SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

MRad-RAG: A Local Multimodal Radiology RAG Framework for Consumer Level Hardware

Liliana Ma, MD, PhD, Stanford Healthcare; Sergios Gatidis, MD; Serena Yeung, PhD;
Shreyas Vasanawala, MD, PhD

Introduction/Background

While large language and vision-language models (LLMs/VLMs) show strong potential in healthcare applications, adoption is hindered by the issue of “hallucinations”. Retrieval-Augmented Generation (RAG) addresses this limitation by integrating high-quality, relevant external data into the model’s response generation. While text-based RAG has shown early success, its extension into the multimodal domain, remains underexplored, despite generative AI's strong potential as an interpretive radiology assistant. Meanwhile, technology companies have released increasingly powerful and efficient open-source models capable of running within HIPAA-compliant environments. We introduce Multimodal Radiology RAG (MRad-RAG), a framework that combines state-of-the-art multimodal foundation models with retrieval-based augmentation. MRad-RAG runs on consumer-level hardware and enables context-aware generation using both image and text.

Methods/Intervention

Content from https://radiologyassistant.nl/, which primarily features radiology image-caption pairs, was embedded into a Chroma vector database using BiomedCLIP, a multimodal embedding model. Google’s Gemma 3 models (4B, 12B parameters), released in March 2025, were selected as the foundational models, for their ability to process multiple text and image inputs. Performance was evaluated using the 2022 Diagnostic Radiology In-Training Exam, comprising multiple-choice questions across 11 subtopics; 41 text-only and 65 image-based. For image-based queries, both images and associated text were embedded, and the most relevant image-caption pairs were retrieved and passed with the query. Implementation and inference were conducted on an Alienware Aurora R16 desktop (NVIDIA GeForce RTX 4090 GPU and a 24-core Intel i9 processor).

Results/Outcome

With Gemma-4B, MRad-RAG improved text-only question accuracy by 12.2% (26.8% vs. 39.0%) but reduced performance on image-based questions by 9.3% (30.8% vs. 21.5%). Conversely, Gemma-12B improved performance on image-based questions (32.3% vs. 29.2%) but slightly reduced accuracy on text-only items (46.3% vs. 51.2%). Performance varied by subtopic.

Conclusion

This pilot study shows the feasibility of a multimodal RAG pipeline on consumer hardware and benchmarks Gemma 3 performance. While results are mixed, MRad-RAG is a customizable, local solution that can be iteratively refined. Future work will expand the reference database, test alternative models, and improve retrieval strategies.

Statement of Impact

This work lays the foundation for a local pipeline that can eventually be scaled and extended to copilot tasks

such as automated reporting and identification of actionable findings.

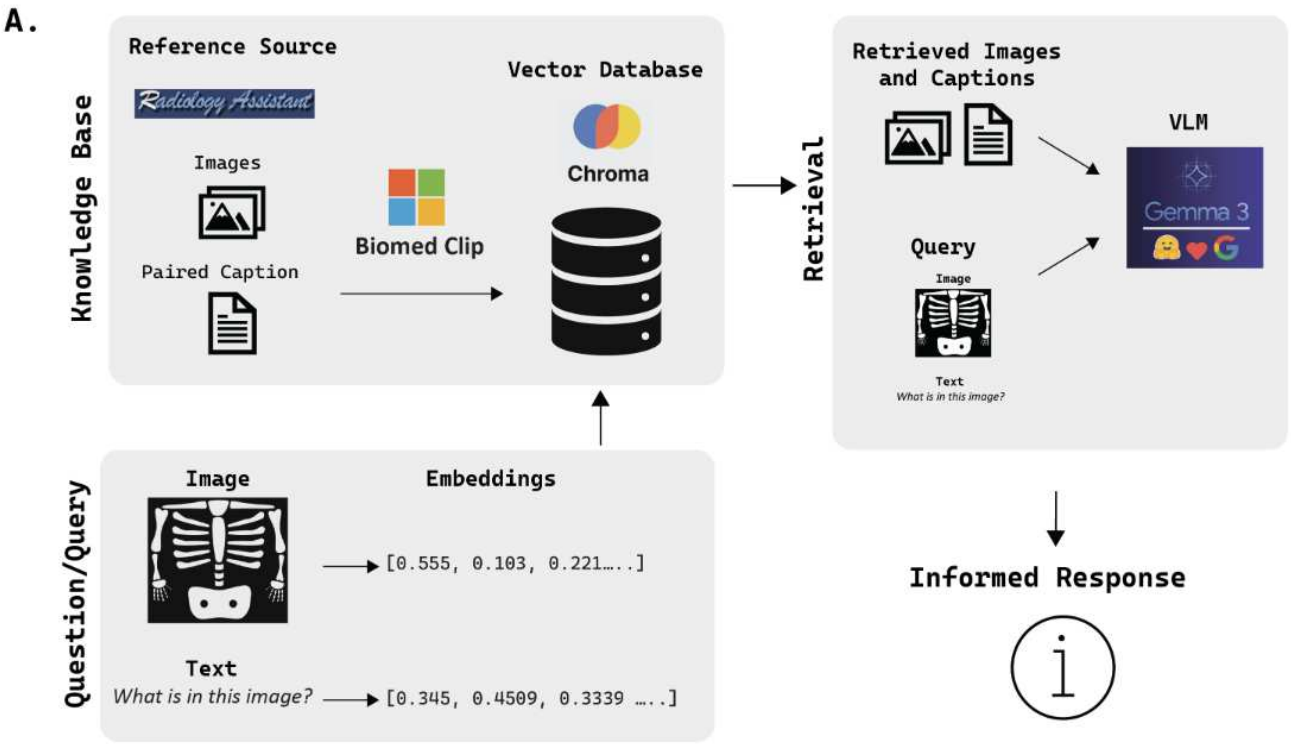

Figure 1: The prototype MRad-RAG pipeline. The Radiology Assistant website was scraped and used as the reference database, as almost all of the data is presented as image-caption pairs. The reference data was then embedded into a vector database. Queries are summarized using the same VLM into summary search terms, embedded and then used to retrieve relevant context (image ad text) from the reference database, which was then fed into the vision language model with a single-shot example. Inference time with the LLM alone was 7.2 s±1.1 s. Inference time with the RAG pipeline was 6.9s ±1.3s.

A.

Example Question:

Which of the following is an EARLY complication of repair of esophageal atresia with distal fistula?

A. Gastroesophageal reflux.
B. Recurrent tracheo-esophageal fistula.
C. Disordered esophageal peristalsis
D. Diverticulum formation.

Correct Answer: B

B.

Summarized terms for retrieval:

Esophageal atresia; distal fistula; tracheo-esophageal fistula, peristalsis, diverticulum formation

↓

Retrieved Context:

Image

Caption

First look at the image and describe the findings.
Then continue reading.
The findings are:
Feeding tube cannot be passed and lies in a dilated proximal esophagus
Normal air in the abdomen.
Diagnosis: esophagus atresia with a distal tracheo-esophageal fistula
Esophagus atresia is an anomaly which arises in the fourth week of the embryogenesis, at a stadium in which the trachea and esophagus should separate from each other.
In case of failure of complete separation esopaghus atresia can occur.
Clinically the neonate cannot swallow saliva, may blow bubbles and will aspirate on feeding.
When a feeding tube is inserted it cannot be passed distally.
A radiograph with a curled up feeding tube will confirm the diagnosis.

C.

MRad-RAG Response	**Pure LLM Response**
B. Recurrent tracheo-esophageal fistula.	C. Disordered esophageal peristalsis.

Figure 2: A, example in-service examination question. B, question is summarized with a single-shot summary prompt, and the summary terms are then used to retrieve relevant image/caption pairs. These are fed in as context. C, in this example, the MRad-RAG selected the correct answer. The explanation is not included here for brevity, but the VLM reasoning included reference to the provided context.

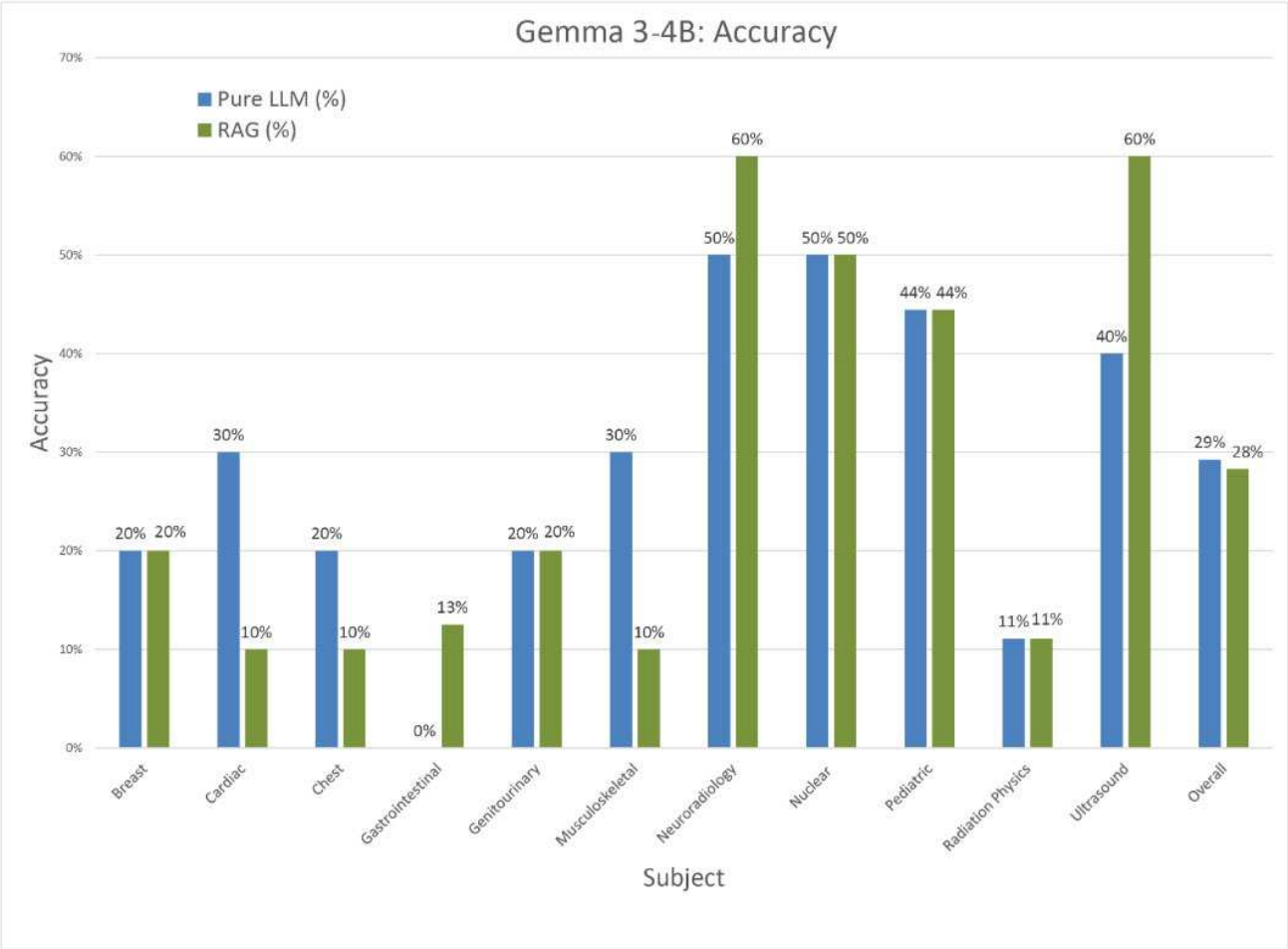

Figure 3: Demonstrates the effects of the VLM alone (Pure VLM) and MRad-RAG (RAG) across different topics. Performance varied widely across topics, and will need to be further tuned and evaluated more extensively on larger datasets.

Keywords

vision language models; large language models; retrieval augmented generation; radiology; foundation models