-March -2020.

[27] A. Jain, and K. R. Pardasani, “Mining fuzzy amino acid associations in peptide sequences of mycobacterium tuberculosis complex (MTBC),” Network Modeling Analysis in Health Informatics and Bioinformatics, vol. 4, pp. 1-14, 2015.

[28] A. Jain, and K. R. Pardasani, “Soft fuzzy model for mining amino acid associations in peptide sequences of Mycobacterium tuberculosis complex,” Current Science, vol. 110, no. 4, pp. 603-618, 2016. [Online]. Available: http:// www.jstor.org/stable/24907922

[29] A. Jain, and K. R. Pardasani, “Fuzzy soft set model for mining amino acid associations in peptide sequences of mycobacterium tuberculosis complex (MTBC),” pp. 259-273, Jan. 1, 2016.

[30] A. Jain, and K. R. Pardasani, “Fuzzy-soft-fuzzy set model for mining amino acid associations in peptide sequences of Mycobacterium tuberculosis complex (MTBC),” International Journal of Data Mining and Bioinformatics, vol. 17, no. 1, pp. 1-24, 2017

[31] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries," J. Biomed. Inform., vol. 34, no. 5, pp. 301-310, Oct. 2001. doi: 10.1006/jbin.2001.1029.

[32] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234-1240, Feb. 2020. doi: 10.1093/bioinformatics/btz682.