Agentic Retrieval-Augmented Generation (RAG) for Healthcare AI

Project proposal & Statement of Work

Meghana Maringanti, Project Manager

Abhimanyu Pandey

Manglam Srivastav

Advisor: Dr. Eung-Joo Lee

Date: October 1st, 2025

Revision History Table

Version	Summary of Changes	Date
V1	*First draft of the project proposal*	*10/01/25*
	Updated with team/mentor feedback	*Date TBD*
	Final review before submission	*Date TBD*
	Version submitted for signatures	*Date TBD*

Contents

1. Executive Summary

The Executive Summary was written by Abhimanyu Pandey.

Our project, **Agentic Retrieval-Augmented Generation (RAG) for Healthcare AI**, is an intelligent, multimodal system that combines medical imaging and clinical text with advanced generative reasoning to deliver accurate, explainable, and user-friendly responses to healthcare queries. The system is built around four core features: **Multimodal Data Indexing** to embed and organize diverse medical data, **Adaptive Context Selection** to dynamically choose the most relevant evidence, **Agentic Orchestration** to coordinate retrieval and reasoning steps, and a **Chatbot Interface** that enables natural, conversational interactions for clinicians and patients alike.

The need for such a system is pressing. Large language models have shown promise in medicine but suffer from hallucinations and unreliable outputs when used in clinical contexts. According to Grand View Research, the global AI in healthcare market was valued at **USD 26.6 billion in 2024** and is projected to grow to **USD 187.7 billion by 2030** at a CAGR of 38.6%. Yet existing multimodal systems, such as Med-Flamingo and LLaVA-Med, lack adaptive retrieval and explainability, leading to errors and clinician mistrust. By integrating agent-driven decision making and provenance tracking, our approach provides a unique solution that addresses reliability, trust, and usability gaps in current offerings.

Development will take place over the semester using Python, Hugging Face, PyTorch, FAISS, and a web-based interface. Table 1 shows the preliminary division of responsibilities across our team. At the end of the project, we will deliver a working prototype that demonstrates real-world healthcare question answering across text and imaging inputs, with evidence logging and role-adapted outputs for patients and clinicians.

Team Member	Feature responsibility
Abhimanyu Pandey	Chatbot Interface, Agentic Orchestration
Mangalam Srivastav	Multimodal Data Indexing, Agentic Orchestration
Meghana Maringanti	Adaptive Context Selection, Agentic Orchestration

Table 1 Preliminary Subsystem Responsibilities

2. User/Market research

The User/Market research was written by Manglam Srivastav.

Overall Market

The healthcare AI sector is rapidly expanding, projected to surpass **USD 180 billion by 2030**. Within this, **Retrieval-Augmented Generation (RAG)** and **multimodal AI** are emerging submarkets, expected to grow from just over USD 1 billion in 2023 to more than USD 11 billion by 2030. This demonstrates strong commercial and clinical demand for systems that can combine medical knowledge retrieval with generative reasoning.

Existing Competitors

Several systems illustrate current progress:

- **Med-Flamingo** and **RadFM**: extend LLMs with vision-language models but remain limited in retrieval adaptivity.
- **OpenEvidence**: physician-only search engine that grounds responses in PubMed but lacks multimodal integration.
- **LLaVA-Med**: provides visual question answering for radiology but does not support provenance or patient-friendly explanations.

  These tools validate the demand but leave clear gaps in adaptive retrieval, agent coordination, and transparency.

User Insights

Early empathy interviews with clinical students and researchers highlight three recurring needs:

1. **Trustworthy evidence** - responses must cite reliable medical sources, not opaque model reasoning.
2. **User-specific communication** - clinicians want precise guideline references, while patients want simplified explanations.
3. **Interactive refinement** - users often want to ask follow-up questions or upload new images without restarting the process.

Our system directly addresses these needs by providing **adaptive retrieval with provenance**, a **chatbot interface with role-specific outputs**, and **agentic orchestration** for flexible reasoning. Together, these ensure adoption in both clinical and patient education contexts.

3. Product Features

This section was written by Meghana Maringanti. Parameter tables were prepared by Abhimanyu Pandey, Manglam Srivastav.

Feature 1: Multimodal Data Indexing

Chunk and embed clinical text (BioClinicalBERT) and images (CLIP / BiomedCLIP) and store multimodal vectors in a FAISS index for fast, scalable retrieval.

Parameter	Min	Max	Comments
Chunk size (tokens text)	512	1024	Balance context vs. index size; smaller chunks improve localization.
Embedding dimension (text)	512	1024	Typical BioClinicalBERT embeddings ≈768; reserve up to 1024 for alternatives.
Embedding dimension (Image)	512	1024	CLIP/BiomedCLIP variants vary; design to accept 512–1024 dims.

Table 2: Multimodal Indexing Parameters

Feature 2: Adaptive Context Selection

Dynamically choose the number *k* of retrieved contexts (and which chunks) using similarity gaps and context diversity heuristics to minimize irrelevant or redundant contexts passed to the LLM.

Parameter	Min	Max	Comments
K (contexts selected)	1	10	Dynamic selection; default 3–6 for LLM prompts.
Similarity gap threshold	0.05	0.50	If top-score − next-score > threshold, pick top only.
Context redundancy filter	0%	50%	Max allowed overlap between chosen contexts.

Table 3: Adaptive Context Selection Parameters

Feature 3: Agentic Orchestration

AI agent(s) orchestrate the pipeline—route queries to retrievers, fuse multimodal outputs, decide whether to call specialist modules (e.g., image annotator), and invoke the LLM with a structured prompt.

Parameter	Min	Max	Comments
Agent routing accuracy	0.7	0.95	Measured on test queries (domain routing + module selection).
Orchestration latency	100 ms	2,000 ms	Target to keep orchestration overhead low compared to inference cost.
Number of parallel agents	1	5	Prototype: single orchestrator; stretch: multi-agent pipelines.
Error/failover rate	0%	5%	Define fallbacks when a module fails (e.g., degrade gracefully).

Table 4: Agentic Orchestration Parameters

Feature 4: Chatbot Interface

A conversational web UI that accepts text and image uploads, shows answers with provenance links, and provides role selection (patient/clinician).

Parameter	Min	Max	Comments
Accessibility level	WCAG 2.1 A	WCAG 2.1 AA	Level A provides basic accessibility. Level AA adds higher contrast, captions, consistent navigation, and resize support.

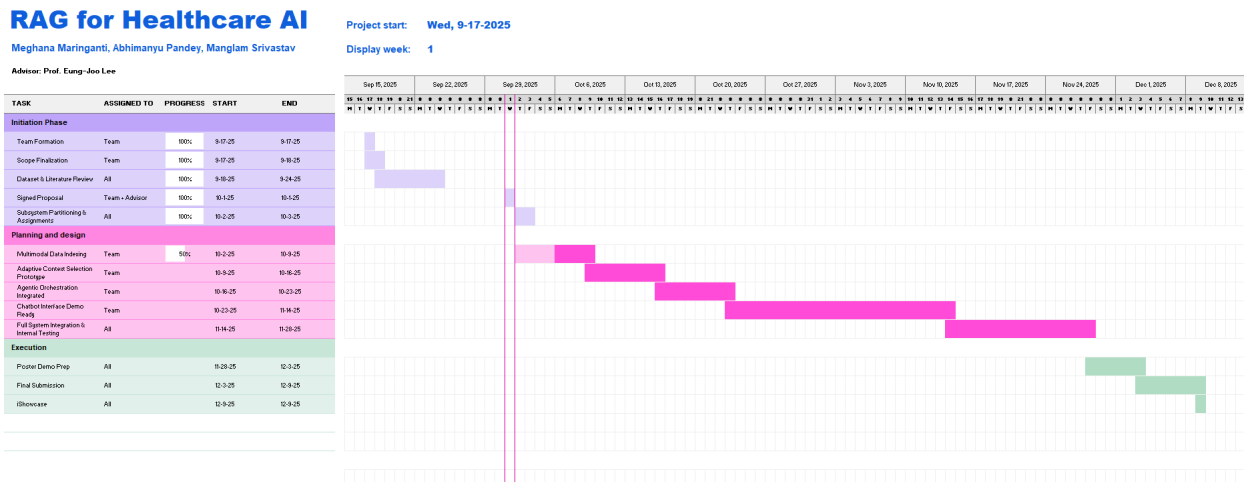
Upload size (Image)	0.5 MB	20 MB	Support typical compressed images
Query text length	50	2000	Ensures queries are concise but detailed enough for retrieval.

Table 5: Chatbot Interface Parameters

4. Project Timeline & Gantt Chart

Milestone	Date
Team Formation	09/17/2025
Scope Finalization	09/18/2025
Dataset and Literature Review	09/24/2025
Signed proposal	10/01/2025
Subsystem Partitioning	10/02/2025
Multimodal Data Indexing Complete	10/09/2025
Adaptive Context Selection Prototype	10/16/2025
Agentic Orchestration Integrated	10/23/2025
Chatbot Interface Demo	11/14/2025
Full System Integration & Internal Testing	11/28/2025
Poster Demo	12/03/2025
Final Submission	12/09/2025
iShowcase	12/10/2025

Table 6: Milestone Schedule



5. Ethics

This section was written by Meghana Maringanti.

We completed a checklist of 21 questions related to ethical concerns for the project Agentic RAG for Healthcare. The full checklist is presented in the table below.

#	Question	Generally (Y/N/M)	Data Breach (Y/N/M)	Notes
1	Could the chatbot provide unsafe or misleading medical advice if misused?	M	N	Possible if LLM hallucinates; mitigated with strong disclaimers.
2	Could users mistake chatbot outputs as professional medical diagnosis/treatment?	Y	N	High risk; must clearly label as prototype/educational only.
3	Could sensitive health-related queries or uploaded images be exposed?	M	Y	Logs/images may contain private data.
4	Does the system store personal health data?	M	Y	Depends on deployment; demo may store temporary logs. Must anonymize.
5	Could the system introduce bias (e.g., some groups better represented)?	M	N	Risk due to training data skew; monitor and document bias.
6	Could outputs include harmful stereotypes (gender, race, culture)?	M	N	Possible from pretrained LLMs
7	Could responses re-traumatize users (e.g., around serious conditions)?	Y	N	Sensitive health topics can trigger distress; use empathetic phrasing
8	Could users become over-reliant on chatbot reassurance?	M	N	Possible for anxious users; mitigate with reminders to consult clinicians.
9	Could malicious users exploit chatbot for misinformation?	M	N	Restrict knowledge base & retrieval to vetted medical literature.
10	Could uploaded medical images be misused?	N	Y	Not a risk in normal use; breach could expose scans. Secure storage needed.
11	Could attackers gain access to user	N	Y	Low general risk; but breach

	logs or interactions?			could expose sensitive logs.
12	Should there be an age restriction for chatbot use?	Y	N	Younger users may misinterpret; restrict access
13	Are users clearly informed how their data is used?	Y	Y	Transparency via consent notice required for trust.
14	Could outputs be misinterpreted and spread as medical truth?	Y	N	High risk if shared out of context; use visible disclaimers/watermarks.
15	Could logs be used for unintended profiling?	N	Y	Not in normal use; breach could allow profiling of queries.
16	Could the chatbot mishandle multimodal inputs (e.g., misinterpret an image or mismatched text-image pair)?	Y	N	Prototype risk; limit to supported formats and add error messages.
17	Could the system fail to return an answer and leave the user with incomplete or confusing guidance?	Y	N	Possible in demo; provide fallback “unable to answer” messages.
18	Could latency or system downtime prevent timely responses, leading to user frustration or risky delays?	M	N	Affects usability; mitigated by monitoring + clear “system unavailable” alerts.
19	Could users upload images or data that the system is not designed to handle (e.g., unrelated photos, non-medical documents)?	Y	N	Likely; add validation + friendly error handling.
20	Could the project’s research data, if released, be misused for commercial exploitation without safeguards?	M	Y	Possible so use licenses and clear scope restrictions in publications.
21	Could the chatbot’s reliance on external sources (papers, databases) lead to copyright/IP issues if misused?	M	N	Possible if raw text is reproduced; restrict to abstracts/public-domain sources.

This ethics review highlights the primary risks of developing and demonstrating an Agentic RAG system for healthcare, with particular emphasis on **data privacy, bias, misinterpretation, and user safety**. While several concerns exist under both normal use and potential data breaches, most risks can be mitigated through **disclaimers, transparency, careful dataset curation, and secure handling of user inputs**. Importantly, this project will be conducted as a **prototype for educational and research purposes only**,

with no use of real patient data. Future work beyond the capstone stage would require additional safeguards, including **regulatory compliance (e.g., HIPAA, FDA guidelines), extensive bias evaluation, and clinical validation** before deployment in real-world healthcare settings.

6. Approvals

The signatures of the people below indicate an understanding of the purpose and content of this document by those signing it. By signing this document, you indicate that you approve of the proposed project outlined in this Statement of Work, the division of work, the Ground Rules and that the next steps may be taken to create a Product Specification and proceed with the project.

Approver Name	Title	Signature	Date
Abhimanyu Pandey	Team Member	Abhimanyu Pandey	10-01-2025
Manglam Srivastav	Team Member	Manglam Srivastav	10-01-2025
Meghana Maringanti	Project Manager	Meghana Maringanti	10-01-2025
Dr Eungjoo Lee	Advisor	Dr Eungjoo Lee	10-01-2025
Dr Nikitha Sharma	Instructor		

Author Contribution

.

Section	Author	Word Count
1. Executive Summary	Abhimanyu Pandey	
2. User /Market Research	Manglam Srivastav	
3. Product Features	Meghana, Abhimanyu, Manglam	
4. Project Timeline & Gantt Chart	Meghana, Abhimanyu, Manglam	
5. Ethics	Meghana Maringanti	
6. Appendix	Team	

7. Appendix

A. Advisor Engagement

1) Project Team Responsibilities

- The Project Manager will set up and facilitate a weekly call/meeting with the Faculty Advisor. The Project Team will provide weekly status updates to the Faculty Advisor including upcoming deliverables, critical issues, and any adjustments to the Project Plan.
- Documents will be provided to the Faculty Advisor with adequate time for review and signature. The time necessary for review will be agreed with the Advisor. The minimum review time will be 3 days prior to the document due date.
- Design files will be provided to the Faculty Advisor as requested in a format agreed to with the Advisor.
- Support requirements will be clearly requested from the Faculty Advisor with the dates required and an adequate time for fulfilling the request.
- Modification requests to the Project Plan by Faculty Advisor will be reviewed and agreed to within 1 week of the request.

2) Faculty Advisor Responsibilities

- The Faculty Advisor will provide knowledge and expertise to help the group stretch their skills.
- The Faculty Advisor will participate in a weekly or bi-weekly call/meeting with the Project Team to review the project status, upcoming deliverables, priorities, issues, and progress to the agreed Project Plan.
- The Faculty Advisor will provide document review, feedback and approval, rejection, approval with contingencies with adequate time for the Project Team to meet the course due dates.
- The Faculty Advisor will provide feedback to requested support requirements from the Project Team. This includes feedback and guidance on design implementations decisions, design files, test plans, test procedures and test results.
- The Faculty Advisor shall provide technical advice and guidance to the Project Team answering inquiries approximately 1 hour per week.
- Modifications to the Project Plan by the Project Team will be resolved and documented within 1 week of the request.
- Grade the finalized project using a skill-based rubric
- Attend iShowcase in May.

B. Ground Rules

As a team and as individual team members, we agree to:

1. **Stay focused on our objectives and goals.**
   Each time the team meets, we will clearly define our objectives and desired outcomes at the beginning of the meeting. We will politely remind team members if we are getting off track.
2. **"Sidebar" any issues that are relevant but not consistent with the immediate objectives.**
   Occasionally, important matters are raised that are not relevant to the immediate goals of the meeting. To keep the group on track, but avoid losing the issue, create a "sidebar" where these topics can be listed and discussed later.
3. **Listen when others are speaking.**
   We will listen and consider others' input before adding our own comments.
4. **All viewpoints will have an opportunity to be heard.**
   We understand that some team members may be quieter than others. We will make an effort to get each team member's viewpoint, and that no one dominates the discussion.
5. **Differences of opinion will be discussed respectfully**
   We will identify areas of agreement before assessing areas of disagreement. We will encourage each other to look beyond our own point of view. We will discuss different ideas respectfully. As a team, we will weigh the merits of different opinions and agree on a process for choosing a direction. All team members will respect and follow the decision or direction.
6. **Look for the good points in new ideas.**
   We will endeavor to explore the value in each idea as we assess and select our path forward.
7. **Focus on the future, not the past.**
   We will use our past experience to inform our decisions, but focus the discussion on the future objectives. Blame for past performance is counterproductive, we will focus on finding solutions.
8. **Agree upon specific action items and next steps.**
   At the end of each meeting and discussion, we will summarize and agree on specific next steps, action items, and assignments.
9. **Accountability**
   As team members, we will each be responsible for our individual assignments and contributions to achieving the team's objectives and goals. We will honor our responsibilities and not let our team members down.