LLM Grounding: Improving Reliability in Clinical AI

Krish Shah
Computer Engineering – 2025
Supervisor: Dr. Vijaya B. Kolachalama
Team Member: Yi Liu

April 21, 2025

Boston University

Recorded Presentation

Project Overview

- **Problem:** LLMs hallucinate facts — unacceptable in clinical decision support.
- **Goal:** Use biomedical retrieval to ground responses in trustworthy sources.
- **Outcome:** Significant improvements in citation accuracy and factual reliability.

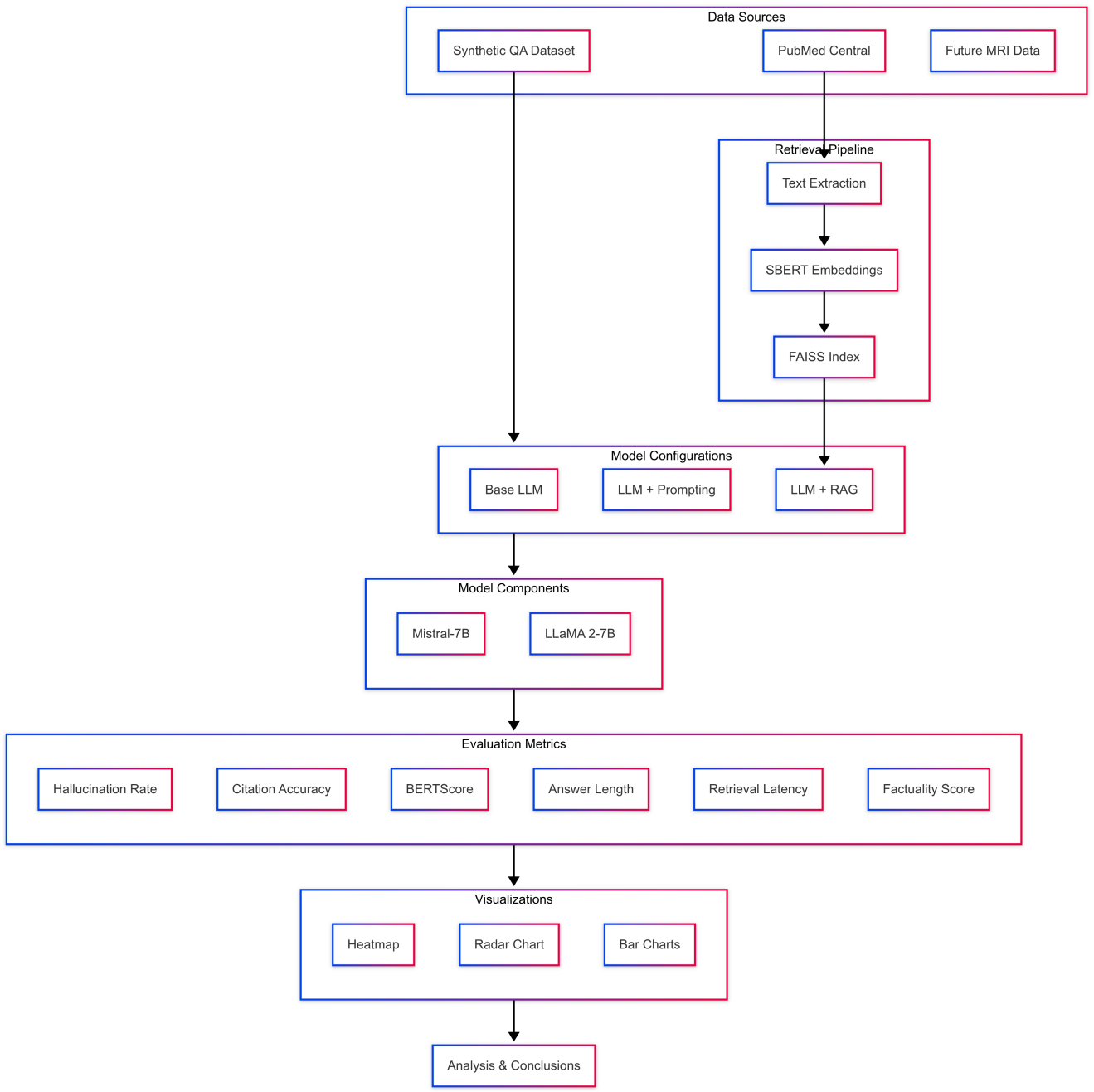
Clinical Motivation

- In high-stakes domains, output hallucination = clinical risk
- AI tools must be interpretable, transparent, and evidence-backed
- Grounded LLMs can support — not replace — medical practitioners

Project Goals

- Design RAG system for medical question answering
- Integrate embedding-based retrieval (SBERT + FAISS)
- Evaluate hallucination, citation quality, semantic similarity, latency
- Benchmark across Base, Prompting, and RAG configurations

System Architecture



Methods and Toolchain

ML Models and Data:

- LLaMA 2 / Mistral-7B (via Ollama)
- PubMedQA, BraTS dataset
- SBERT (sentence-transformers)

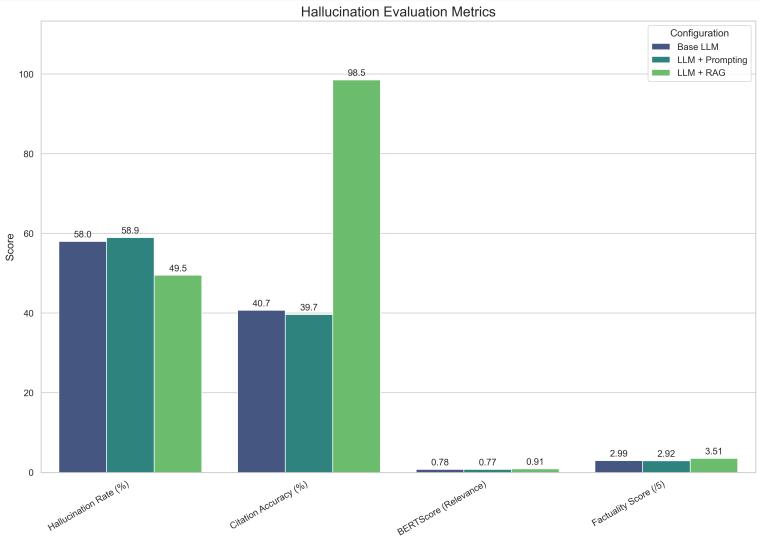
Engineering Stack:

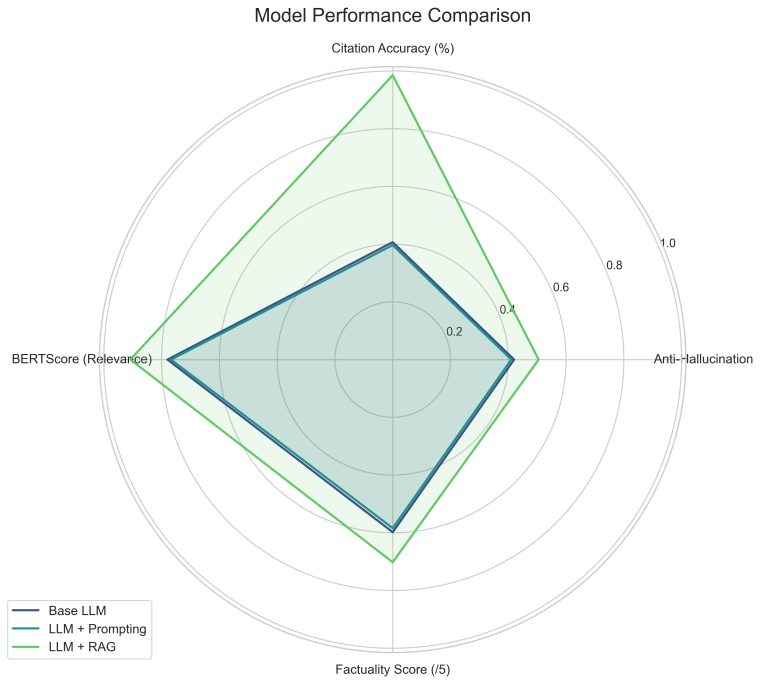
- Python, PyTorch
- Hugging Face, LangChain, FAISS
- PDFMiner, spaCy, Jupyter

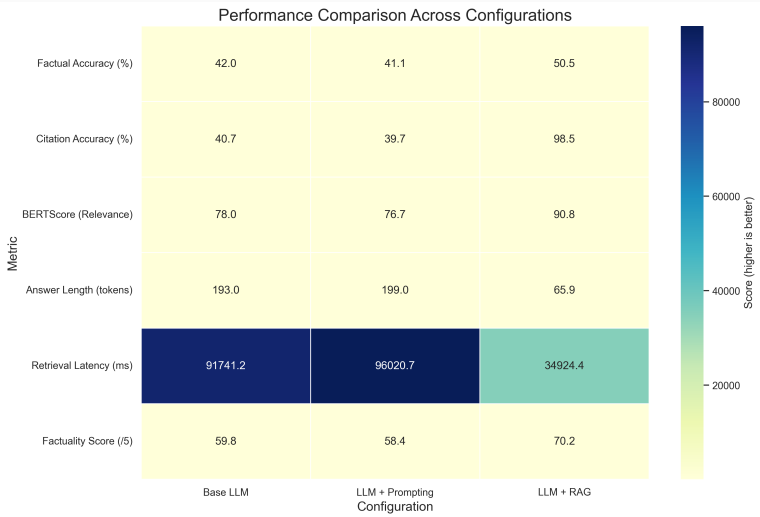
Quantitative Results

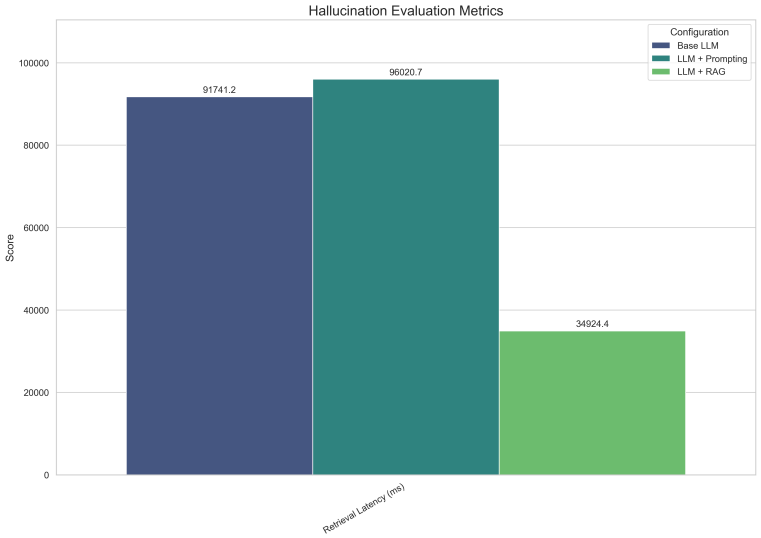
Metric	Base	Prompting	RAG
Hallucination Rate (%)	57.97	58.94	49.49
Citation Accuracy (%)	40.69	39.66	98.55
BERTScore (Relevance)	0.78	0.77	0.91
Answer Length (tokens)	193	199	66
Retrieval Latency (ms)	91741	96020	34924
Factuality Score (/5)	2.99	2.92	3.51

Visualization Highlights









Technical Reflection

Successes

- RAG improved citation accuracy by 58%
- Factuality gains confirmed through human scoring
- Robust evaluation pipeline created

Limitations

- RAG latency 35s/query
- Prompting alone was ineffective
- No multimodal integration yet

Conclusion and Future Work

- **Impact:** Grounding improves factual trust in LLMs
- **Next:** Multimodal integration (MRI/labs), retrieval optimization
- **Vision:** Safe, transparent AI support tools in medicine