Development and Evaluation of a Locally-Hosted Large Language Model with Retrieval-Augmented Generation for CT Decision Support

Gabriel K Lau, BA; Tien Comlekoglu, PhD; Ayush Doshi, MD; Arun Krishnaraj MD, MPH, MBA

Department of Radiology and Medical Imaging
University of Virginia School of Medicine, Charlottesville, VA, USA

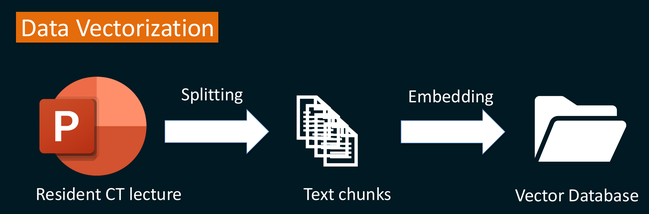
Background

- Choosing the right imaging protocol in a timely manner is a crucial part of patient care and involves collaboration between radiologists and referring providers
- To address potential delays in patient care, we developed and evaluated a locally-hosted Large Language Model (LLM) utilizing Retrieval-Augmented Generation (RAG) as a clinician support tool.
- Our aim was to enhance workflow efficiency, comply with internal documentation, and offer decision support to referring providers.

Methods

- We utilized a lecture presentation to non-radiology residents covering relevant CT protocols for various indications to serve as a knowledge base for an LLM and for implementing a RAG pipeline.
- We evaluated six open-source LLMs with parameter counts ranging from 3-billion to 8-billion parameters from the Llama (Meta), Qwen, Microsoft Phi, and Mistral model families.
- LLMs were evaluated using a standard set of 26 questions, and outputs were scored for semantic similarity to the expected answer using the f1 BERTscore.
- All outputs were evaluated for accuracy by radiologists.

RAG Framework



What is RAG?

- Combines information retrieval + LLM generation
- Reduces hallucinations and improves domain-specific accuracy

How Our RAG Pipeline Works

- Step 1: Query Input - User submits a clinical question
- Step 2: Document Retrieval - RAG searches lecture materials and identifies the most relevant content "chunks"
- Step 3: Context Augmentation - Retrieved text is added to the LLM prompt and provides grounded, institution-specific guidance
- Step 4: LLM Generation - Locally hosted LLM generates a response using: retrieved context and system prompt instructions
- Step 5: Output - Returns targeted imaging recommendation aligned with internal guidelines

Models Tested in Development of RAG Pipeline

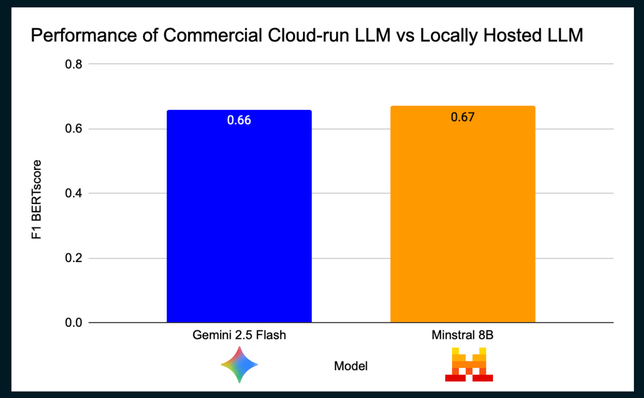
Locally run models

Qwen · Mistral · Llama · Phi

Cloud commercial model

Gemini

Results



- The highest performing local LLM was the Mistral 8B with a BERTscore of 0.67 (Mistral 8B).
- The Mistral model matched the commercial Gemini 2.5 Flash (f1 score 0.66) given an equivalent system prompt, RAG pipeline, and questions.
- The Mistral model answered 21/26 questions correctly.

Conclusions

- An open-source locally-hosted LLM utilizing a RAG framework has potential to be used to provide decision support for imaging orders.
- This approach represents a promising and secure solution to augment clinical workflows and assist in timely and effective delivery of patient care.
- Future directions could incorporate RAG in ACR appropriateness criteria and institutional imaging policies.