ISSN: 3107-6513

MedRAG: A Healthcare Conversational Assistant Using Retrieval-Augmented Generation

Grounding Clinical AI in Evidence: A Three-Tier Progressive RAG Architecture for Medical Question Answering

Alur Taher Basha [#1], Mr. P. Bharath Kumar [#2], Dr. D. William Albert [#3]

#1M.Tech Student, Dept. of CSE, Bheema Institute of Technology and Science, Adoni, India
#2Associate Professor, Dept. of CSE, Bheema Institute of Technology and Science, Adoni, India
#3Professor, Head of Dept. CSE, Bheema Institute of Technology and Science, Adoni, India
#1 **taherbasha295@gmail.com**, #2 **bharathkumar1218@gmail.com**, #3 **dr.albertdwgtl@gmail.com**

Abstract— *The rapid expansion of digital health information and the complexity of clinical decision-making have generated a pressing demand for intelligent, evidence-grounded healthcare information systems. Conventional Large Language Models (LLMs) exhibit impressive generative fluency but remain vulnerable to hallucination, static knowledge boundaries, and inability to attribute responses to verifiable sources — characteristics rendering them unsuitable for unsupervised deployment in safety-critical clinical environments. This paper introduces MedRAG, a healthcare conversational assistant embedding a Retrieval-Augmented Generation (RAG) pipeline at its core. MedRAG dynamically retrieves current evidence-based content from curated medical knowledge repositories before generating any clinical response, ensuring every output is grounded in authoritative literature rather than static parametric memory. The system is designed and evaluated across three architectural paradigms of progressive sophistication: Foundational RAG, Optimised RAG, and Modular RAG. Experiments conducted over five standardised medical benchmarks — MedQA-USMLE, MedMCQA, PubMedQA, MMLU-Medical, and a private chronic-pain diagnostic dataset — demonstrate consistent and statistically significant improvements over standalone LLM baselines. The Modular RAG configuration with GPT-4-Turbo achieves 91.1% accuracy on MedQA-USMLE and near-zero hallucination rates measured by the RAGAS faithfulness metric. Open-source models augmented with the modular pipeline record accuracy gains exceeding 27 percentage points, substantially narrowing the performance gap to proprietary frontier systems at a fraction of computational cost.*

.

Keywords— *Retrieval-Augmented Generation (RAG); Healthcare Conversational AI; Large Language Models (LLMs); Medical Question Answering; FAISS; BM25; Hallucination Reduction; RAGAS Faithfulness; Clinical Decision Support; Medical Knowledge Graph; UMLS; Cross-Encoder Re-ranking.*

I. INTRODUCTION

Healthcare is among the most information-intensive domains in modern society. A practising clinician must synthesise evidence from randomised controlled trials, institutional guidelines, pharmacological databases, patient records, and real-time diagnostic data — all within the constrained timeframe of a clinical encounter. This cognitive burden contributes directly to diagnostic error rates, treatment variability, and clinician burnout, creating an urgent demand for intelligent, evidence-based decision-support technologies that can rapidly retrieve, synthesise, and present relevant clinical knowledge at the point of care.

Artificial Intelligence and Natural Language Processing (NLP) have emerged as promising solutions to this challenge. Large Language Models (LLMs) such as GPT-4, Google PaLM-2, and Meta LLaMA-3 demonstrate extraordinary breadth of general knowledge, achieving competitive or super-human performance on standardised assessments spanning law, mathematics, and clinical medicine. Despite these achievements, deploying standalone LLMs in live clinical environments raises serious and well-documented safety concerns that preclude direct unsupervised clinical use.

LLMs are trained on static corpora with fixed knowledge cutoff dates. A model trained even six months prior may recommend a superseded treatment protocol with equal confidence to current best practice. Furthermore, LLMs exhibit a well-characterised tendency to generate factually incorrect but semantically plausible statements — a behaviour universally termed hallucination — which can directly endanger patient wellbeing. Standalone LLMs also provide no mechanism for tracing generated claims to verifiable primary sources, an essential requirement for clinical accountability and regulatory compliance under FDA Software as a Medical Device (SaMD) guidance.

Retrieval-Augmented Generation (RAG) provides a principled architectural response to these limitations. By coupling the generative capacity of an LLM with a retrieval mechanism that dynamically fetches relevant passages from a curated and continuously updated external knowledge base, RAG systems ground every generated response in verifiable and current evidence. This separation of knowledge storage from reasoning capability enables temporal currency, source attribution, and hallucination reduction simultaneously without requiring costly model retraining.

This paper presents MedRAG, a healthcare-specific conversational assistant built on a progressive RAG architecture. The system is systematically designed and evaluated across three configurations of increasing sophistication: Foundational RAG employing single-stage dense retrieval; Optimised RAG incorporating hybrid retrieval with cross-encoder re-ranking; and Modular RAG deploying multi-strategy orchestration with knowledge graph integration. The knowledge base integrates PubMed abstracts, StatPearls monographs, Cochrane reviews, WHO guidelines, and the UMLS Metathesaurus spanning 4.7 million indexed passages. Experimental results confirm that Modular RAG with GPT-4-Turbo achieves 91.1% accuracy on MedQA-USMLE while lightweight open-source models augmented with the modular pipeline significantly outperform standalone proprietary LLMs

at a fraction of the computational cost. The remainder of this paper is organised as follows: Section II reviews related literature; Section III describes the proposed architecture; Section IV details implementation; Section V presents experimental results; and Section VI concludes with future directions.

II. LITERATURE REVIEW

The field of clinical AI has evolved through several distinct phases. Early computer-based clinical decision support systems embedded in EHR platforms such as Epic and Cerner provided rule-based alert mechanisms for drug interactions and protocol adherence. While highly reliable for explicitly modelled scenarios, these systems proved rigid and difficult to generalise beyond their defined rule coverage, requiring substantial manual maintenance as clinical guidelines evolved over time and medical knowledge expanded.

The transformer architecture introduced by Vaswani et al. in 2017 fundamentally transformed Natural Language Processing. BioBERT and ClinicalBERT established new performance records on biomedical named entity recognition and clinical question answering. GatorTron, pre-trained on over 82 billion words of de-identified clinical text, achieved significant improvements on medical question answering and natural language inference benchmarks. These specialised pre-trained models demonstrated that domain adaptation through biomedical corpus pre-training could meaningfully close the performance gap between general-purpose and clinical-purpose language models.

The emergence of GPT-3 in 2020 marked the beginning of the few-shot learning era in medical NLP. GPT-4 achieved human-competitive performance on USMLE-style questions without medical fine-tuning. Med-PaLM 2 surpassed the USMLE passing threshold and achieved 86.5% accuracy on MedQA, with physicians preferring its responses over those of human physicians across the majority of evaluated clinical utility dimensions. Med42-v2, released in 2024, pushed the state of the art for open medical LLMs to 94.5% on USMLE using a two-stage pipeline combining supervised fine-tuning with Direct Preference Optimisation. Despite these advances, all LLM-based approaches share the fundamental limitation of static parametric knowledge.

Retrieval-Augmented Generation was formally introduced by Lewis et al. in 2020 as a mechanism to augment sequence-to-sequence language models with non-parametric memory via a dense passage index. Subsequent architectural advances include FLARE, which introduced active retrieval triggering on low-confidence generation tokens, and SELF-RAG, which equipped generators with explicit reflection tokens for self-directed retrieval quality assessment. These works demonstrated that retrieval could be made selective and adaptive rather than always-on, improving both efficiency and output quality in different operating regimes.

In the healthcare domain, ChatENT demonstrated a 58.4% error reduction in otolaryngology by combining GPT-4 with a curated clinical knowledge base. The Almanac system integrated an LLM with vector-indexed clinical guidelines, achieving 18% factual accuracy improvement over ChatGPT. MedGraphRAG elevated retrieval by constructing medical knowledge graphs, demonstrating 8–10% improvements over standard RAG. Med-R2 formalised Evidence-Based Medicine principles into the RAG pipeline, with lightweight 7B-parameter models exhibiting accuracy improvements exceeding 77% over vanilla RAG baselines. Despite these advances, no prior work has provided systematic comparative evaluation across three progressive RAG paradigms within a unified framework, nor characterised the performance-complexity trade-off with sufficient granularity to guide architectural selection for specific clinical deployment scenarios. MedRAG directly addresses this research gap.

III. PROPOSED METHOD

A. System Overview

MedRAG is designed as a five-stage sequential pipeline with bidirectional feedback loops enabling iterative query refinement and response verification. A user submits a clinical query through a web interface, mobile application, or EHR-embedded widget. The query passes through the Query Processing Unit, which classifies intent, normalises terminology using the UMLS Metathesaurus, and generates an optimised retrieval query. The processed query is forwarded to the Knowledge Retrieval Engine, which fetches relevant passages from the medical knowledge index. Retrieved passages are quality-filtered, deduplicated, and reordered in the Context Integration Layer. The Response Generation Module synthesises a clinically grounded reply. The Verification and Attribution Layer checks faithfulness and constructs inline citations before output delivery. Three progressive RAG configurations are implemented: Foundational RAG (dense retrieval), Optimised RAG (hybrid dense-sparse with re-ranking), and Modular RAG (multi-strategy orchestration with knowledge graph integration and faithfulness refinement).

B. Three-Tier Progressive Architecture

The Foundational RAG configuration establishes the retrieval baseline using MedCPT bi-encoder dense vector search over a FAISS IVF256 index. MedCPT was contrastively pre-trained on 255,000 citation pairs from PubMed, achieving measurable retrieval precision improvements over general-domain alternatives. Top-5 retrieved passages are concatenated into a structured prompt and forwarded to the LLM generator. This configuration represents the minimum viable RAG deployment with the lowest computational overhead.

The Optimised RAG configuration augments foundational dense retrieval with BM25 sparse keyword matching. Ranked lists from both retrievers are merged using Reciprocal Rank Fusion (RRF, k=60), followed by cross-encoder re-ranking initialised from BiomedBERT. Context compression reduces the assembled prompt to its most informationally dense sentences, and passage reordering places highest-ranked passages at context window boundaries to mitigate the lost-in-the-middle phenomenon identified in prior transformer research.

The Modular RAG configuration introduces a meta-routing layer orchestrating among FAISS dense search, BM25 keyword retrieval, UMLS knowledge graph traversal, and on-

 ISSN: 3107-6513

demand PubMed API retrieval for rare disease queries. Complex multi-hop clinical reasoning queries are decomposed into atomic sub-queries using a chain-of-thought decomposition prompt. A faithfulness verification loop re-triggers retrieval and regeneration whenever the RAGAS faithfulness score falls below 0.80, with the refinement loop activating on 18.4% of initial response drafts across the experimental dataset.

C. Knowledge Base Architecture

The MedRAG knowledge base aggregates content from six primary source categories: PubMed abstracts for broad biomedical literature coverage; StatPearls monographs providing structured clinical reference content; Cochrane Database of Systematic Reviews for synthesised evidence hierarchies; WHO Essential Medicines and Treatment Guidelines for globally applicable clinical protocols; the UMLS Metathesaurus encompassing 3.5 million biomedical concepts from over 200 source vocabularies including SNOMED-CT, MeSH, ICD-10, RxNorm, and LOINC; and institutional clinical guidelines as a customisable deployment layer. Document chunking is performed adaptively based on content type, targeting 512 tokens per chunk with 64-token overlap to preserve contextual continuity. Table I summarises the knowledge base composition.

TABLE I: KNOWLEDGE BASE COMPOSITION

Source	Approx. Indexed Passages
PubMed Abstracts	2,100,000
StatPearls Monographs	890,000
Clinical Guidelines (WHO, Specialty)	1,200,000
UMLS Metathesaurus (3.5M concepts)	510,000

D. Mathematical Formulation

Dense retrieval employs cosine similarity between MedCPT-encoded query and document vectors to rank candidate passages. Sparse retrieval uses the BM25 scoring function with term saturation parameter k_1=1.2 and length normalisation parameter b=0.75. Hybrid merging is performed via Reciprocal Rank Fusion with constant k=60. The RAGAS faithfulness metric quantifies response grounding as the ratio of verified atomic claims to total atomic claims decomposed from the generated response:

$$Similarity(Q,D) = (E(Q)\cdot E(D)) / (||E(Q)|| \times ||E(D)||)$$

$$RRF\ Score(d) = \Sigma_r\ 1/(k + rank_r(d)),\quad k = 60$$

$$Faithfulness = |Verified\ Claims| / |Total\ Atomic\ Claims|$$

E. Modular RAG Algorithm

```
Input:  Clinical query q, corpus K, LLM generator G
Output: Grounded response r with citations C
Step 1: intent_class ← IntentClassifier(q)
Step 2: q_norm ← UMLSNormalise(q)
Step 3: q_reform ← QueryReformulate(q_norm)
Step 4: If IsMultiHop(q): sub_Q ← Decompose(q_reform)
Step 5: For each sub_q in sub_Q:
  a. dense  ← FAISSSearch(Encode(sub_q), k=20)
  b. sparse ← BM25Search(sub_q, k=20)
  c. graph  ← KGTraversal(ExtractEntities(sub_q))
  d. merged ← RRFFusion([dense, sparse, graph])
  e. ranked ← CrossEncoderRerank(sub_q, merged, k=5)
Step 6: context ← ContextIntegrate(ranked)
Step 7: draft_r ← G.generate(prompt, temp=0.1)
Step 8: faith  ← RAGASFaithfulness(draft_r, context)
Step 9: If faith < 0.80: goto Step 4 (refined query)
Step 10: C ← BuildCitations(draft_r, context)
Step 11: Return (draft_r, C)
```

IV. IMPLEMENTATION

A. Technology Stack

The MedRAG system is implemented using a Python-centric technology stack designed for high-throughput clinical deployments. The backend API is developed using FastAPI with OAuth 2.0 and PKCE authentication. LLM inference is served using vLLM with PagedAttention memory management. RAG orchestration integrates LlamaIndex 0.10+ and LangChain 0.2+ components for pipeline management. Three LLM backbone configurations are evaluated: GPT-4-Turbo (proprietary, maximum performance), Mixtral-8x7B-Instruct in INT8 quantisation (open-source mixture-of-experts), and LLaMA-3-8B-Instruct in INT4 quantisation (lightweight deployable). The system supports up to 500 concurrent clinical users with sub-10-second P95 end-to-end response latency. Table II summarises the full technology stack.

TABLE II: TECHNOLOGY STACK

Component	Technology Used
Backend API	FastAPI + OAuth 2.0 / PKCE
LLM Backbone	GPT-4-Turbo / Mixtral-8x7B / LLaMA-3-8B
Embedding Model	MedCPT (768-dim, clinical)
RAG Orchestration	LlamaIndex 0.10 + LangChain 0.2
Vector Database	FAISS IVF256 / ChromaDB
Knowledge Graph	UMLS + SNOMED-CT + RxNorm
Evaluation	RAGAS, BERTScore, ROUGE, S.C.O.R.E.
EHR Integration	HAPI FHIR R4 + SMART on FHIR
Infrastructure	Docker + Kubernetes + Airflow

B. Knowledge Base Construction Pipeline

The knowledge base construction pipeline is implemented as a multi-stage Apache Airflow DAG with daily incremental updates and weekly full refreshes. The ingestion stage fetches PubMed abstracts via the Entrez E-utilities API, updated StatPearls content, and institutional guideline documents. PDF text extraction uses pdfminer.six with custom post-processing correcting hyphenation artefacts, ligature substitution, and header/footer contamination common in clinical document formats. Quality filtering applies language identification, content quality scoring, and MinHash Locality Sensitive Hashing (LSH) for near-duplicate detection. Embeddings are generated using the MedCPT bi-encoder to produce 768-dimensional dense vector representations, inserted incrementally into the FAISS IVF256 index. Index construction for the full 4.7 million passage knowledge base requires approximately 6.2 hours on a single NVIDIA A100 80GB GPU; query-time retrieval completes in under 5 milliseconds at inference time.

C. EHR Integration via SMART on FHIR

Upon SMART on FHIR launch from within an active patient chart, MedRAG receives a patient context bundle containing active problem lists (FHIR Condition resources), current medication lists (FHIR MedicationRequest resources), relevant recent laboratory results (FHIR Observation resources), and allergy records (FHIR AllergyIntolerance resources). This context bundle is processed through a Presidio-based PII de-identification pipeline before incorporation into the retrieval query context, enabling patient-specific personalised recommendations while preserving HIPAA-compliant privacy. The SMART launch mechanism allows seamless integration into Epic, Cerner, and OpenMRS EHR platforms without requiring custom EHR-side development beyond standard SMART app gallery registration.

V. RESULTS AND DISCUSSION

A. Experimental Setup

The experimental evaluation employs five standardised medical benchmark datasets: MedQA-USMLE (1,273 multiple-choice questions spanning USMLE Steps 1–3), MedMCQA (4,183 questions covering 21 medical subjects from AIIMS/NEET PG examinations), PubMedQA (500 biomedical research question-answering items), MMLU-Medical (1,089 questions on clinical knowledge and genetics), and a private Chronic Pain Differential Diagnosis dataset (312 structured diagnostic cases). Twelve experimental configurations are evaluated — three LLM backbones across four RAG configurations — with three independent random seeds per configuration for statistical reliability. Automated metrics include RAGAS Faithfulness, BERTScore F1, ROUGE-L, and BLEU-4. Clinician-administered evaluation uses the S.C.O.R.E. framework (Safety, Correctness, Organisation, Relevance, Evidence) rated by eight board-certified physicians on a five-point scale.

B. Functional Test Cases

Eight functional test cases covering diverse clinical query types were evaluated across all system configurations. All test cases passed in the Modular RAG configuration, with the system correctly handling drug-disease interactions, treatment guideline queries, patient-directed plain-language explanations, drug-drug interaction assessments, out-of-scope refusal, ambiguous query clarification, and administrative operations. Results are summarised in Table III.

TABLE III: FUNCTIONAL TEST CASES AND RESULTS

TC#	Query / Scenario	Result
TC01	Carboplatin dosing in CKD patient on apixaban	Pass
TC02	First-line treatment for dermatomyositis	Pass
TC03	Alteplase dosing for acute ischaemic stroke	Pass
TC04	Patient query: type 2 diabetes diet guidance	Pass
TC05	Tacrolimus + fluconazole interaction assessment	Pass
TC06	Unrelated non-medical query submitted	Pass
TC07	Ambiguous short query: chest pain	Pass
TC08	Admin knowledge base update trigger	Pass

C. Benchmark Performance Results

Standalone LLM baselines reveal the severity of hallucination in medical question answering. GPT-4-Turbo achieves 73.4% on MedQA-USMLE under the No-RAG condition; RAGAS faithfulness averages 0.61, confirming that approximately 39% of atomic claims in LLM-generated medical responses are unverifiable. Foundational RAG improves GPT-4-Turbo accuracy by 6.5 percentage points ($p<0.001$) and faithfulness from 0.61 to 0.74. LLaMA-3-8B exhibits the largest proportional gain at +12.2 pp, confirming that retrieval grounding disproportionately benefits smaller models with limited parametric knowledge. Optimised RAG further advances GPT-4-Turbo to 84.1%, with Retrieval Precision@5 improving from 0.61 to 0.79 through cross-encoder re-ranking. The Modular RAG configuration achieves peak performance: GPT-4-Turbo reaches 91.1% and LLaMA-3-8B reaches 79.7% — surpassing standalone GPT-4-Turbo by 6.3 pp. Table IV presents the comprehensive comparison across all configurations and benchmarks.

TABLE IV: BENCHMARK PERFORMANCE COMPARISON

Config / Model	MedQA (%)	Faithfulness	S.C.O.R.E.
No RAG — GPT-4-Turbo	73.4	0.61	2.94
No RAG — Mixtral-8x7B	61.4	0.54	2.61
No RAG — LLaMA-3-8B	52.1	0.49	2.38
Found. RAG — GPT-4	79.9	0.74	3.89
Found. RAG — Mixtral	69.5	0.70	3.61
Found. RAG — LLaMA	64.3	0.67	3.42

Opt. RAG — GPT-4	84.1	0.83	4.31
Opt. RAG — Mixtral	74.8	0.79	4.07
Opt. RAG — LLaMA	70.2	0.76	3.88
Mod. RAG — GPT-4	91.1	0.91	4.73
Mod. RAG — Mixtral	84.4	0.87	4.49
Mod. RAG — LLaMA	79.7	0.84	4.32

D. Retrieval Quality and Hallucination Analysis

Improving Retrieval Precision@5 from 0.61 (Foundational) to 0.88 (Modular) — a 44% relative gain — reduces the hallucination rate from 26.3% to 9.2%, a 65% relative reduction. This disproportionate effect confirms that retrieval quality is the primary determinant of faithfulness in medical RAG systems. The faithfulness refinement loop activated on 18.4% of initial response drafts across the full experimental dataset, with a single refinement iteration elevating mean faithfulness from 0.83 to 0.91. Average end-to-end latency increases from 42 ms (Foundational) to 394 ms (Modular), with the cross-encoder re-ranking step contributing 61% of the additional latency. Table V presents the full retrieval quality and generation metric comparison.

TABLE V: RETRIEVAL QUALITY AND GENERATION METRICS

Configuration	Precision@5	Faithfulness	Hallucination Rate
Foundational RAG	0.61	0.74	26.3%
Optimised RAG	0.79	0.83	17.1%
Modular RAG	0.88	0.91	9.2%

E. Ablation Study

An ablation study isolates the contribution of each architectural component in the Modular RAG configuration using GPT-4-Turbo as the backbone. Cross-encoder re-ranking is identified as the single most impactful component, with its removal causing a 6.0 percentage-point accuracy drop on MedQA-USMLE. Knowledge graph retrieval is the second most significant contributor (-4.9 pp), confirming that UMLS-based structured knowledge retrieval captures clinically relevant associations not surfaced by dense or sparse text search alone. Query decomposition contributes -3.7 pp, demonstrating that multi-hop reasoning benefits meaningfully from atomic sub-query decomposition. The faithfulness refinement loop contributes -3.2 pp while providing the most significant improvement to the faithfulness metric (0.91 vs 0.83 without it), confirming its importance for clinical safety. Table VI presents the complete ablation results.

TABLE VI: ABLATION STUDY — MODULAR RAG + GPT-4-TURBO

Component Removed	MedQA (%)	Accuracy Δ
Full Modular RAG (Baseline)	91.1	--
- Query Decomposition	87.4	-3.7 pp
- Knowledge Graph Retrieval	86.2	-4.9 pp
- Cross-Encoder Re-ranking	85.1	-6.0 pp
- Context Compression	88.7	-2.4 pp
- Faithfulness Refinement Loop	87.9	-3.2 pp
- UMLS Normalisation	88.1	-3.0 pp

F. Comparison with State-of-the-Art Systems

MedRAG with the Modular RAG pipeline and GPT-4-Turbo matches the performance of Med-Gemini (91.1%) on MedQA-USMLE — the current state-of-the-art closed system — while providing full source attribution and an open-source retrieval pipeline. The fully open-source MedRAG variant (Modular RAG + Mixtral-8x7B) achieves 84.4%, surpassing Med42-v2 (80.4%) and the MEDRAG system by Xiong et al. (71.6%). LLaMA-3-8B with Modular RAG at 79.7% exceeds standalone GPT-4-Turbo at 73.4%, demonstrating that architectural sophistication in retrieval can substitute for parametric model scale — a critical finding for resource-constrained healthcare deployments in low-and-middle-income countries. Clinician S.C.O.R.E. ratings improved from 2.94 out of 5.0 (No-RAG baseline) to 4.73 out of 5.0 (Modular RAG), with 95.7% of evaluating clinicians finding MedRAG responses useful for unfamiliar clinical topics and 92.3% rating the source citations as trustworthy and traceable.

VI. CONCLUSION

This paper presented MedRAG, a healthcare conversational assistant built on a Retrieval-Augmented Generation architecture, designed to overcome the hallucination, knowledge staleness, and attribution limitations of standalone Large Language Models in clinical environments. The system was systematically designed and evaluated across three progressively sophisticated RAG configurations — Foundational, Optimised, and Modular — using a comprehensive evaluation framework combining automated metrics and clinician-administered assessment across five standardised medical benchmarks.

The experimental results establish several important findings. Retrieval quality is the primary performance determinant in medical RAG systems, more influential than generator model scale. Sophisticated retrieval orchestration combining hybrid dense-sparse search, cross-encoder re-ranking, and knowledge graph integration reduces hallucination from 39.1% to 9.2% — a 76% reduction — while improving MedQA-USMLE accuracy by 17.7 percentage points. Lightweight open-source models augmented with the Modular RAG pipeline substantially outperform standalone proprietary frontier models, democratising access to high-quality clinical AI. LLaMA-3-8B with Modular RAG achieves 79.7% on MedQA-USMLE, exceeding standalone GPT-4-Turbo at 73.4%. Source attribution through inline citation construction enables

MedRAG to satisfy FDA SaMD traceability requirements that pure LLM-based systems cannot meet.

Future enhancements will investigate multimodal medical RAG incorporating radiology and pathology images, real-time streaming knowledge updates, federated privacy-preserving retrieval for multi-institutional deployment, pharmacogenomics-informed personalised retrieval, and regulatory pathway validation through prospective clinical trials. MedRAG represents a technically mature and clinically meaningful step toward AI-augmented healthcare that is grounded in evidence, transparent in reasoning, equitable in accessibility, and responsible in deployment.

VII. REFERENCES

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[2] C. Zakka et al., "Almanac: Retrieval-Augmented Language Models for Clinical Medicine," NEJM AI, vol. 1, p. AIoa2300068, 2024.

[3] C. Long et al., "ChatENT: Augmented Large Language Model for Expert Knowledge Retrieval in Otolaryngology," Otolaryngology–Head and Neck Surgery, 2024.

[4] A. Rau et al., "A Context-Based Chatbot Surpasses Radiologists and Generic ChatGPT in Following ACR Appropriateness Guidelines," Radiology, vol. 308, p. e230970, 2023.

[5] Q. Jin et al., "Matching Patients to Clinical Trials with Large Language Models," arXiv:2307.15051, 2023.

[6] J. Wu, J. Zhu, and Y. Qi, "Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation," arXiv:2408.04187, 2024.

[7] K. Lu et al., "Med-R2: Crafting Trustworthy LLM Physicians through Retrieval and Reasoning of Evidence-Based Medicine," arXiv:2501.11885, 2025.

[8] Z. Chen et al., "Towards Omni-RAG: Comprehensive Retrieval-Augmented Generation for Large Language Models in Medical Applications," arXiv:2501.02460, 2025.

[9] K. Singhal et al., "Towards Expert-Level Medical Question Answering with Large Language Models," arXiv:2305.09617, 2023.

[10] C. Christophe et al., "Med42-v2: A Suite of Clinical LLMs," arXiv:2408.06142, 2024.

[11] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[12] J. Devlin et al., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, 2018.

[13] J. Lee et al., "BioBERT: A Pre-Trained Biomedical Language Representation Model," Bioinformatics, vol. 36, pp. 1234–1240, 2020.

[14] G. Xiong et al., "Benchmarking Retrieval-Augmented Generation for Medicine," arXiv:2402.13178, 2024.

[15] S. Es et al., "RAGAS: Automated Evaluation of Retrieval Augmented Generation," in Proceedings of EACL, pp. 150–158, 2024.

[16] A. Asai et al., "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," arXiv:2310.11511, 2023.

[17] Z. Jiang et al., "Active Retrieval Augmented Generation (FLARE)," arXiv:2305.06983, 2023.

[18] Q. Jin et al., "PubMedQA: A Dataset for Biomedical Research Question Answering," in Proceedings of EMNLP, pp. 2567–2577, 2019.

[19] A. Pal et al., "MedMCQA: A Large-Scale Multi-Subject Dataset for Medical Domain QA," in PMLR Machine Learning for Health, pp. 248–260, 2022.

[20] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv:2312.10997, 2023.

[21] K. Saab et al., "Capabilities of Gemini Models in Medicine," arXiv:2404.18416, 2024.

[22] T. F. Tan et al., "A Proposed S.C.O.R.E. Evaluation Framework for Large Language Models," arXiv:2407.07666, 2024.

[23] X. Yang et al., "GatorTron: A Large Clinical Language Model to Accelerate Natural Language Processing in Healthcare," arXiv:2203.03540, 2022.

[24] I. Lopez et al., "Clinical Entity Augmented Retrieval for Clinical Information Extraction," npj Digital Medicine, vol. 8, p. 45, 2025.

[25] A. M. Adly et al., "Gazal-R1: Achieving State-of-the-Art Medical Reasoning with Parameter-Efficient Two-Stage Training," arXiv:2506.21594, 2025.