Agentic Multimodal Retrieval-Augmented Generation (RAG) for Healthcare

1. Introduction

Large Language Models (LLMs) have demonstrated impressive abilities in reasoning, summarization, and general knowledge retrieval across diverse domains. However, when these models are deployed in safety-critical environments such as healthcare, their limitations become more pronounced. LLMs are known to hallucinate producing confident but incorrect statements particularly when they lack explicit grounding in clinical evidence. This makes them unreliable in medical contexts, where accuracy, transparency, and interpretability are essential.

Although recent multimodal medical AI systems such as Med-PaLM and LLaVA-Med can process both text and images, they still rely on largely static retrieval pipelines and opaque inference mechanisms. They seldom provide adaptive retrieval, detailed provenance tracking, or modality-specific preprocessing. As a result, clinicians or researchers cannot easily determine which cases, images, or textual evidence contributed to a model's recommendation.

This capstone project attempts to address these limitations by designing and implementing an **Agentic Multimodal Retrieval-Augmented Generation (RAG) system**. The system integrates radiology images with clinical text, performs multimodal fusion, and employs a retrieval-driven agent capable of generating verifiable, evidence-grounded explanations. Instead of acting as a diagnostic tool, the prototype serves as a research and educational framework for analyzing how multimodal RAG pipelines behave under realistic medical queries. The system aims to generate multimodal embeddings, perform adaptive similarity-based retrieval, produce transparent clinical summaries, and support conversational interactions where every generated claim is tied to retrieved evidence.

---

2. Problem Statement and Objectives

Unreliable reasoning from ungrounded LLMs poses significant risks in clinical settings. Medical imaging introduces additional complexity, as models must interpret heterogeneous modalities, metadata, and textual descriptions simultaneously. Several challenges emerge from this landscape. Existing systems provide limited transparency into their retrieval processes, making it difficult for users to verify which evidence influenced a prediction. Image and text processing pipelines are frequently treated independently, which diminishes the advantage of multimodality. Public datasets, such as MedPix, are relatively small and inconsistently structured, complicating preprocessing workflows. Furthermore, typical RAG systems retrieve documents only once and lack iterative, agentic verification mechanisms that detect inconsistencies or refine outputs.

This project therefore focuses on five core objectives: to construct a unified multimodal pipeline capable of embedding medical images and clinical text; to build a scalable vector index supporting fast similarity search; to design a retrieval layer capable of handling text-only, image-only, and combined queries; to integrate this retrieval layer with an LLM-based agent that generates provenance-tracked clinical explanations; and to evaluate the system using both classical information-retrieval metrics and RAG-specific measures such as faithfulness and context relevance.

3. System Architecture Overview

The system architecture consists of four major modules implemented in Python: a data loading and preprocessing module, an embedding generation module, a vector indexing module, and a retrieval module. The data loading component ingests heterogeneous MedPix data including case-level JSON files, image descriptions, and file paths then merges them into a unified tabular dataset that includes diagnoses, findings, anatomical locations, imaging modalities, and derived body-region categories.

The embedding module encodes both images and text. Images are processed through BiomedCLIP, while clinical text is encoded using BioClinicalBERT. Because these models produce embeddings of different dimensionalities, an alignment step projects them into a common latent space. Multimodal fusion is performed either by concatenating the embeddings or by producing a weighted average, enabling the system to represent cases in a unified format.

To enable large-scale retrieval, the system builds FAISS-based approximate nearest-neighbor indices over the hybrid embeddings. These indices are optimized for fast query responses and support adjustable clustering and probing strategies.

At the front of the system is a retrieval layer that accepts text, image, or multimodal queries from the user, constructs a comparable embedding through the same preprocessing and encoding pipeline, and performs similarity search against the index. Results are returned with associated metadata and fed into an LLM router that communicates with external model providers such as OpenAI, OpenRouter, or Ollama. The LLM ultimately generates a grounded explanation referencing the retrieved evidence.

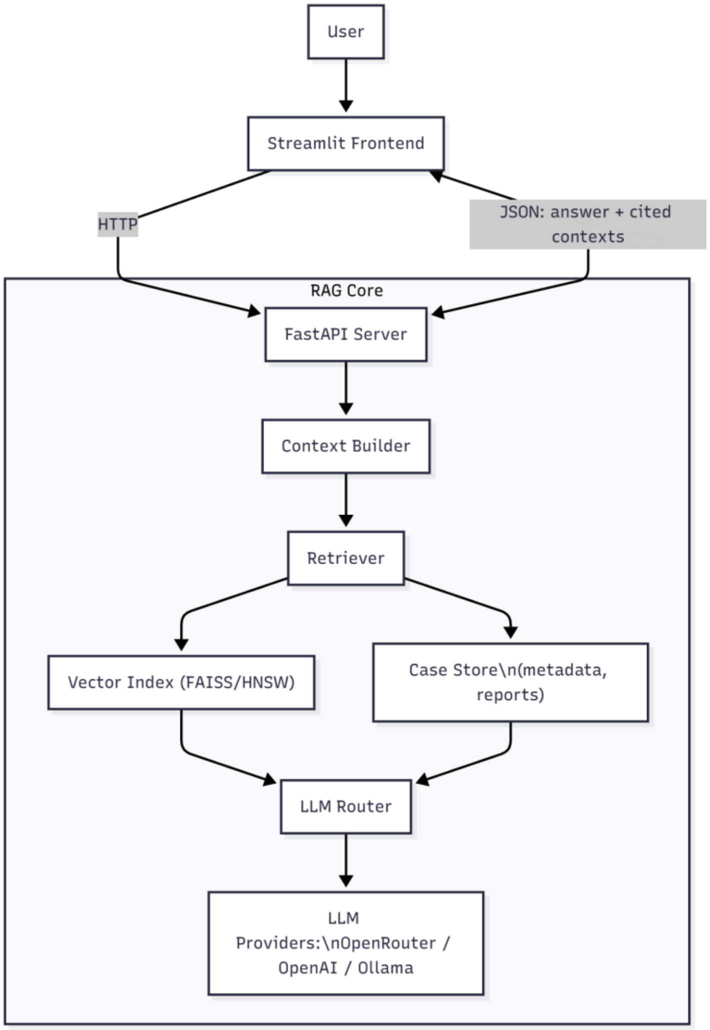

User
Streamlit Frontend
HTTP
JSON: answer + cited contexts
RAG Core
FastAPI Server
Context Builder
Retriever
Vector Index (FAISS/HNSW)
Case Store\n(metadata, reports)
LLM Router
LLM Providers:\nOpenRouter / OpenAI / Ollama

Figure 1. High-level Agentic Multimodal RAG Architecture
(User interacts with a Streamlit frontend; queries are sent over HTTP to the FastAPI-based RAG core, which builds context from the vector index and case store, then routes requests to external LLM providers via the LLM router and returns JSON answers with cited contexts.)

4. Methodology and Detailed Module Design

4.1 Data Loading and Preprocessing (data_loader.py)

The data loader is responsible for transforming fragmented MedPix data into a structured dataset. It handles directory resolution, loads case-level and image-level JSON files, and unifies them into a denormalized DataFrame where each row represents a single medical image paired with relevant clinical context. A key component of preprocessing is body-region classification, where diagnoses are analyzed to infer anatomical categories such as brain, spine, cardiac, or musculoskeletal regions. This allows for more meaningful evaluation of retrieval tasks across clinically relevant subdivisions.

During integration, each image is associated with its corresponding descriptive text, metadata, modality, and patient information. Images lacking sufficient information are skipped, and summary statistics are produced to highlight dataset imbalance or missingness. Utility functions allow researchers to extract modality-specific, diagnosis-specific, or region-specific subsets for targeted experiments. The design favors denormalization to simplify downstream embedding and indexing operations.

4.2 Multimodal Embedding Generation (embeddings.py)

Embedding generation is a central component of the system. Image preprocessing is modality-aware: CT scans undergo Hounsfield Unit windowing tailored to soft-tissue visualization, while MRI scans are standardized through z-score normalization. These transformations ensure that the model receives consistent and clinically meaningful intensity distributions.

Images are encoded using BiomedCLIP or, if unavailable, a general CLIP model. Text fields including diagnoses, findings, captions, and history are concatenated into a single clinical description and encoded using BioClinicalBERT or a fallback sentence-transformer. After encoding, PCA-based alignment ensures that image and text embeddings occupy a comparable vector space, enabling fusion. The system supports both concatenation and weighted-average fusion, allowing experimentation with different multimodal combination strategies.

A coordinating `MultimodalEmbedder` orchestrates this process by producing a dictionary of embedding types, each of which can be saved and indexed. The module is robust to missing images or malformed text and includes configuration options for batch size, device placement, and fusion parameters.

4.3 Vector Indexing and Similarity Search (indexing.py)

Similarity search is facilitated through FAISS, using IVF-Flat indices with inner-product similarity. The indexing module validates configuration parameters, constructs clustering structures, trains centroids, and inserts all hybrid embeddings into the index. Metadata is stored separately in JSON format to preserve detail without bloating the index file.

Queries involve two stages: identifying candidate clusters and then performing fine-grained similarity search within them. The index returns ranked results with similarity scores and accompanying metadata. The design balances retrieval quality with computational efficiency, making interactive exploration feasible even on modest hardware.

4.4 Retrieval System and Query Handling (retrieval.py)

The retrieval module serves as the system's primary interface. It processes user inputs, constructs embeddings consistent with those in the index, and executes similarity search. Three query modes are supported. When both image and text are provided, their embeddings are fused to match the hybrid index. When only an image or only text is provided, the missing modality is replaced with a zero vector to maintain dimensional consistency.

The module includes wrapper functions for different query styles and a batch retrieval mode useful for evaluation experiments. It also formats results into a readable structure, displaying similarity scores and key metadata for downstream interpretation. Usage examples embedded in the script serve as informal regression tests, ensuring correct integration across modules.

---

5. Evaluation and Results

The system is evaluated along two complementary axes: classical retrieval metrics and RAG-oriented evaluation of generated answers.

5.1 Retrieval Metrics

Using held-out queries derived from the MedPix dataset, we compute:

- **Precision@k** – fraction of retrieved cases in the top-k that share the correct diagnosis or body region.
- **Mean Reciprocal Rank (MRR)** – how high the first relevant item appears in the ranking.
- **F1 Score** – harmonic mean of precision and recall over relevant items.
- **ROUGE-L** – compares generated textual summaries with reference findings or diagnoses where available.

5.2 RAGAS-Based Generation Evaluation

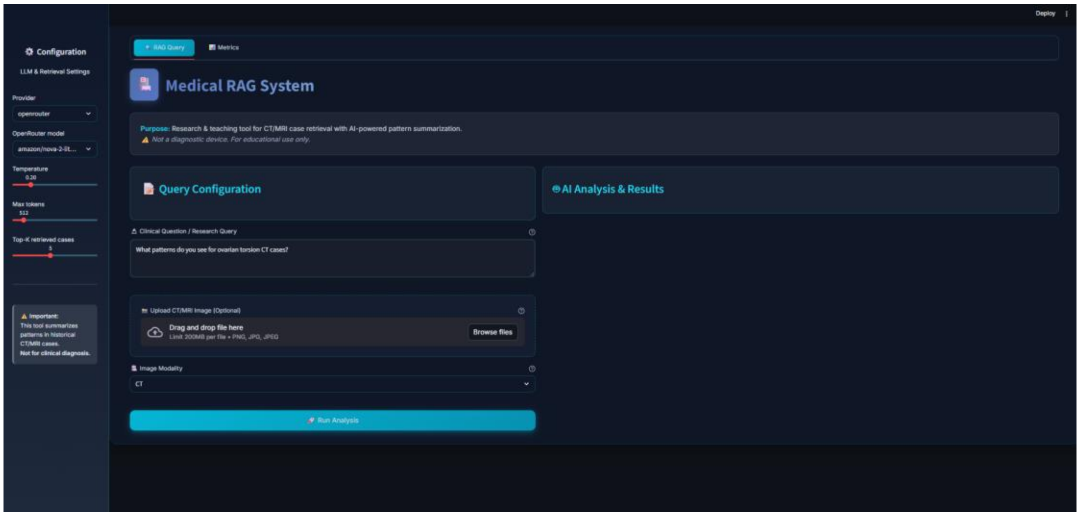
To assess the quality of LLM-generated responses when conditioned on retrieved evidence, we use RAGAS-style metrics:

- **Faithfulness** – whether the answer stays consistent with the retrieved context, penalizing hallucinations.
- **Answer Relevance** – alignment of the answer with the original query.
- **Context Relevance** – whether the retrieved context is actually useful for answering the question.
- **Latency** – wall-clock time from user query to final answer, reflecting both retrieval and generation.

5.3 Qualitative Demonstration

Three main UI views were created to demonstrate the system:

1. A **query screen** that lets the user upload an image, enter a free-text question, and select the modality.



Configuration

LLM & Retrieval Settings

Provider: openrouter

OpenRouter model: amazon/nova-2-lit...

Temperature: 0.20

Max tokens: 512

Top-K retrieved cases: 5

⚠ Important: This tool summarizes patterns in historical CT/MRI cases. **Not for clinical diagnosis.**

Medical RAG System

Purpose: Research & teaching tool for CT/MRI case retrieval with AI-powered pattern summarization.

⚠ *Not a diagnostic device. For educational use only.*

Query Configuration

AI Analysis & Results

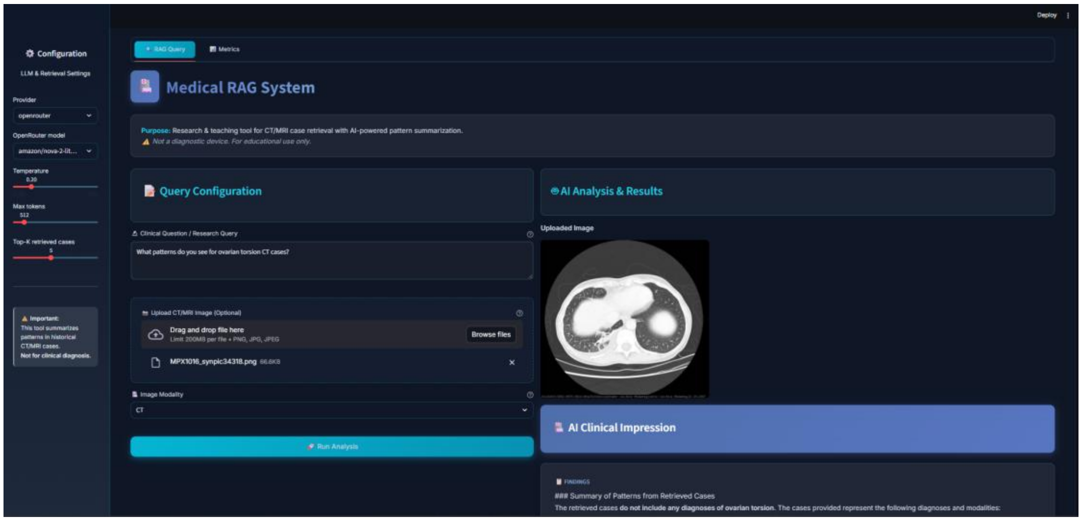
Clinical Question / Research Query: What patterns do you see for ovarian torsion CT cases?

Upload CT/MRI Image (Optional)

Drag and drop file here
Limit 200MB per file • PNG, JPG, JPEG

Image Modality: CT

2. A **result viewer** that shows similarity scores, image thumbnails, key metadata, and a grounded explanation from the LLM.



Configuration

LLM & Retrieval Settings

Provider: openrouter

OpenRouter model: amazon/nova-2-lit...

Temperature: 0.20

Max tokens: 512

Top-K retrieved cases: 5

⚠ Important: This tool summarizes patterns in historical CT/MRI cases. **Not for clinical diagnosis.**

Medical RAG System

Purpose: Research & teaching tool for CT/MRI case retrieval with AI-powered pattern summarization.

⚠ *Not a diagnostic device. For educational use only.*

Query Configuration

AI Analysis & Results

Clinical Question / Research Query: What patterns do you see for ovarian torsion CT cases?

Upload CT/MRI Image (Optional)

Drag and drop file here
Limit 200MB per file • PNG, JPG, JPEG

MPX1016_synpic34318.png 66.6KB

Image Modality: CT

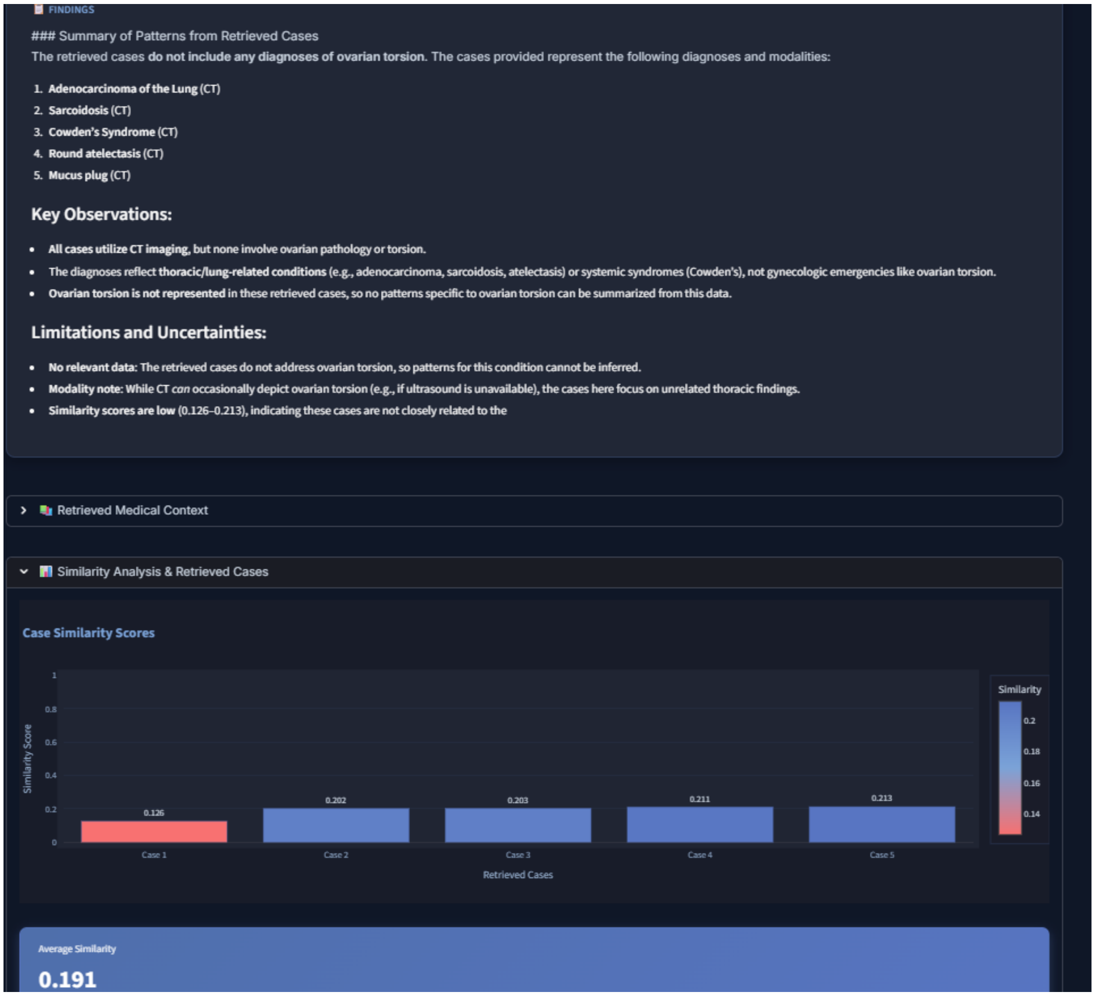
Uploaded Image

AI Clinical Impression

FINDINGS

Summary of Patterns from Retrieved Cases

The retrieved cases **do not include any diagnoses of ovarian torsion.** The cases provided represent the following diagnoses and modalities:



FINDINGS

Summary of Patterns from Retrieved Cases

The retrieved cases **do not include any diagnoses of ovarian torsion.** The cases provided represent the following diagnoses and modalities:

1. **Adenocarcinoma of the Lung** (CT)
2. **Sarcoidosis** (CT)
3. **Cowden's Syndrome** (CT)
4. **Round atelectasis** (CT)
5. **Mucus plug** (CT)

Key Observations:

- **All cases utilize CT imaging,** but none involve ovarian pathology or torsion.
- The diagnoses reflect **thoracic/lung-related conditions** (e.g., adenocarcinoma, sarcoidosis, atelectasis) or systemic syndromes (Cowden's), not gynecologic emergencies like ovarian torsion.
- **Ovarian torsion is not represented** in these retrieved cases, so no patterns specific to ovarian torsion can be summarized from this data.

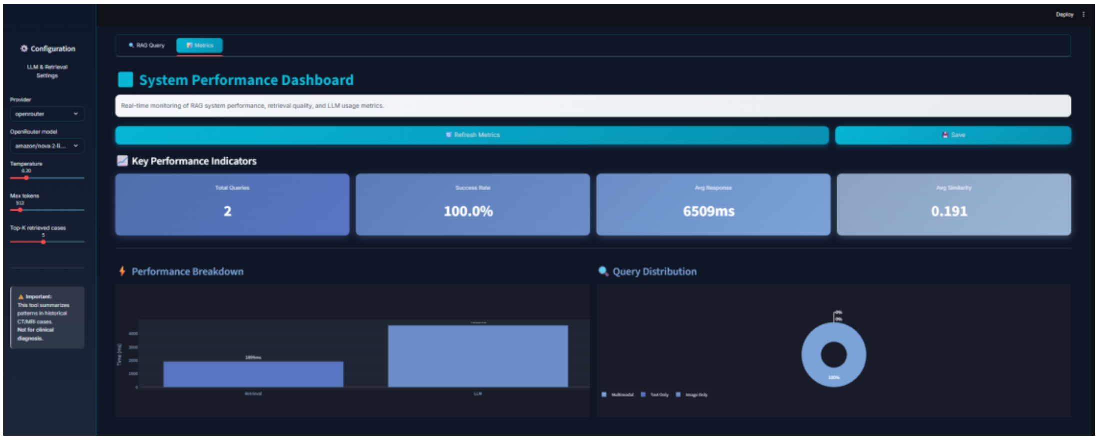
Limitations and Uncertainties:

- **No relevant data:** The retrieved cases do not address ovarian torsion, so patterns for this condition cannot be inferred.
- **Modality note:** While CT *can* occasionally depict ovarian torsion (e.g., if ultrasound is unavailable), the cases here focus on unrelated thoracic findings.
- **Similarity scores are low** (0.126–0.213), indicating these cases are not closely related to the

Average Similarity

0.191

3. A **monitoring dashboard** containing charts for latency distributions, embedding dimensionality, and basic model-behavior diagnostics.



Configuration

LLM & Retrieval Settings

Provider: openrouter

OpenRouter model: amazon/nova-2-li...

Temperature: 0.20

Max tokens: 512

Top-K retrieved cases: 5

⚠ Important: This tool summarizes patterns in historical CT/MRI cases. Not for clinical diagnosis.

System Performance Dashboard

Real-time monitoring of RAG system performance, retrieval quality, and LLM usage metrics.

Key Performance Indicators

Total Queries	Success Rate	Avg Response	Avg Similarity
2	100.0%	6509ms	0.191

Together, these evaluations show that the retrieval layer returns clinically coherent neighbors and that grounding information can be surfaced to the user alongside generated explanations.

6. Limitations

Despite its contributions, the current system has several limitations:

- **Dataset size and coverage** – The MedPix subset used in this project contains roughly two thousand CT and MRI images, which is small compared to real hospital PACS archives. Many pathologies are under-represented or absent.
- **Modality scope** – Only CT and MRI scans are included; other common modalities such as X-ray, ultrasound, and PET are not yet supported.
- **Lack of clinical validation** – No formal user study with clinicians has been conducted. All evaluation is based on offline metrics and qualitative inspection.
- **Single-agent reasoning** – The current agent performs a single retrieve-then-generate workflow, with optional iterative refinement, but does not involve multiple specialized agents (e.g., one for retrieval, one for verification, one for summarization).
- **Domain shift** – Public teaching datasets differ from real hospital data in prevalence, acquisition protocols, and annotation style, limiting direct transferability of performance numbers to real-world environments.

7. Future Work

Future extensions fall into three broad categories: technical, design, and research.

7.1 Technical Scope

- Incorporate additional modalities such as X-ray, ultrasound, and PET, with modality-specific preprocessing and normalization.
- Extend the embedding module to support more powerful **multimodal transformers** that jointly attend over image patches and text tokens.
- Integrate **graph-based RAG** and long-term memory so that the agent can reason over patient-level trajectories and knowledge graphs, not just single studies.
- Experiment with alternative indexing structures (e.g., HNSW) and **dynamic indices** that support insertion of new clinical cases without full rebuilds.

7.2 Design and Product Scope

- Evolve the UI into a **deployment-ready system** with authentication, study selection, and exportable reports.
- Introduce an **interactive feedback loop** where clinicians can up-vote, down-vote, or correct retrieved cases and generated summaries, and feed that signal back into the retrieval or prompting strategy.

- Provide richer **visualizations of similarity** (for example, t-SNE or UMAP projections of embeddings by body region or diagnosis).

7.3 Research Scope

- Conduct a structured **clinical evaluation** where radiologists or residents use the system for case-based teaching or decision support, and collect usability scores and qualitative feedback.
- Quantify **hallucination rate** and **citation correctness** by manually auditing a subset of generated answers.
- Study retrieval behavior and robustness under **distribution shift** by introducing synthetic noise, adversarial queries, or new pathologies.
- Explore **procedural reasoning** using surgical guidelines—from retrieving the correct guideline segments to verifying each generated reasoning step against those segments.

---

8. Ethical and Societal Considerations

Because the system deals with medical content, several ethical principles guide its design:

- **Research-only disclaimer** – The prototype is explicitly not a diagnostic tool and must not be used to make individual patient decisions. All outputs are for research, teaching, and demonstration purposes.
- **Bias and fairness** – Public radiology datasets under-represent certain demographics and pathologies, which can bias retrieval results. Any future deployment would require more representative data and explicit fairness analysis.
- **Explainability and accountability** – Every generated answer is linked to the images and text that supported it, improving transparency and enabling human oversight. Logs of queries and retrieved cases help reconstruct why the model produced a given output.
- **Data governance** – While MedPix is a teaching dataset, real clinical deployment would require strict access controls, de-identification procedures, and alignment with HIPAA and institutional review policies.

---

9. Conclusion

This capstone project implemented an end-to-end **Agentic Multimodal RAG system for healthcare**. Starting from raw MedPix JSON files and image folders, we designed data loaders, multimodal embedding pipelines, FAISS indices, and a flexible retrieval API that jointly support text-only, image-only, and multimodal queries.

The resulting framework demonstrates that it is feasible to build a transparent, retrieval-driven assistant that grounds its responses in radiology images and associated clinical text. Although the current prototype is limited to a small dataset and has not yet undergone clinical validation, it provides a strong foundation for future work on multimodal clinical decision support, case-based teaching tools, and research on retrieval-augmented reasoning in high-risk domains.

10. References

1. M. Douze, G. Szilvasy, P.-E. Mazaré, and H. Jégou, "The Faiss library," *arXiv preprint*, 2024.
   Available at: https://arxiv.org/abs/2401.08281
2. S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," *arXiv preprint*, 2023.
   Available at: https://arxiv.org/abs/2309.15217
3. 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," *NeurIPS*, 2020.
   Available at: https://arxiv.org/abs/2005.11401
4. A. Carré et al., "Standardization of brain MR images across machines and protocols: Bridging the gap for MRI-based radiomics," *Scientific Reports*, vol. 10, no. 12340, 2020.
   Available at: https://www.nature.com/articles/s41598-020-69298-z
5. Z. Abboud and S. Kadoury, "Impact of train- and test-time Hounsfield unit window variation on CT segmentation of liver lesions," *Medical Imaging 2023: Image Processing*, Proc. SPIE 12464, 2023.
   Available at: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12464/2653974