HeyDocAI

Radiology Report Explainer with Evidence-Backed Q&A

INFO 7375 - Prompt Engineering for Generative AI
Generative AI Project - Final Report

Done by Ranjithnath Karunanidhi

1. Introduction

Radiology reports are critical medical documents but are often written in highly technical language that is difficult for patients and non-medical users to understand. Misinterpretation can lead to anxiety, misinformation, or incorrect assumptions about health conditions.

HeyDocAI is a Generative AI system designed to bridge this gap by:

- Translating radiology reports into plain English
- Extracting structured clinical information
- Answering user questions using evidence-backed retrieval

The system leverages **Retrieval-Augmented Generation (RAG)** and **Prompt Engineering** to ensure accuracy, transparency, and reliability.

This system is intended for **educational purposes only** and does not provide medical diagnosis or treatment advice.

2. Objectives

The primary objectives of this project are:

1. Build a real-world Generative AI application in the healthcare domain
2. Implement **Retrieval-Augmented Generation (RAG)** using a vector database
3. Design robust **prompt engineering workflows**
4. Provide evidence-backed answers with citations
5. Evaluate system performance using quantitative metrics

3. Core Generative AI Components Used

This project implements **two required components**:

3.1 Prompt Engineering

- System-level instructions defining assistant behavior
- Task-specific prompts:
  - Explanation (Simple / Normal / Clinician)
  - Structured extraction (JSON)
  - Evidence-based Q&A
- Context injection with retrieved evidence
- Guardrails for edge cases and hallucination prevention

3.2 Retrieval-Augmented Generation (RAG)

- Domain-specific knowledge base (radiology PDFs)
- Vector storage using Pinecone
- Semantic retrieval using OpenAI embeddings
- Evidence ranking and filtering before generation
- Citation formatting with source and page numbers

4. System Architecture

The system follows modular, evidence-first architecture.

Architecture Flow

- User inputs a radiology report or question via Streamlit UI
- Input is validated using guardrails
- RAG orchestration retrieves relevant evidence from Pinecone
- Evidence + context injected into prompts
- OpenAI LLM generates grounded responses
- Citations are displayed alongside answers

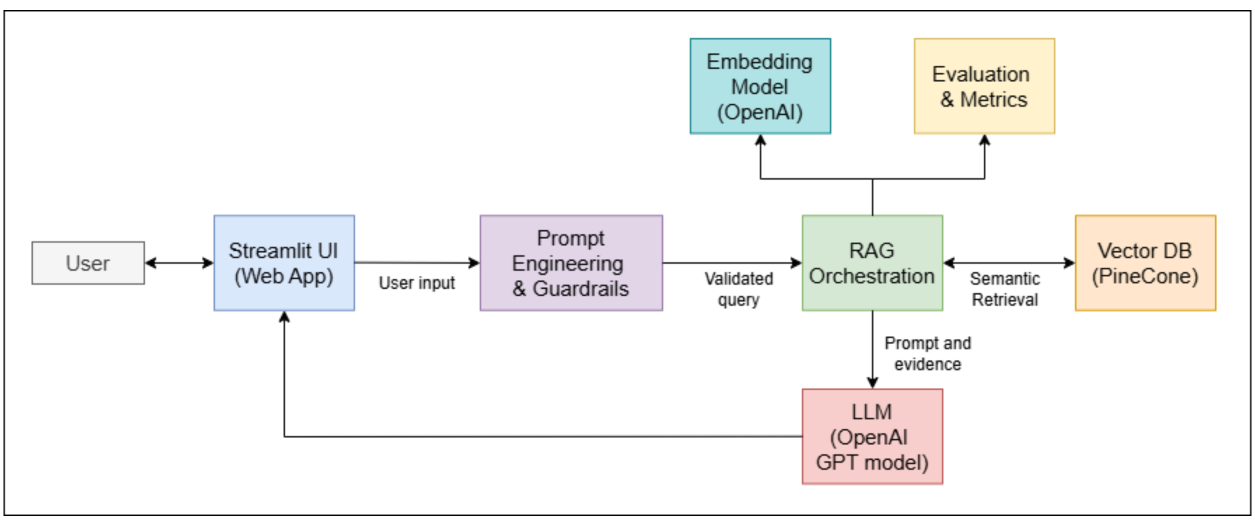


Figure 1. System Architecture Diagram

5. Knowledge Base Construction

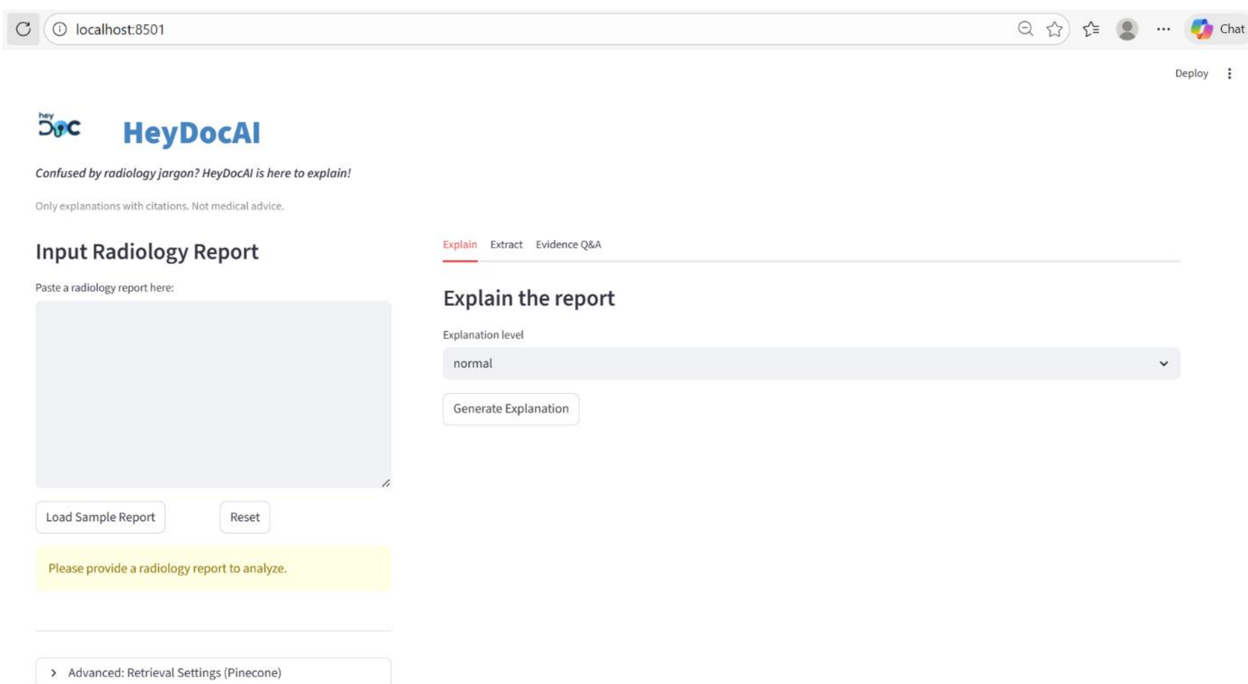
The knowledge base consists of curated, authoritative radiology reference PDFs, including:

- Chest X-ray interpretation guides
- Thoracic imaging glossaries
- Lung pathology references (effusion, consolidation, pneumothorax)

Processing Steps

- PDF text extraction
- Page-level chunking
- Overlapping chunk strategy
- Embedding generation using OpenAI
- Storage in Pinecone vector database

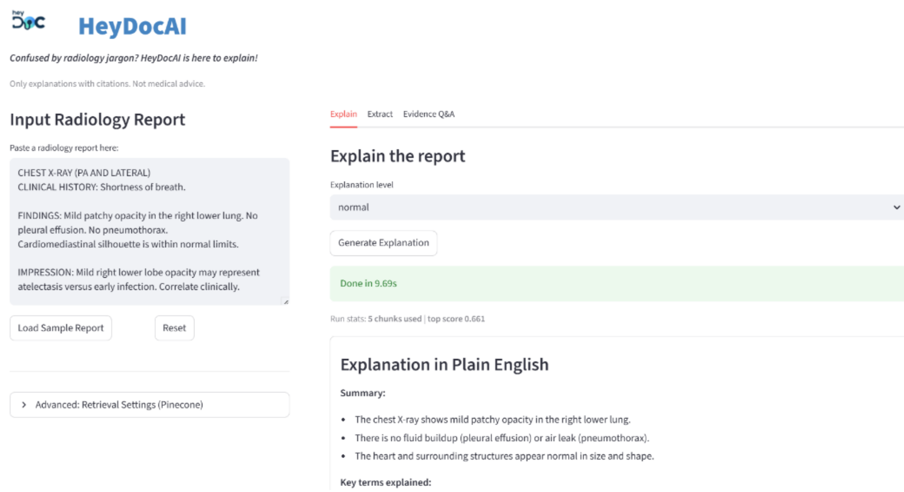
This enables precise semantic retrieval for downstream generation.



6. Application Features

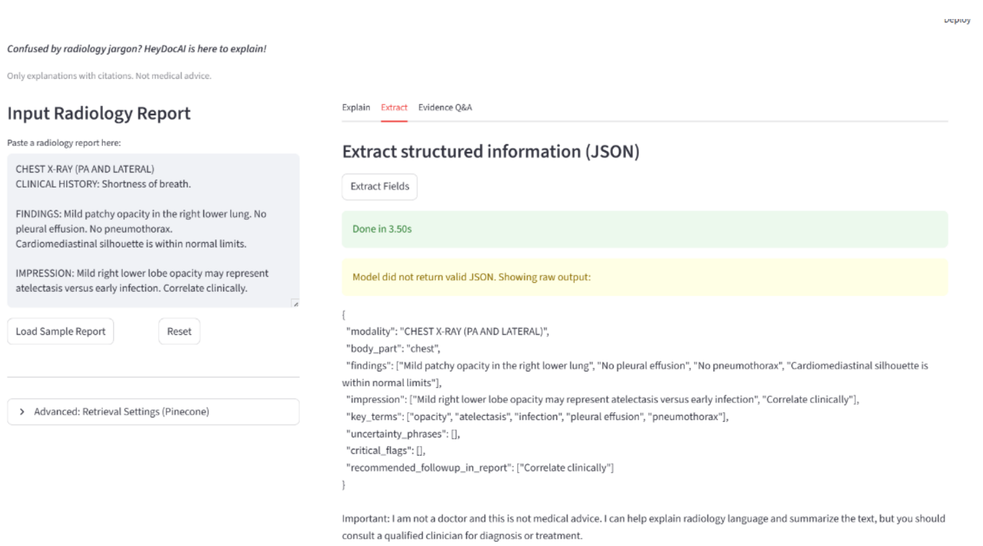
6.1 Explain Radiology Reports

- Converts reports into plain-English explanations
- Supports three explanation levels:
    - Simple (patient-friendly)
    - Normal
    - Clinician-oriented
- Displays retrieval statistics and citations



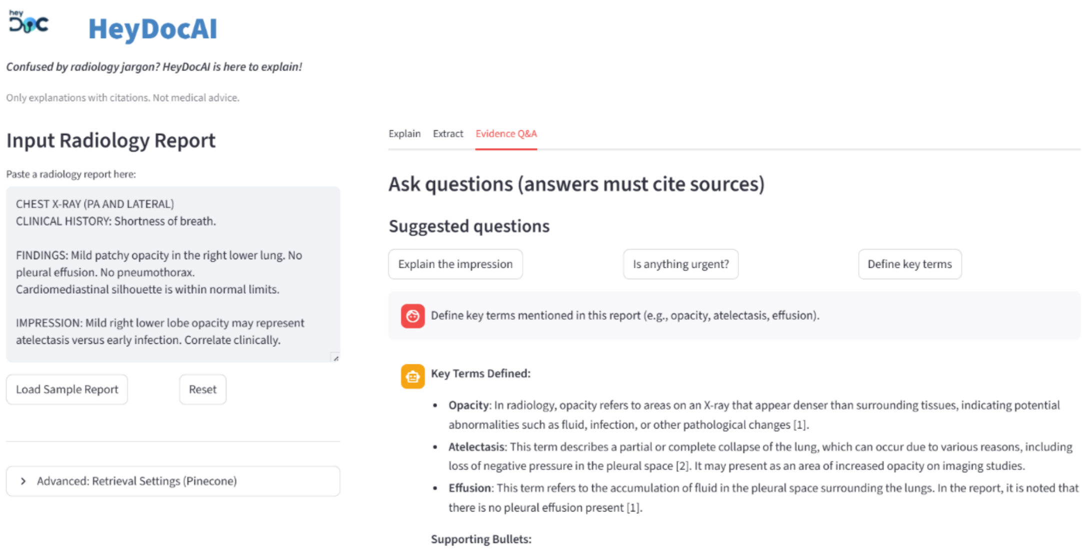
6.2 Structured Information Extraction

- Extracts findings, impressions, and key entities
- Outputs JSON and structured tables
- Useful for downstream clinical workflows



6.3 Evidence-Backed Q&A

- Users ask questions about the report
- Answers generated only if sufficient evidence is retrieved
- All responses include citations
- Chat history maintained for context



7. Guardrails and Safety Measures

To ensure reliability and ethical use, the system implements:

- Empty / invalid input validation
- Minimum similarity score thresholds
- “I don’t know” responses when evidence is insufficient
- Mandatory medical disclaimers
- Citation enforcement for generated answers

These measures significantly reduce hallucination risk.

8. Performance Evaluation

Automated evaluation was conducted using a standardized query set.

Metrics Measured

- Average response latency
- Evidence availability rate

- Citation coverage
- Citation coverage when evidence is available

Results Summary

- Average latency: ~4–5 seconds
- Evidence availability: ~90%
- Citation coverage (given evidence): **100%**

Snipped of eval/results.json:


```
"id": "q1_ggo_definition",
    "type": "definition",
    "question": "What does ground-glass opacity mean in a radiology report?",
    "latency_sec": 12.714,
    "retrieved_chunks_count": 5,
    "citations_present": true,
    "citations_ui": [
      "[1] thoracic_imaging_glossary_fleischner.pdf \u2014 page 35 (score: 0.600)",
      "[2] thoracic_imaging_glossary_fleischner.pdf \u2014 page 6 (score: 0.592)",
      "[3] thoracic_imaging_glossary_fleischner.pdf \u2014 page 9 (score: 0.576)",
      "[4] thoracic_imaging_glossary_fleischner.pdf \u2014 page 19 (score: 0.575)",
      "[5] thoracic_imaging_glossary_fleischner.pdf \u2014 page 47 (score: 0.560)"
    ],
    "top_sources": [
      {
        "source": "thoracic_imaging_glossary_fleischner.pdf",
        "page": 35,
        "score": 0.6001
      },
      {
        "source": "thoracic_imaging_glossary_fleischner.pdf",
        "page": 6,
        "score": 0.5917
      },
```


9. Challenges and Solutions

Challenge 1: Hallucinated Medical Responses
Solution: Strict evidence validation and refusal to answer without sufficient retrieval.

Challenge 2: Overly Technical Explanations
Solution: Multi-level prompt templates tailored to user expertise.

Challenge 3: Long, Complex Radiology Documents
Solution: Chunking + overlap strategy with ranking and filtering.

10. Ethical Considerations

- No personal health data is stored
- No diagnosis or treatment recommendations
- Transparent citation of sources
- Bias minimized by grounding answers in trusted references
- Clear disclaimers shown in UI

11. Future Improvements

- Multimodal support (X-ray images + text)
- Expanded medical domains (CT, MRI, ultrasound)
- Improved ranking strategies (MMR, re-ranking models)
- Cloud deployment with authentication
- Multilingual support

12. Conclusion

HeyDocAI demonstrates how Generative AI can be safely and effectively applied in healthcare education. By combining **Prompt Engineering** with **Retrieval-Augmented Generation (RAG)**, the system delivers transparent, reliable, and user-friendly explanations of complex medical documents.

This project showcases a practical, real-world application of modern Generative AI techniques while addressing ethical and reliability concerns.

13. Appendix

- GitHub: *https://github.com/Ranjithnathk/HeyDoc-AI-RAG-Medical-Report-Explainer*
- Screenshots: docs/screens/
- Evaluation Results: eval/results.json
- Video Demonstration