Machine Learning Concentration Experiential Component Report

Grounding Large Language Models for Reliable Clinical AI

Student: Krish Shah
Major: Computer Engineering – Machine Learning Concentration

Supervisor: Dr. Vijaya B. Kolachalama
Affiliation: Boston University School of Medicine
Email: vkola@bu.edu

Team Member: Yi Liu
Role: PhD Student, Computer Science
Email: yliu22@bu.edu

Date: April 21, 2025

1. Overview

Large language models (LLMs) like GPT and LLaMA have demonstrated strong performance across general-purpose tasks, but their application in critical domains like healthcare remains limited by their tendency to hallucinate—i.e., generate plausible but incorrect or unverifiable information. This project explores how hallucinations can be reduced through a technique known as grounding, where the LLM output is explicitly tied to trustworthy sources such as biomedical literature and structured patient data.

Working at the intersection of natural language processing and biomedical informatics, our goal was to build a Retrieval-Augmented Generation (RAG) system to improve factual accuracy in LLMs used for healthcare question answering. We integrated biomedical information retrieval, prompt engineering, and domain-specific evaluation techniques. This work contributes to the design of trustworthy clinical decision support tools powered by AI.

2. Contributions

- Designed and implemented a RAG pipeline using SBERT embeddings and FAISS-based retrieval.
- Developed domain-specific prompt templates to adapt open-source LLMs without full fine-tuning.
- Created automated and manual evaluation metrics including hallucination rate, citation accuracy, and BERTScore.

- Contributed to the design of a multimodal encoder architecture to enable future integration of MRI and lab data.

The core system was implemented and tested, and early results are promising. The multimodal module is still under development and will be integrated in a future research phase.

3. Methods of Study Relevant to Machine Learning

Data:

- Biomedical Literature: Extracted from PubMed Central using PDFMiner + spaCy.
- Synthetic Clinical QA Pairs: Created based on medical guidelines and diagnostic examples.
- Imaging (in progress): BraTS MRI datasets will be incorporated for multimodal integration.

ML Models:

- LLMs: Mistral-7B and LLaMA 2-7B (via Ollama, no fine-tuning).
- Retrieval: Sentence-BERT embeddings + FAISS dense similarity search.
- Prompting: Instructional and few-shot prompt engineering templates were tested across configurations.
- Evaluation: BERTScore, BLEU, and citation/hallucination annotations.

Tools: Python, PyTorch, HuggingFace Transformers, LangChain, sentence-transformers, FAISS, Ollama, Jupyter, and cloud + local GPU infrastructure (NVIDIA V100).

System Pipeline:

To facilitate reliable generation, we implemented a retrieval-augmented generation (RAG) pipeline that integrates dense vector search with language model conditioning. The system embeds text chunks from biomedical literature using SBERT and indexes them using FAISS. When presented with a question, the system retrieves top-k relevant documents based on cosine similarity. These documents are injected into a carefully templated prompt, which is passed to a frozen LLM such as LLaMA or Mistral to generate a grounded response. All configurations — Base, Prompting, and RAG — run through this pipeline with different levels of retrieval involvement.

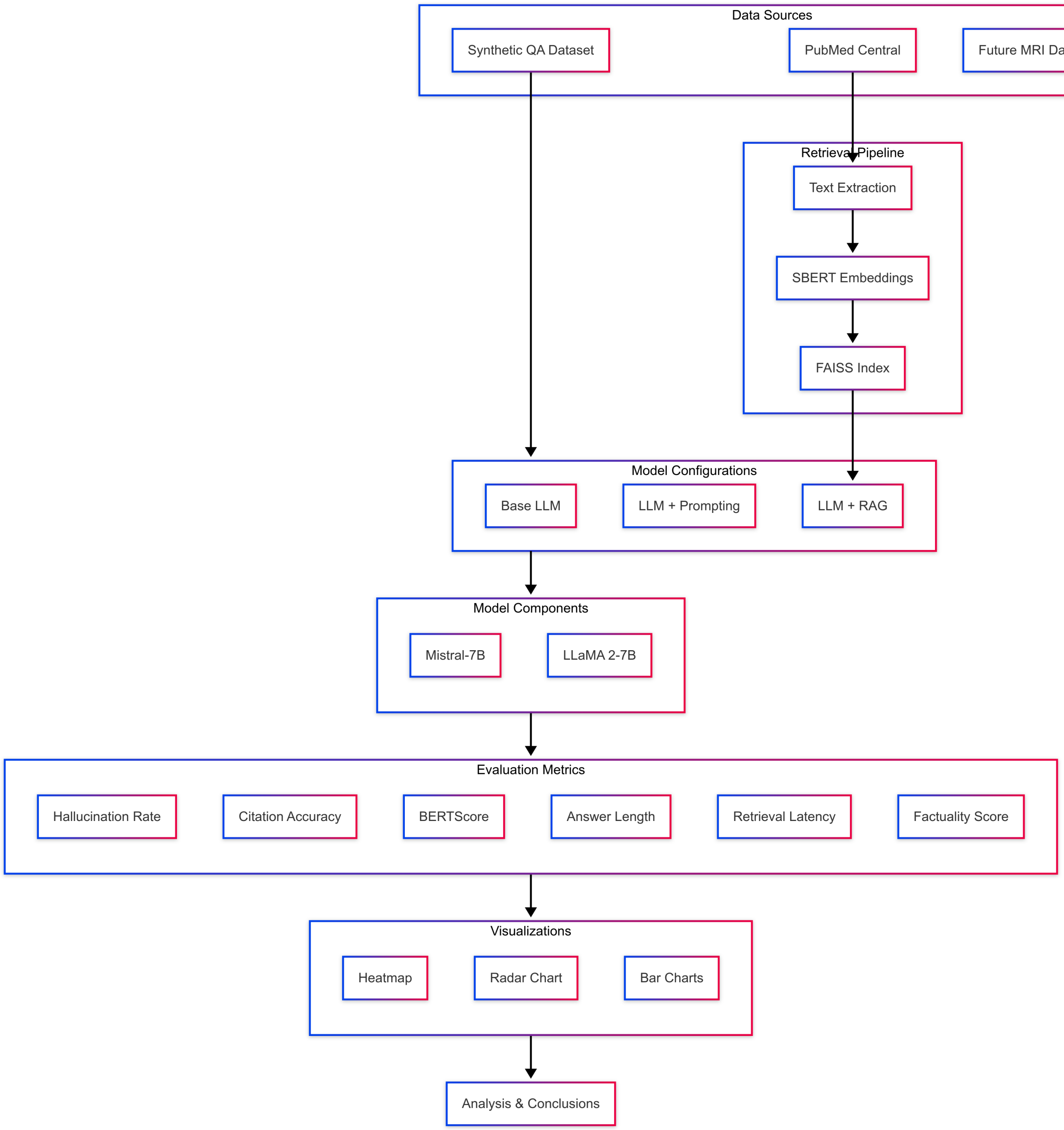

Figure 1: System architecture of the Retrieval-Augmented Generation (RAG) pipeline. Biomedical QA inputs are embedded, indexed, retrieved, and passed into an LLM as context. This ensures outputs are grounded in verifiable evidence.

4. Evaluation of Contributions

We tested three configurations: Base LLM, LLM + Prompting, and LLM + RAG.

Table 1: Real-World Evaluation Metrics Across Configurations

Metric	**Base LLM**	**LLM + Prompting**	**LLM + RAG**
Hallucination Rate (%)	57.97	58.94	49.49
Citation Accuracy (%)	40.69	39.66	98.55
BERTScore (Relevance)	0.7798	0.7674	0.9082
Answer Length (tokens)	193.0	199.0	65.9
Retrieval Latency (ms/query)	91741.2	96020.7	34924.4
Factuality Score (/5)	2.99	2.92	3.51

Results Analysis and Discussion:

Our results show that integrating Retrieval-Augmented Generation (RAG) into the LLM workflow led to a significant improvement in both factuality and grounding. The RAG-enhanced model achieved the highest citation accuracy (98.55%) and BERTScore relevance (0.91), while also receiving the best human-assigned factuality score (3.51/5). This supports our hypothesis that anchoring language models in biomedical evidence sources can drastically reduce hallucinations and improve trust in clinical response generation.

Interestingly, the Base LLM and Prompting-only configurations produced similarly long answers (193–199 tokens) but showed little difference in performance. Their hallucination rates were nearly identical (57.97% and 58.94%, respectively), and both struggled to cite or support responses with verifiable sources. This suggests that, while prompt engineering can steer tone and structure, it alone is not sufficient for factual grounding in specialized domains like medicine.

However, RAG did introduce a tradeoff — retrieval latency increased by over 30 seconds per query (34,924ms), which would not be acceptable in real-time clinical systems. Additionally, the shorter average answer length (65.9 tokens) from the RAG system may have affected the level of detail provided, which could be improved with better prompt-template tuning or chunk selection strategies.

Conclusion: Our approach was successful in demonstrating that grounding LLMs using biomedical retrieval pipelines can substantially enhance factuality and citation accuracy. While the current system is not yet production-ready due to performance overhead and limited multimodal capabilities, it provides a strong proof-of-concept foundation.

Future Improvements:

- Optimize retrieval time through document caching and index compression.
- Experiment with hybrid reranking models to improve context selection.
- Integrate multimodal inputs (e.g., lab results, imaging reports) for more context-aware generation.
- Explore instruction-tuned or lightweight fine-tuned models on biomedical QA datasets to reduce hallucinations without retrieval.

5. Visual Evaluation

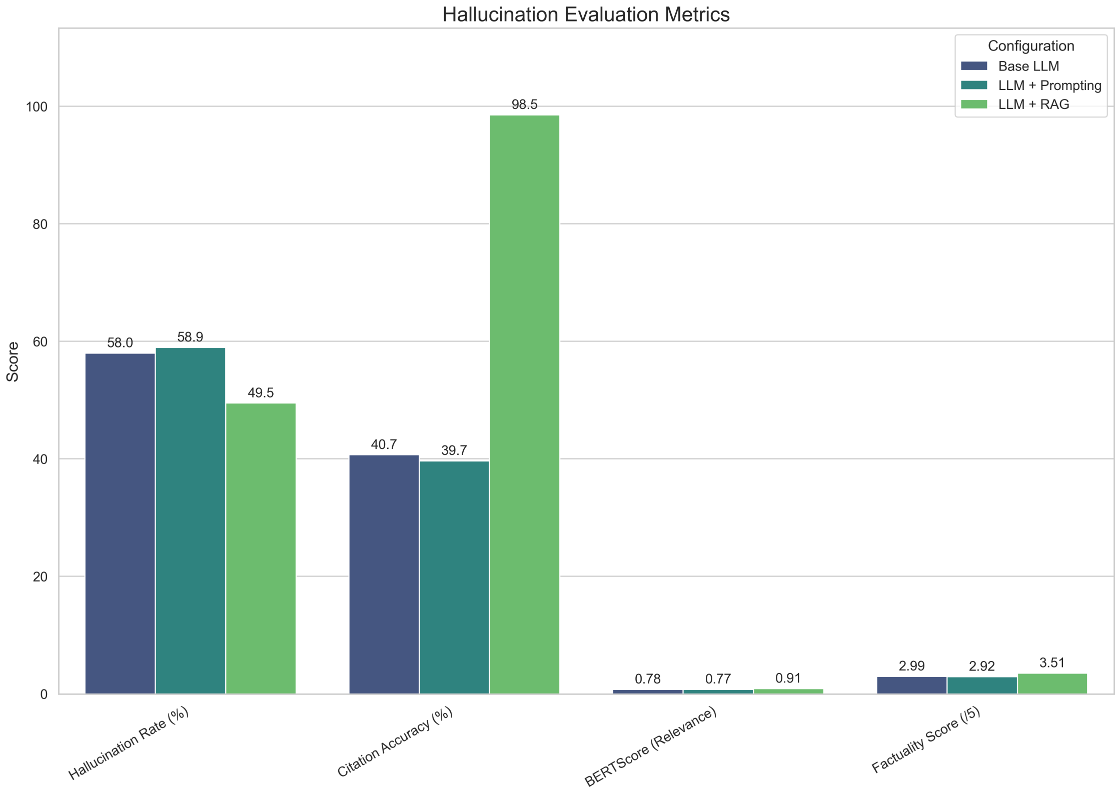

Figure 2: Expanded bar chart comparing all metrics across configurations.

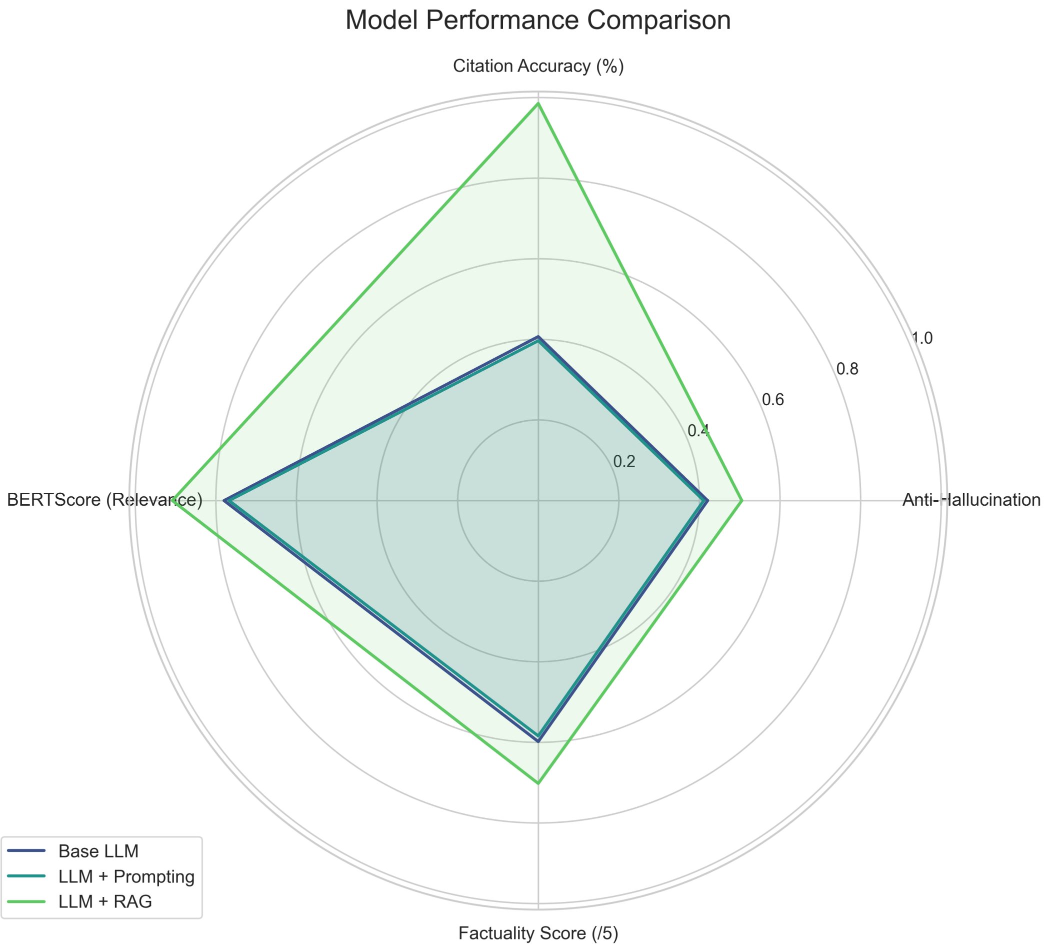

Figure 3: Radar chart showing tradeoffs across key metrics. RAG outperforms on factuality and citation accuracy.

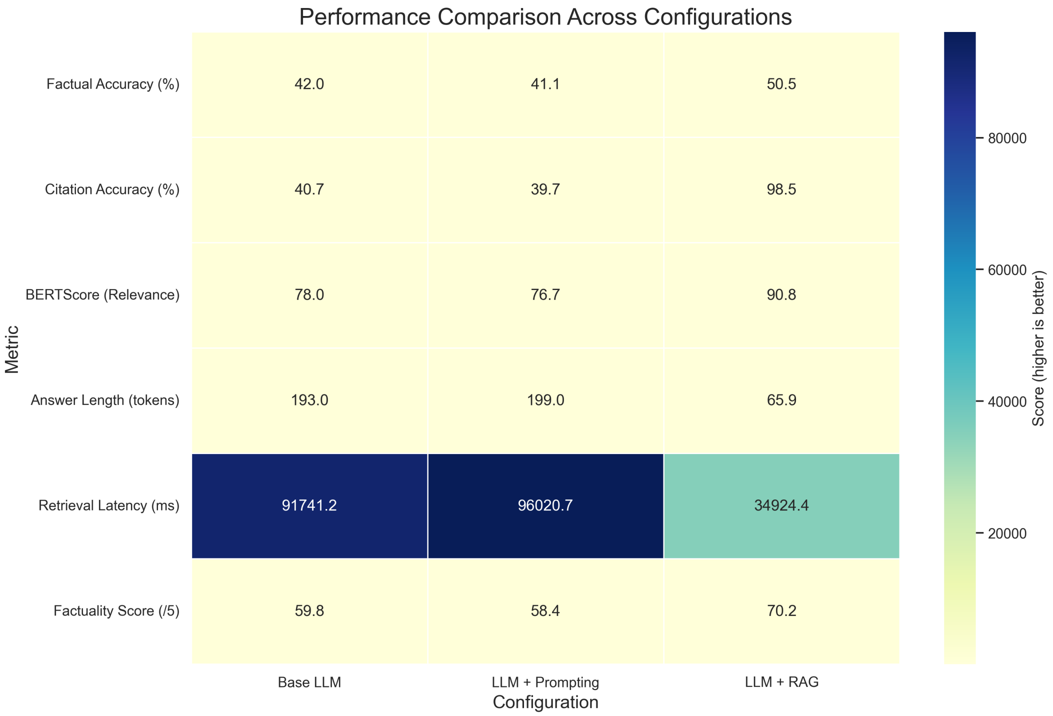

Figure 4: Heatmap of metric values, highlighting optimal configuration per dimension.

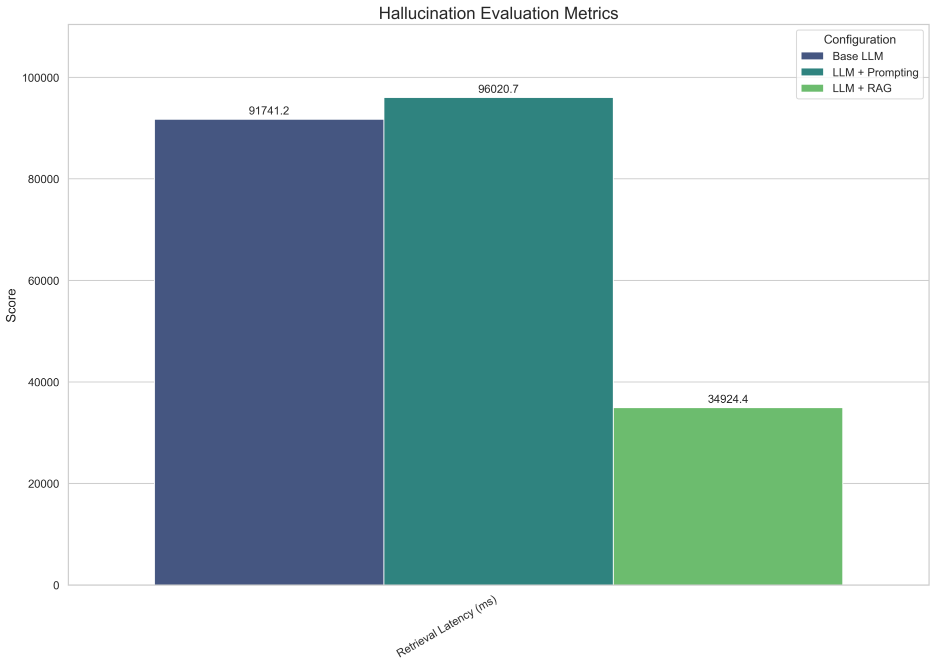

Figure 5: Retrieval latency comparison across configurations. Prompting adds little cost, but RAG introduces notable overhead.

6. References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*. https://proceedings.neurips.cc/...

2. Xue, C., Kowshik, S. S., et al. (2024). AI-based differential diagnosis of dementia etiologies on multimodal data. *Nature Medicine*, 30, 1234–1245. https://www.nature.com/...

3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, pp. 4171–4186. https://aclanthology.org/N19-1423/

4. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of EMNLP 2019*, pp. 3982–3992. https://aclanthology.org/D19-1410/

7. Group Contributions

Krish Shah (Lead):

- Built the RAG system and SBERT-FAISS retriever
- Engineered prompts and created benchmark datasets
- Designed evaluation framework and visualizations

Yi Liu (Support):

- Helped tune retrieval pipeline and experiment structure
- Collaborated on metric design and analysis interpretation
- Led multimodal integration planning for next research phase