MIRAGE

Medical Imaging Retrieval Augmented Generation Environment

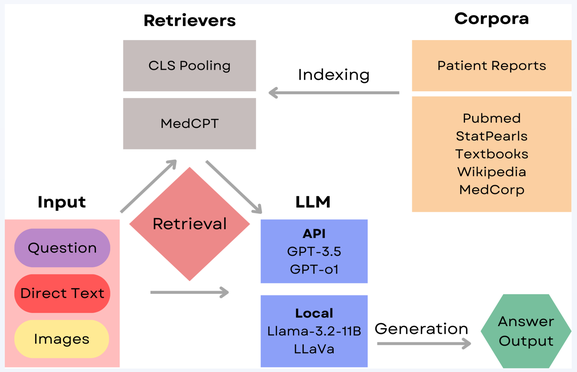
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Ahmed Alqalam, Andrew Cassarino, Nicolo Krueger

Introduction

MIRAGE is an AI system designed to assist radiologists by providing instant, evidence-based responses to complex clinical questions. The system leverages Retrieval Augmented Generation (RAG) to pull relevant data from a large corpus (database) and generate context-specific insights, improving diagnostic efficiency in X-ray interpretation.

Figure 1 – RAG System Architecture [2]



Objectives

1. Develop an intelligent assistant that can respond to radiology-related questions.
2. Improve diagnostic decision-making and efficiency for radiologists at the Children's National Hospital

Collaboration

Partnered with Dr. Syed Anwar at Children's National Hospital.

Children's National

Department of Biomedical Engineering
School of Engineering & Applied Science

Materials and Methods

Technology Stack:

- **Retrieval-Augmented Generation (RAG)**
- Large Language Model (LLM) via GPT API as well as Llama
- Corpus of ~300 MIMIC-CXR patient reports, capped at 50,000 tokens

Corpus Setup:

- MIMIC-CXR dataset formatted and structured using custom scripts
- Each patient entry includes impression, findings, and associated metadata

Input:

The API can receive text plainly as well as images in 2 specific ways.

- URL to an uploaded image
- 64-bit encoded image

64-bit Encoder

{'type': 'image_url', 'image_url': {'url': ' …

When encoded, this image went from 1.36MB to 1,906,780 base-64 characters long.

Testing Strategy:

CORPUS

INPUT

study_id	subject_id	ref_id	question_typ	question	answer
58065422	10018423	51545426	presence	is there atelectasis in the lung bases?	yes
58815716	10018423	50526690	presence	is there consolidation?	no
50526690	10018423	51545426	presence	is there pneumothorax?	no
51545426	10018423	58065422	abnormalit	is there evidence of any abnormalities in this image?	yes

QUESTIONS

There are 7 question types derived from the MIMIC Excel database to ensure a focused and controlled evaluation of MIRAGE's performance: abnormality, presence, location, level, type, view, and difference.

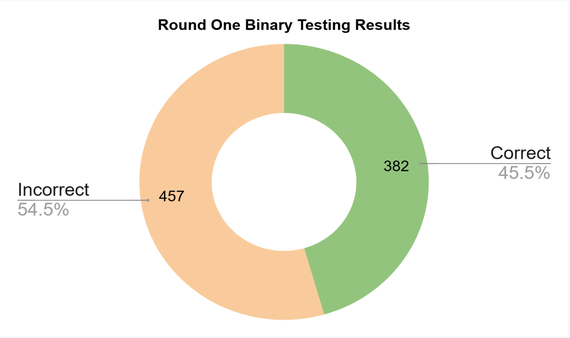
Evaluation Method:

- We ran MIRAGE on these question types for each patient report
- We manually reviewed model outputs and logged accuracy across each category.

Results

First Round of Testing:

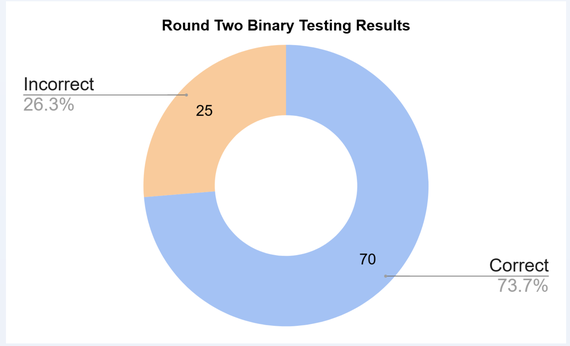
We tested MIRAGE using **three binary question types** derived from the MIMIC-CXR Excel database: abnormality, presence, and view.



Accuracy dropped due to mismatches caused by the "View" question type, as MIRAGE analyzed only the first X-ray for patients with multiple images. Since it couldn't associate views correctly, this question type will be removed from future tests.

Second Round of Testing:

We tested MIRAGE using **two binary question types** derived from the MIMIC-CXR Excel database: abnormality and presence in tandem with an upgraded LLM from GPT-3.5 to GPT-o1.



With a 73.7% accuracy, the second round of testing shows that MIRAGE performs well enough for further evaluation. This result aligns with other early-stage clinical NLP systems, which typically achieve 70–75% accuracy [4].

Changes and Progress

1. Future testing will involve sentence-level eval.
2. Implement Llama 3.2 11B model for local deployment at the hospital.
3. Evaluating MIRAGE using real clinical cases.

Conclusion

The MIRAGE system demonstrates strong foundational potential for assisting radiologists. By integrating a multimodal framework, MIRAGE aims to reduce diagnostic turnaround times, alleviate radiologist burnout and provide immediate support for physicians. Our testing, although limited in scope, achieved an accuracy of 73.68% using binary classification. These results highlight the promise of our system.

Looking ahead, our team is planning on deploying Llama-3.2-11B for localized hospital use and starting evaluation on real clinical cases. This will bring MIRAGE closer to real-world implementation. With planned improvements, we believe MIRAGE can evolve into a clinically viable assistant that enhances the efficiency and quality of radiological care.

References

[1] Fawzy NA, Tahir MJ, Saeed A, Ghosheh MJ, Alsheikh T, Ahmed A, Lee KY, Yousaf Z. Incidence and factors associated with burnout In radiologists: A systematic review. Eur J Radiol Open. 2023 Oct 23;11:100530. doi: 10.1016/j.ejro.2023.100530. PMID: 37920681; PMCID: PMC10618688.

[3] Ian A. Weissman, Peter Van Geertruyden, Anand M. Prabhakar, David Fessell, Mark Alson, Frank J. Lexa, Practice Resources to Address Radiologist Burnout,Journal of the American College of Radiology,Volume 20, Issue 5,2023, Pages 494-499, ISSN 1546-1440, https://doi.org/10.1016/j.jacr.2023.03.007.

[2] Guangzhi X, Qiao J, Zhiyong L, Aidong Z. Benchmarking Retrieval-Augmented Generation for Medicine. 2024. https://arxiv.org/abs/2402.13178

[4] Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ Health Care Inform. 2021 Mar;28(1):e100262. doi: 10.1136/bmjhci-2020-100262. PMID: 33653690; PMCID: PMC7929894.

GitHub

Acknowledgments

We thank Dr. HyungSok Choe, Dr. Syed Anwar, Zhenhoa Zhoa, Nishad Kulkarni, and Faisal Al Munajjed for their continued support and collaboration.