Hallucination Mitigation Strategies in Medical Question Answering (MQA)

Data Science Institute
COLUMBIA UNIVERSITY

Team: Prajna Girish, Rithika Shankar, Kanisha Shah, Tushar Bura
Industry Mentors: Elia Lima-Walton, Harsh Sindhwa

ELSEVIER

Background

Hallucinations in medical question answering systems pose significant risks to patient safety and healthcare decision-making. Our project addresses this critical issue by developing robust mitigation strategies for LLMs in medical QA applications. We propose novel approaches to detect and reduce hallucinations, ensuring more reliable and trustworthy AI-assisted medical information retrieval.

Understanding the data

The dataset contains information from several sources, most notably ADAM. Focus analysis further revealed a vast array of unique medical topics and frequency of question types. These insights helped in assessing the dataset's diversity in features which may impact the model type.

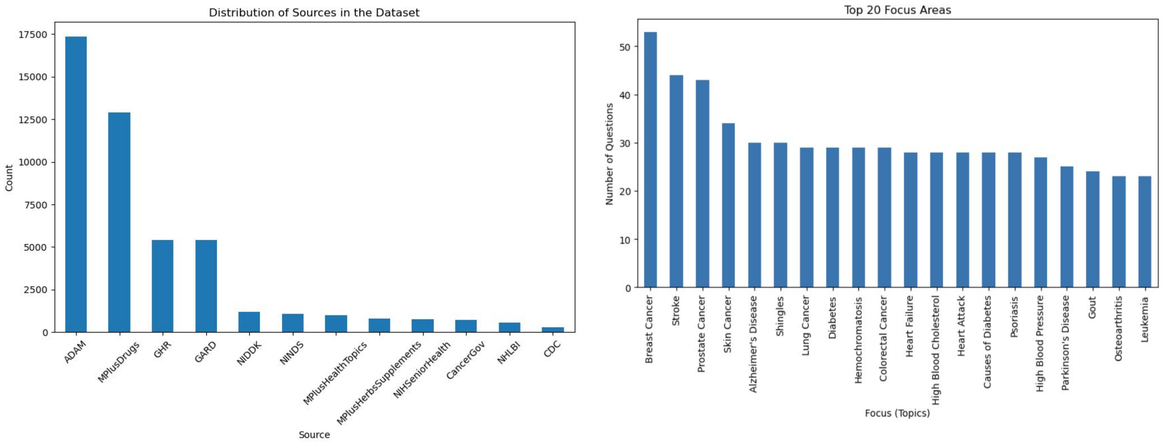

Figure 1. Exploratory Data Analysis

Architecture

Our system employs a multi-stage approach:

- **Medical knowledge retrieval - RAG**
- **LLM-based answer generation**
- **Hallucination detection module**
- **Answer refinement and verification**

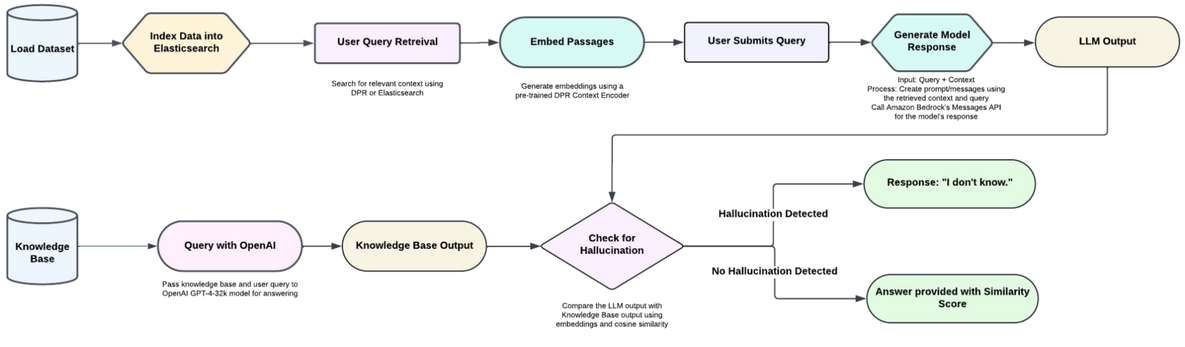

Figure 2. Model Architecture

Findings / Conclusion

The RAG architecture in our system combines a retrieval layer with a generation layer.

- **In the retrieval layer, we use Elasticsearch and FAISS for similarity-based retrieval of relevant context. Retrieved context is embedded using a Dense Passage Retrieval (DPR) encoder and combined with MedQUAD examples to construct a structured prompt.**
- **We use Claude to generate responses based on the query and retrieved context. To ensure accuracy, we validate the LLM output using cosine similarity, comparing it with the retrieved knowledge base output.**
- **This architecture ensures that responses are grounded in reliable knowledge while minimizing hallucination.**

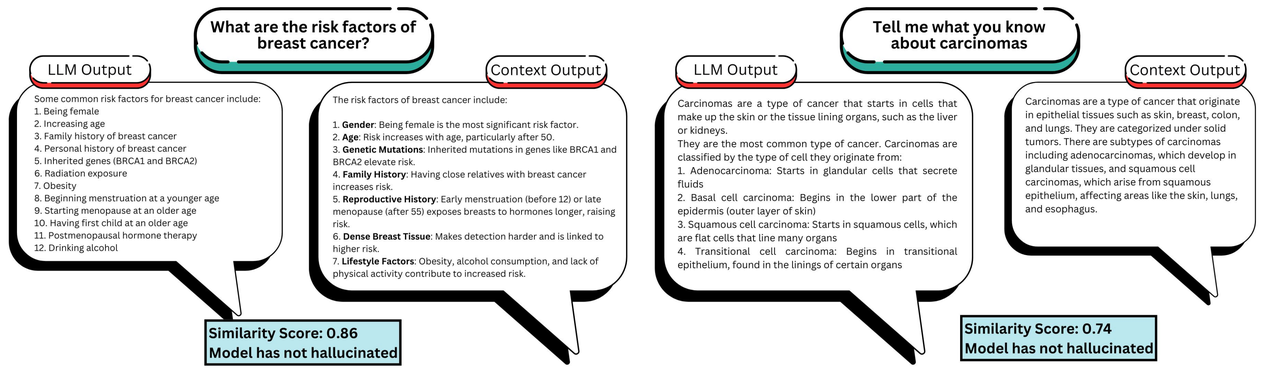

Figure 3. Comparing LLM output to Context with similarity score

Future Improvements

- **Fine-tune the Bedrock or OpenAI model using several domain-specific datasets to improve accuracy and adaptability to medical queries, enhancing the generation layer's relevance and precision.**
- **Strengthen hallucination detection by integrating multiple similarity metrics and external validation against trusted medical knowledge bases like PubMed or UMLS.**

Acknowledgments

We thank our mentors from Elsevier for their guidance and support throughout this project. We also express our gratitude to the Data Science Institute at Columbia University for providing the resources and academic environment that made this project possible.

References

1. **MedExQA: Medical Question Answering Benchmark with Multiple Explanations Yunsoo Kim, Jinge Wu, Yusuf Abdulle, Honghan Wu**
2. **DelucionQA: Detecting Hallucinations in Domain-specific Question Answering; Mobashir Sadat**
3. **https://github.com/KhalilMrini/Medical-Question-Answering**
4. **https://github.com/abachaa/MedQuAD**
5. **A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions**