MedRAG: A Multi-Agent Retrieval-Augmented System for Clinically Accurate Radiology Report Generation

Anish Tiwari	Nehal Shah	Ritika	Vedant B	Sasidhar Reddy
CSA, IISc	CSA, IISc	CSA, IISc	CSA, IISc	CSA, IISc
SR NO. 24423	*SR NO. 24623*	*SR NO. 24244*	*SR NO. 25024*	*SR NO. 24053*

Abstract

'Radiology report generation remains one of the most challenging tasks in medical AI due to hallucinations, missing findings, and lack of factual grounding. Large Language Models (LLMs) often generate fluent but clinically incorrect summaries, limiting safe deployment. To address this, we propose **MedRAG**, a multi-agent, retrieval-augmented generation framework designed for producing clinically accurate radiology reports. MedRAG integrates a 4-agent pipeline—Retrieval, Generation, Validation, and Refinement—coordinated using LangGraph in a stateful, iterative loop. A hybrid dense–sparse retrieval module (Bio-E5 + ColBERTv2 + BM25) ensures high-recall and high-precision evidence selection from biomedical corpora. Report generation is performed using Qwen2-72B-Instruct, while MedGemma and SapBERT-based validation enforce factual correctness through both semantic and evidence-based checks. A Coarse-to-Fine Decoding (C2FD) loop regenerates corrected reports until no errors remain. The system will be evaluated using RadGraph-F1, CheXbert F1, and retrieval recall metrics. Expected outcomes include significantly reduced hallucinations, improved clinical accuracy, and a reliable, modular framework suitable for real-world radiology workflows.

1. Introduction

Radiology departments generate millions of imaging studies every year, and timely, accurate reporting is essential for clinical decision making. While recent advances in Vision-Language Models (VLMs) and Large Language Models (LLMs) have shown promising capability in generating radiology reports, a major challenge remains: **clinical hallucinations**. These occur when models produce confident but incorrect statements, potentially compromising patient safety.

Retrieval-Augmented Generation (RAG) has emerged as a robust strategy to address hallucination by grounding LLM outputs in authoritative medical evidence. However, traditional RAG pipelines are limited by single-shot generation, noisy retrieval, and lack of validation.

To overcome these limitations, we propose **MedRAG**, a multi-agent, iterative, state-aware architecture designed specifically for radiology. MedRAG incorporates hybrid retrieval, multi-stage validation, and loop-based refinement to ensure that every generated report is verifiably supported by retrieved biomedical evidence.

Our objectives are:

- Develop a multi-agent medical RAG system for radiology report generation.
- Integrate hybrid dense + sparse retrieval optimized for medical terminology.
- Implement iterative refinement and validation to eliminate hallucinations.
- Evaluate the system using biomedical NLP metrics such as RadGraph-F1 and CheXbert F1.

2. Methodology

Our proposed MedRAG system is built as a multi–agent, retrieval-augmented medical report generation pipeline. The architecture follows a coarse-to-fine workflow coordinated through LangGraph, where each agent contributes a specialized capability.

2.1 Retrieval Pipeline

The retrieval module follows a hybrid dense–sparse strategy to ensure both semantic and lexical coverage. Dense embeddings are produced using the BioE5 model and indexed using FAISS, while a BM25 index provides sparse retrieval. For each query, the system retrieves a broad pool of passages through weighted fusion of dense and sparse scores. The top candidates are then refined using ColBERTv2, which performs late-interaction token-level re-ranking. This two-stage process ensures that the final evidence is clinically relevant and specific to the patient context.

2.2 Vision Agent

Since radiology reports require image alignment, the Vision Agent processes input X-ray images using Biomed-CLIP to obtain image embeddings. These embeddings are aligned with text embeddings and injected into the retrieval and generation stages. This enables the downstream LLM to generate findings that are image-grounded rather than purely text-based.

2.3 Generation Agent

The Generation Agent uses a quantized Qwen2-72B model served via vLLM. Using the retrieved textual and visual evidence, the agent performs a "coarse" generation step that produces an initial radiology report. Prompts are constructed to enforce factual grounding while allowing the model to synthesize natural-language clinical descriptions.

2.4 Safety–Validation Agent

To ensure factual correctness, we introduce a two-layer validation module. First, each sentence of the generated report is checked for semantic support using a medical embedding model (SapBERT or BioSentVec). Sentences with low similarity to retrieved evidence are flagged. Second, flagged claims are passed to MedGemma, which performs instruction-based medical fact-checking to determine whether each claim is supported, contradicted, or unverifiable. This yields structured validation feedback.

2.5 Coarse-to-Fine Decoding Loop

The system employs an iterative refinement mechanism. If the validation agent detects unsupported statements, the workflow loops into a Refinement Agent, which updates the state with validation feedback and prompts the Generation Agent to regenerate a corrected version of the report. This coarse-to-fine decoding (C2FD) continues until all validations pass or a maximum number of iterations is reached. This ensures that the final report is both clinically accurate and evidence-aligned.

2.6 Infrastructure and Serving

All models are served using vLLM on Wells Fargo server, enabling efficient batched inference. Quantized weights are used for large models to meet memory constraints. LangGraph is used to orchestrate the agents and maintain

determinism across retries. FAISS and BM25 indices support fast, low-latency retrieval across medical corpora.

2.7 Evaluation

We will evaluate MedRAG using:

- **RadGraph-F1**: clinical entity and relation correctness.
- **CheXbert F1**: radiology label classification accuracy.
- **Retrieval Recall@k**: retrieval effectiveness.

These metrics allow us to assess both retrieval quality and final report correctness.

3. Individual Member Contributions

The project is executed by a 5–member team. Each contributor is responsible for a distinct, non-overlapping subsystem of the MedRAG architecture.

- **Sasidhar: Vision Agent**
  Responsible for the radiology image understanding component using BiomedCLIP. Handles image embedding extraction, image-text alignment, integrating embeddings into retrieval and generation, and ensuring cross-modal consistency.
- **Ritika: Retrieval Pipeline**
  Design and implement the hybrid retrieval module. Responsibilities include dataset acquisition, text preprocessing, chunking, Bio-E5 dense embedding generation, FAISS index construction, BM25 sparse indexing and ColBERTv2 re-ranking integration.
- **Vedant: Agent Architecture + Orchestration**
  Designs the multi-agent workflow using LangGraph, defines the `MedRAGState` structure, implements the Coarse-to-Fine Decoding loop, and ensures robust state transitions between Retrieval, Generation, Validation, and Refinement agents.
- **Nehal: Generation & LLM Integration**
  Implements the Qwen2-72B (AWQ-quantized) report-generation pipeline using vLLM. Develop prompting templates for coarse and refined generation, integrate retrieved context, and optimize inference performance.
- **Anish: Safety-Validation + Evaluation**
  Build the two-stage validation module using SapBERT/BioSentVec semantic alignment and MedGemma fact-checking. Design and implement evaluation metrics including RadGraph-F1, CheXbert F1, Consistency Score, and retrieval Recall@k. Ensure output safety and clinical correctness.