

Evidence-Grounded Clinical Decision Support via Patient-Scoped Hybrid Retrieval-Augmented Generation and Claim-Level Verification

Raksha¹, Prathiksha Shetty², Nisha L³, Rajesh M Rai⁴, Ms. Deepthi⁵,

^{1,2,3,4,5} Assistant Professor, Department of Computer Science and Engineering, Shree Devi Institute of Technology, Visvesvaraya Technological University

Abstract—Large language models (LLMs) are strong at generating natural language, but using them for clinical decision support demands factual grounding, traceable evidence, and protection of patient data. This paper presents a clinical decision support system (CDSS) built on Retrieval-Augmented Generation (RAG). It combines dense semantic retrieval, BM25 lexical retrieval, intent-aware Reciprocal Rank Fusion (RRF), cross-encoder reranking, and evidence deduplication. Generated answers are checked at the sentence level using semantic similarity, cross-encoder scoring, and clinical contradiction rules, and the system adds confidence scoring and patient-scoped access control. Clinical documents are segmented hierarchically, enriched with contextual metadata, and indexed with FAISS and BM25. Retrieved evidence is reranked and filtered before generation, and each generated claim is then verified against its supporting evidence. Evaluation used 25 benchmark queries over six clinical documents with 31 indexed chunks. The full system reached a Recall@10 of 1.0000, a faithfulness of 1.0000 under deterministic offline (mock) generation, a citation precision of 0.3636, a claim-verification F1-score of 0.8511, and a mean end-to-end latency of 5194.71 ms. The evaluation also covered retrieval performance, evidence grounding, claim verification, and per-component latency. Cross-encoder reranking and NLI verification together made up about 99.0% of total latency, making them the main computational bottlenecks. These results are preliminary: the corpus is small, faithfulness does not reflect live LLM generation, and BM25-only and Hybrid retrieval matched or exceeded the full system on MRR at a fraction of the latency. Further optimization and evaluation with live language models and larger clinical datasets are needed.

Index Terms—Clinical Decision Support System, Retrieval-Augmented Generation, Hybrid Retrieval, Evidence Verification, Clinical NLP

I. INTRODUCTION

Clinical decision support systems require reliable access to patient information, clinical guidelines, laboratory results, and medical evidence. Traditional retrieval systems can locate relevant documents but do not necessarily synthesize evidence

into a response, while large language models (LLMs) may generate unsupported or inaccurate information. Retrieval-Augmented Generation (RAG) combines language models with external knowledge sources to improve evidence-based responses [1]. Dense retrieval supports semantic matching [2], whereas BM25 is effective for exact medical terminology and numerical information [3]. However, clinical applications additionally require evidence traceability and verification. MedNLI provides a foundation for clinical inference [7], while RAGTruth, FActScore, and ARES address hallucination detection and factuality evaluation in generated responses [9][11].

To address these challenges, this work proposes a multi-stage clinical RAG system that combines dense and lexical retrieval using Reciprocal Rank Fusion (RRF) [4]. A cross-encoder reranks retrieved evidence, and an NLI-based verification stage checks generated claims against the retrieved context. The system also incorporates patient-scoped retrieval and role-based citation filtering. These mechanisms are designed to improve the relevance, reliability, and traceability of generated clinical responses. The verification stage further distinguishes supported information from unsupported or contradictory

claims, reducing dependence on unsupported model knowledge.

The main contributions of this work are: Hybrid clinical retrieval using dense and BM25 methods with intent-aware rank fusion. Cross-encoder evidence reranking and contextual evidence deduplication. Sentence-level evidence verification using semantic similarity, cross-encoder scoring, and clinical contradiction rules. Confidence scoring based on claim grounding, evidence coverage, and retrieval quality. Patient-scoped retrieval isolation and role-based citation filtering (implemented; not empirically tested in this work).

II. LITERATURE REVIEW

Lewis et al. introduced Retrieval-Augmented Generation as a framework that combines a parametric sequence-to-sequence model with a non-parametric external memory [1]. This provides a foundation for grounding generated responses in retrieved information. The proposed CDSS follows the same general principle while introducing additional clinical evidence filtering and post-generation verification. Dense Passage Retrieval demonstrated that learned dense representations can effectively retrieve semantically relevant passages for question answering [2]. However, semantic retrieval may be insufficient for exact clinical terms, abbreviations, drug names, or numerical information. Therefore, the proposed system complements dense retrieval with BM25.

BM25 is a probabilistic lexical retrieval method that considers term frequency, document length, and inverse document frequency [3]. Although dense retrieval captures semantic similarity, lexical retrieval can provide strong performance for exact terminology and numerical expressions. The proposed architecture executes dense and BM25 retrieval and combines their rankings using weighted Reciprocal Rank Fusion. Cormack et al. introduced Reciprocal Rank Fusion (RRF), a ranking combination method that aggregates multiple ranked lists using reciprocal rank contributions [4]. The proposed system applies RRF after dense and lexical retrieval, with weights selected by a deterministic clinical query router. Cross-encoder reranking with BERT-style models [5] and late-interaction retrieval [6] have demonstrated the value of deeper query-passage interaction for ranking.

Romanov and Shivade introduced MedNLI, an expert-annotated dataset for natural language inference in the clinical domain [7]. Their work demonstrated that clinical NLI presents domain-specific challenges and that domain knowledge can improve inference performance.

Self-RAG demonstrated that retrieval and self-reflection can improve factuality and citation accuracy in generated responses [8]. RAGTruth provides a corpus for studying hallucinations in retrieval-augmented language models and emphasizes that unsupported or contradictory claims can still occur after retrieval [9]. FactScore evaluates factuality by decomposing generated text into atomic facts and determining the proportion supported by reliable evidence [10]. ARES provides automated evaluation of RAG systems using context relevance, answer faithfulness, and answer relevance [11]. These approaches motivate the claim-level evaluation and grounding metrics used in this work.

The proposed system integrates these components into one pipeline:

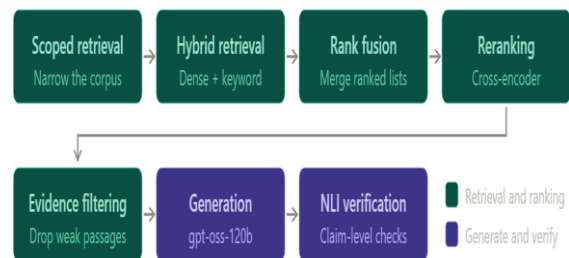


Fig. 1. Proposed pipeline: scoped and hybrid retrieval, rank fusion and reranking feed evidence filtering, generation and NLI verification.

This integrated design addresses the gap between retrieval quality and clinical response reliability by incorporating post-generation evidence verification. Claim-level verification helps identify unsupported or contradictory statements before presenting the final response.

III. METHODOLOGY

A. System Architecture

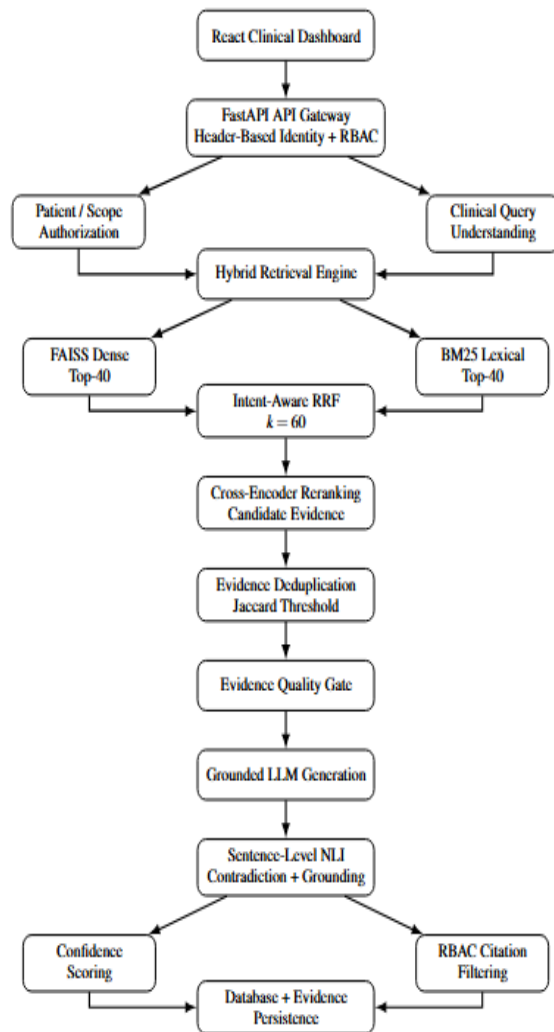


Fig. 2. Architecture of the proposed clinical decision support system, showing patient-scoped retrieval, hybrid evidence search, reranking, grounded generation, and claim verification.

The process begins when a citizen submits a grievance through the public interface. The citizen provides the required complaint details, including a description of the issue and its geographical location. The location can be selected using the map by searching for a place, using the available location option, or selecting the exact position through the map marker. The system records the corresponding latitude and longitude along with the grievance.

B. Clinical Document Ingestion

The ingestion pipeline accepts clinical PDF documents and validates them before indexing.

Content is extracted page by page, and a medical relevance gate checks whether sufficient clinical terminology is present. Documents containing fewer than two distinct clinical terms are rejected, documents with intermediate relevance are held for review, and documents that meet the approval threshold are indexed.

C. Text Extraction and Chunking

Documents are hierarchically segmented into sections and subsections, with the page number and source file name preserved. Long sections are divided at sentence boundaries into chunks of 500 words with an overlap of 100 words, so that clinical statements spanning a chunk boundary are not lost. Each chunk is enriched with contextual metadata, including document identifier, patient scope, source page, and section heading, before embedding. This metadata is later used for patient-scoped filtering and for tracing every generated claim back to its source.

D. Embedding Generation and Index Construction

Each chunk is converted into a 384-dimensional vector using the all-MiniLM-L6-v2 sentence-transformer model. The vectors are L2-normalized and stored in a FAISS IndexFlatIP index, so that inner product equals cosine similarity. In parallel, the chunks are indexed with BM25 using $k1 = 1.5$ and $b = 0.75$. Maintaining both indexes is important in the clinical domain, because dense retrieval handles paraphrased queries well, while BM25 preserves exact matches for drug names, abbreviations, and laboratory values that embeddings often blur.

E. Patient-Scoped Hybrid Retrieval

When a query arrives, the system first identifies the user, verifies role and patient access, and extracts clinical entities and intent. Patient and scope filters are applied before retrieval, so evidence from other patients never enters the context. Within this scope, the FAISS dense retriever and the BM25 retriever each return their top 40 candidates. The two ranked lists are merged using weighted Reciprocal Rank Fusion (RRF), in which each chunk receives a score from both retrievers based on its rank position, weighted by the retriever's importance, and the two contributions are added to give the fused score. A smoothing constant of 60 is used so that no single top-ranked position dominates the fused ranking. A deterministic query

router assigns a higher BM25 weight to keyword-heavy queries and a higher dense weight to diagnostic or semantically oriented queries. Because RRF depends only on rank positions, it avoids normalizing the very different score scales of the two retrievers.

F. Cross-Encoder Reranking and Evidence Filtering

The fused candidates are reranked using the ms-marco-MiniLM-L-6-v2 cross-encoder, which reads the query and each chunk together and gives a more precise relevance score than independent embeddings. The raw score is passed through a sigmoid function, then adjusted by a section-alignment boost and a source-authority factor. Near-duplicate passages are then removed using token-level Jaccard similarity, discarding any candidate whose similarity to an already admitted passage exceeds 0.82. The final context is limited to a small set of non-redundant evidence passages.

G. Evidence Quality Gate and Grounded Generation

Before generation, an evidence gate checks whether the retrieved evidence is sufficient, using dense similarity, BM25 evidence, and cross-encoder quality. The dense similarity threshold is 0.35 and the cross-encoder raw score cutoff is -3.0 . If sufficient evidence cannot be established, generation stops and an insufficient-evidence response is returned. Otherwise, the selected passages are assembled into a provenance-aware prompt carrying the source, page, section, and reference identifier of each passage. The language model is instructed to answer only from this evidence and to cite the supporting passages. Generation uses openai/gpt-oss-120b through the Groq API with a temperature of 0.05, and a Gemini-based fallback is configured.

H. Sentence-Level Claim Verification

The generated response is divided into individual factual claims, and each claim is evaluated against the candidate evidence passages at multiple granularities. The verifier combines semantic similarity, cross-encoder evidence scores, and explicit clinical contradiction rules based on antonym pairs and negation patterns. A claim is labeled supported when its similarity to the evidence is at least 0.55, or when its cross-encoder score is at least 0.50 and its similarity is at least 0.38. Claims that fall below this level but meet the partial-evidence conditions are labeled partially supported, and claims triggering a

contradiction rule receive a substantially lower score. This stage is NLI-inspired rather than a trained entailment model. Supported claims are treated as grounded in the retrieved evidence, while partially supported claims are flagged as having weaker backing. Because the stage is not a trained entailment model, its labels are heuristic and should be read as approximate judgments rather than definitive verdicts.

I. Confidence Scoring

The overall confidence combines three signals: claim grounding and coverage (weight 0.70), the mean of the top three dense retrieval scores (weight 0.15), and the mean of the top three reranking scores (weight 0.15). If no supported claims exist, confidence is capped at a low value, and if a contradiction is detected it is strongly reduced. The final answer is reported as High, Medium, or Low confidence.

J. Evidence Attribution, Filtering and Storage

For every supported claim, the system extracts the supporting evidence with its source document and page number. Citations are then filtered according to the user's role, so unauthorized evidence is never exposed. The query, verified response, confidence, and evidence references are stored in the database to keep an auditable history, and the verified response is presented with inline citations.

The complete methodology of the proposed system is summarized in Fig. 2. The workflow begins with the upload of a medical PDF, followed by text extraction, document chunking, and embedding generation, after which the chunks are indexed in the FAISS vector database. When a user submits a query, the system performs semantic retrieval, reranks the candidates with a cross-encoder, and selects the most relevant chunks as evidence. The language model generates a response from this evidence, and each sentence is then verified and assigned a confidence score. Finally, the supporting evidence is extracted with its source document and page number, the interaction is stored in the database, and the verified clinical response is returned to the user.

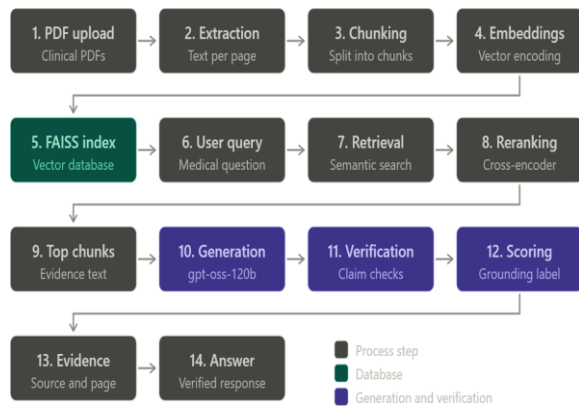


Fig. 3. Overall methodology of the proposed RAG-based clinical decision support system.

IV. IMPLEMENTATION AND SYSTEM ENGINEERING

The proposed clinical decision support system is implemented as an integrated web-based platform that combines clinical document ingestion, patient-scoped hybrid retrieval, evidence-grounded generation, claim-level verification, and access-controlled citation delivery. This integrated approach is motivated by prior work on retrieval-augmented generation [1], hybrid retrieval [3], [4], and hallucination and factuality evaluation [9]-[11], which have generally been studied as separate components.

A. System Architecture

The system follows a layered architecture consisting of a React-based clinical dashboard, a FastAPI backend, a retrieval and verification engine, and a persistence layer. The dashboard communicates with the backend through protected API endpoints. The backend handles user identification, role checking, query processing, and response generation, while the retrieval and verification engine performs hybrid search, reranking, and claim verification. The implementation is divided into modules for authentication and authorization, clinical entity extraction and query expansion, deterministic retrieval routing, retrieval and reranking, the end-to-end pipeline, claim verification and confidence calculation, database persistence, and document storage. Keeping these responsibilities in separate modules allows each stage of the pipeline to be tested and evaluated independently.

B. Document Indexing Configuration

Clinical documents are indexed using the following configuration. Chunks of 500 words with an overlap of 100 words are embedded with the all-MiniLM-L6-v2 model into 384-dimensional vectors and stored in a FAISS IndexFlatIP index. The dense and BM25 retrievers each return up to 40 candidates, which are fused with Reciprocal Rank Fusion using a constant of 60. The fused candidates are reranked with the ms-marco-MiniLM-L-6-v2 cross-encoder, and redundant passages are removed at a Jaccard threshold of 0.82. For verification, a claim is labeled supported at a similarity of 0.55 and partially supported at 0.38. The evidence gate uses a dense similarity threshold of 0.35 and a cross-encoder raw quality cutoff of -3.0

C. Authentication and Role-Based Access

The Protected API operations require a user context. The current implementation obtains the user identifier from the X-User-Id request header and maps it to the corresponding user record. Role-checking dependencies then restrict access to operations and clinical data according to the user's assigned role. This header is an identity claim that is not cryptographically verified, so it does not by itself authenticate the caller. It is therefore a development-stage mechanism, and its limitations are discussed in Section VII.

D. Patient Data Isolation

Patient-specific evidence is associated with its patient scope, while hospital-wide guidelines and textbooks are kept separate from individual patient records. Patient and scope filters are applied before retrieval, so evidence outside the requesting user's permitted scope is never included in the generation context. This prevents patient-specific information from reaching the language model or appearing in the response. Citation filtering after verification acts as a secondary control.

E. End-to-End Processing

Query processing follows the implemented sequence. The system first identifies the requesting user, verifies role and patient access, and extracts clinical entities and intent, generating query variants where applicable. Patient and scope filters are then applied, and candidates are retrieved with FAISS and BM25 and combined using weighted RRF. The fused candidates

are reranked with the cross-encoder, and redundant passages are removed. The evidence-quality gate then decides whether generation can proceed. If it can, an evidence-grounded response is generated, its factual claims are extracted, and each claim is verified against the retrieved evidence. Confidence is then calculated, citations are filtered according to the user's permissions, and the response and verification metadata are stored in the database. This sequence corresponds to the implemented flow and is illustrated in Fig. 1 and Fig. 2.

F. Persistence and Storage

Chat history, generated responses, verification results, confidence values, and evidence references are stored in a relational database accessed through SQLAlchemy, which provides an auditable record of each interaction. Uploaded documents are handled through a storage abstraction that applies file-type and size validation and path sanitization before processing.

G. Offline Generation Environment

The intended generation configuration uses openai/gpt-oss-120b through the Groq API with a Gemini-based fallback, at a temperature of 0.05. During the reported evaluation, the external API keys were unavailable, so the repository's deterministic mock-generation path was used to evaluate the retrieval and verification pipeline. Because the mock generator composes its output from the retrieved evidence, the reported faithfulness and grounding results reflect this offline setup and should not be read as measurements of live language-model performance. Evaluation with a live model under the same conditions remains future work. Since the mock output is built directly from the retrieved passages, high grounding scores are expected by construction and mainly confirm that the retrieval and verification stages work as designed. A live-model run would additionally test how often the generator introduces unsupported claims, which this offline setup cannot capture.

V. RESULTS AND DISCUSSION

A. Retrieval Results

The proposed clinical RAG system was evaluated on 25 benchmark queries over six clinical documents containing 31 indexed chunks. Retrieval metrics were

calculated over the 23 queries with positive ground truth, while generation and grounding metrics were calculated over all 25 queries. The dense FAISS retriever achieved a Recall@10 of 1.0000 with an MRR of 0.8406, whereas BM25 alone achieved a higher MRR of 0.9783 with an average latency of only 0.17 ms. The hybrid configuration retained a Recall@10 of 1.0000 and an MRR of 0.9783 at 29.58 ms, showing that combining semantic and lexical evidence preserves high recall without losing ranking quality. Adding weighted RRF gave an MRR of 0.8913 and an NDCG@10 of 0.9180, which were lower than plain hybrid retrieval on this small corpus. Adding the cross-encoder reranker raised Recall@5 from 0.9783 to 1.0000 and Precision@5 from 0.2174 to 0.2261. This Recall@5 gain of 0.0217 corresponds to a single query on the 23-query positive set. At the same time, MRR dropped from 0.9783 to 0.9710 and NDCG@10 from 0.9721 to 0.9659, and latency rose to 3762.09 ms. The full system kept a Recall@10 of 1.0000 and an MRR of 0.9710 while adding evidence verification and safety mechanisms, but its MRR and NDCG@10 stayed below those of the hybrid configuration, and it ran at roughly 100 times the latency of BM25 alone. The Direct LLM baseline scored zero on all retrieval metrics because it does not retrieve from the indexed corpus. These zeros are valid baseline measurements, not missing values.

B. Generation and Grounding Results

The Direct LLM baseline scored zero faithfulness and zero citation precision, with an unsupported claim rate of 1.0000, because in the offline setup it had no access to the indexed corpus. All RAG configurations achieved a faithfulness of 1.0000 and an unsupported claim rate of 0.0000. Because generation used the deterministic mock path described in Section IV-G, this result is expected by construction and does not indicate the faithfulness of a live language model. The identical scores for RAG without NLI and RAG with NLI also show that these metrics do not capture any contribution from the verification stage. The evidence quality gate suppressed both out-of-scope queries before any claims were generated in the full system. Citation precision was about 0.36 across the RAG configurations (0.3667 without the full verification pipeline and 0.3636 for the full system), which means roughly 64% of citations did not match the ground truth. Citation selection therefore needs improvement.

For calibration, RAG without NLI had an Expected Calibration Error of 0.5789, while RAG with NLI scored 0.0000. This should not be read as perfect calibration, since all evaluated claims were supported and ten bins over so few samples are unreliable. The full system's ECE of 0.2288 is higher because its confidence also incorporates retrieval, coverage, and reranking signals.

C. Claim Verification Results

Claim verification was evaluated on 40 clinically labeled claim-premise pairs, consisting of 20 supported claims, 10 neutral or unsupported claims, and 10 contradictory claims. Similarity alone gave a precision of 0.6786, a recall of 0.9500, and an F1-score of 0.7917. The cross-encoder alone gave a precision of 0.6667, a recall of 1.0000, and an F1-score of 0.8000. Combining similarity and the cross-encoder did not help, giving an F1-score of 0.7692. The full validator, which adds clinical antonym and negation rules, achieved the best result with a precision of 0.7407, a recall of 1.0000, and an F1-score of 0.8511.

A precision of 0.7407 with a recall of 1.0000 corresponds to 20 true positives and 7 false positives, meaning 7 of the 20 neutral or contradictory pairs were incorrectly accepted as supported. In a clinical setting, this false-accept rate is a significant limitation. The results also show why similarity alone is insufficient for clinical claim validation, since claims with substantial lexical overlap with the evidence may still reverse its clinical meaning. Explicit contradiction handling therefore adds a useful layer of verification.

D. Latency Analysis

The measured end-to-end latency was 5194.71 ms, averaged over warm iterations on the 25-query benchmark. Entity extraction took 12.53 ms, dense retrieval 28.46 ms, BM25 retrieval 0.85 ms, RRF fusion 0.09 ms, evidence deduplication 0.28 ms, and mock LLM generation 10.22 ms. The two dominant components were cross-encoder reranking at 2675.72 ms and NLI verification at 2466.55 ms. Together they account for approximately 99.0% of the total latency on the evaluated CPU environment, which identifies transformer-based evidence validation as the main computational bottleneck.

The generation time reflects the deterministic mock path, so a live LLM call would add substantially to the

end-to-end latency. NLI verification time will also grow with the number of generated claims. The retrieval latencies in the ablation study (three warm runs per query) and the per-component latencies here (five warm iterations per query) come from separate experiments and are not directly additive.

E. Discussion

The results show that the pipeline retrieves relevant evidence reliably on this small corpus, with a Recall@10 of 1.0000 for the dense, hybrid, and full configurations. However, the added stages did not give a measurable ranking benefit over BM25 alone or plain hybrid retrieval. Reranking improved Recall@5 by a single query but lowered MRR and NDCG@10, and it cost far more latency. The main value of the full system therefore lies in what follows retrieval, namely the evidence gate, claim verification, confidence scoring, and access-controlled citations, rather than in better ranking.

The claim validator reached an F1-score of 0.8511 and detected the contradictions covered by its rules, but it still accepted 7 of 20 neutral or contradictory pairs. Its rules do not explicitly cover numeric, unit, or laterality contradictions, which are especially important in clinical text. Citation precision of about 0.36 also shows that choosing the right source for each claim remains the weakest part of the system.

These findings must be read within the limits of the evaluation. The corpus had only six documents and 31 chunks, so with a candidate depth of 40 each retriever could return the entire corpus, which makes Recall@10 non-discriminative. Faithfulness reflects mock generation and not a live model, no held-out split, annotator agreement, or confidence intervals are reported, and patient scoping and access control were not empirically tested. The results therefore show that the pipeline is feasible under the evaluated conditions, but they do not establish live-LLM performance, clinical validity, or generalizability. Larger, independently annotated datasets, evaluation with a live language model, access-control testing, and latency optimization are needed in future work.

VI. CONCLUSION

This paper presents a scoped hybrid Retrieval-Augmented Generation (RAG) Clinical Decision Support System for evidence-grounded, traceable

clinical question answering. The system combines dense FAISS retrieval, BM25, intent-aware Reciprocal Rank Fusion, cross-encoder reranking, evidence deduplication, quality gating, grounded generation, and sentence-level NLI-inspired claim verification. Unlike basic RAG pipelines, it verifies generated claims against retrieved evidence and uses contradictions to set confidence scores and citations. Patient-scoped retrieval and role-based citation filtering support access control.

On 25 benchmark queries over six clinical documents, the Full System achieved a Recall@10 of 1.0000 and a faithfulness of 1.0000 under deterministic mock generation. The claim validator obtained an F1-score of 0.8511 but falsely accepted 7 of 20 neutral or contradictory pairs. End-to-end latency was 5194.71 ms, with reranking and verification accounting for about 99.0% of it. On this small corpus, BM25-only and Hybrid retrieval matched or exceeded the Full System on MRR at far lower latency, and citation precision was only about 0.36. The results show the pipeline is feasible under the evaluated conditions but do not establish live-LLM performance or clinical validity. Future work includes larger datasets, live-LLM evaluation, access-control testing, and latency optimization.

ACKNOWLEDGEMENT

The authors thank their project guide, faculty members, and institution for their guidance and support. Their valuable suggestions and feedback contributed to the development, evaluation, and completion of this work.

REFERENCES

- [1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. 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, 2020.
- [2] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," in *Proc. EMNLP*, 2020, pp. 6769–6781.
- [3] 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.
- [4] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods," in *Proc. 32nd ACM SIGIR Conf. Research and Development in Information Retrieval*, 2009, pp. 758–759.
- [5] R. Nogueira and K. Cho, "Passage Re-ranking with BERT," *arXiv preprint arXiv:1901.04085*, 2019.
- [6] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in *Proc. 43rd Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 2020, pp. 39–48.
- [7] A. Romanov and C. Shivade, "Lessons from Natural Language Inference in the Clinical Domain," in *Proc. EMNLP*, 2018, pp. 1586–1596.
- [8] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," in *Proc. International Conference on Learning Representations (ICLR)*, 2024.
- [9] C. Niu, Y. Wu, J. Zhu, S. Xu, K. S. Shum, R. Zhong, J. Song, and T. Zhang, "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models," in *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, pp. 10862–10878.
- [10] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FACTScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation," in *Proc. EMNLP*, 2023, pp. 12076–12100.
- [11] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," in *Proc. NAACL-HLT*, 2024, pp. 338–354.