Hybrid Memory-Retrieval Chatbot: Enhancing Trust in Medical Chatbot

Rochester Institute of Technology

Sagarika Singh
ss3028@rit.edu
MS AI

Github Repo

LinkedIn

INTRODUCTION

- Medical chatbots help with symptom checking and education but often lack accuracy and context-awareness.
- This project enhances chatbot reliability using Phi-2 with improved memory (ChromaDB) and retrieval systems (BM25 + MedCPT + RRF)

MOTIVATION

- LLMs often hallucinate and lose context, which can mislead users in medical chats.
- We combine memory context, hybrid RAG, and Phi-2 to reduce these risks.

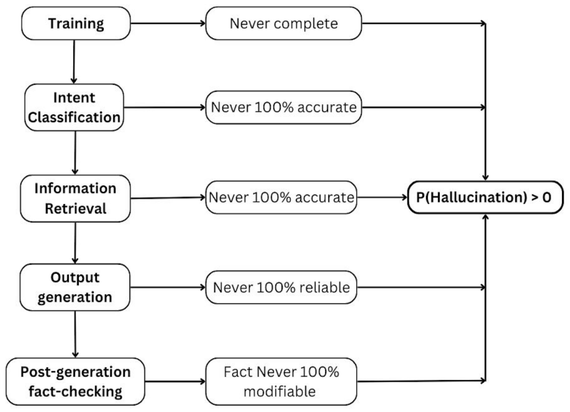

Fig 1: Limitations associate at each step of LLM leading to non-zero probability of hallucinations [2]

RESEARCH QUESTIONS

- How well does Selective RAG reduce hallucinations and ChromaDB-based memory improve context in medical chatbots?
- Can a hybrid of ChromaDB memory, RAG, and Phi-2 improve overall response reliability?

METHODOLOGY

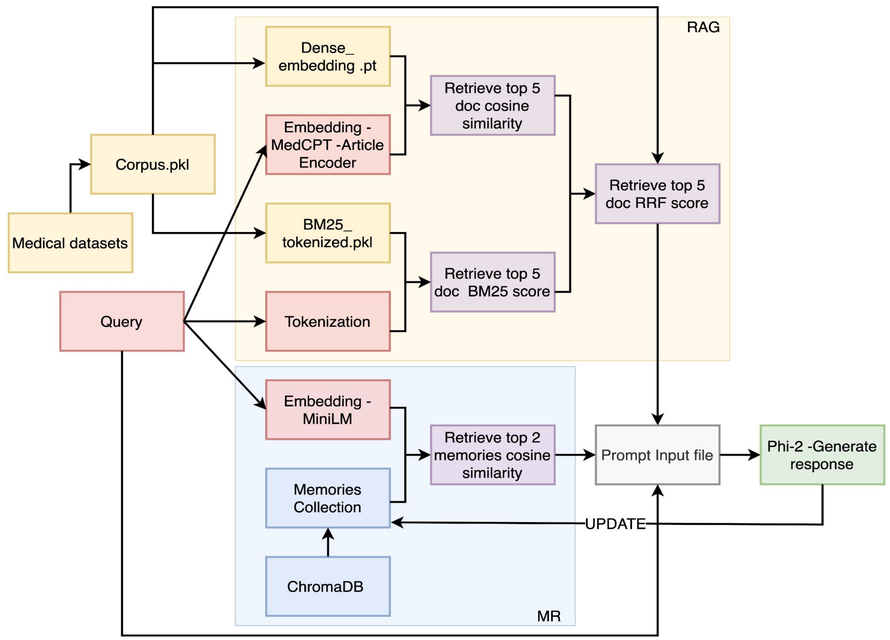

Fig 2: Detailed Architecture of our proposed hybrid system

- A unified corpus of 216,102 QA samples is created from BioASQ, MedQuAD, MedMCQA, and Kaggle.
- Documents are encoded using MedCPT embeddings and BM25 tokenization.
- RRF aggregation combines rankings, prioritizes overlaps, resolves ties with BM25, and always includes one MedCPT result.
- Top-2 memories are retrieved if cosine similarity ≥ 0.4.
- Prompt file includes instructions, query, top-5 docs (≤150 tokens each), top-2 memories (≤200 tokens), and an answer placeholder.
- Prompt length is limited to <1024 tokens for Phi-2 compatibility.
- If no relevant content is retrieved, the chatbot outputs: "no response".

Component	Type	Details
ChromaDB	Vector Database	Stores & retrieves user memories
BM25	Lexical Retriever	Token based keyword search
MedCPT	Semantic Retriever	Embedding based document retrieval
RRF	Rank Aggregator	Combines BM25 & MedCPT rankings
MiniLM	Embedding Model	Vector dimension = 384
MedCPT Encoder	Embedding Model	Vector dimension = 768
Prompt file	Input Format	Merged memories + documents + query
Phi-2	Language Model	2.7 B parameters

Table 1: Details about the frameworks used

RESULTS

Metrics	Mistral with RAG (baseline) [1]	FT Mistral with RAG (baseline) [1]	Our System
Dataset	Meadow-MedQA		MedQuAD
BERTScore F1	0.181	0.221	**0.8644**
Rouge-L	0.2512	0.221	0.2273
Perplexity	6.4691	4.84	**12.8758**
Avg. Time (s)	78	150	**28**

High BERTScore F1: strong semantic alignment with ground truth.

Higher perplexity: slightly reduced fluency.

Table 2: Comparative performance of our medical chatbot system against baseline (Q/A - 20 samples)

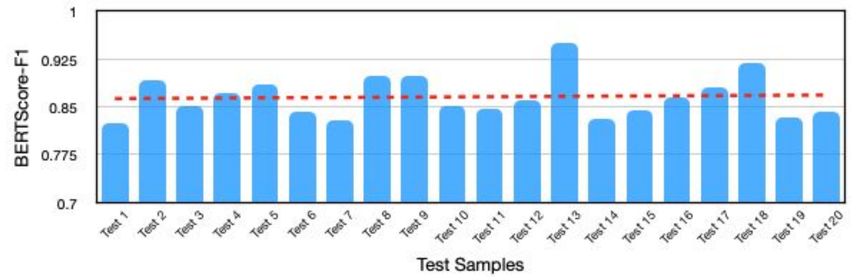

Fig 3: BERTScore F1 across 20 test Q/A (MedQuAD) from our hybrid system

Consistently high scores (+0.84) indicates strong semantic alignment with ground truth, suggesting low hallucination tendencies.

Memory Context	BERTScore F1	Perplexity
None	0.869	11.59
1 Memory	0.852	10.23
2 Memory	**0.847**	**8.69**

Memory context improves fluency but does not consistently enhance semantic accuracy.

Table 3: Testing how memory context affects response on a small test queries

CONCLUSION

- Advanced RAG with ChromaDB memory retrieval helps reduce hallucinations by anchoring responses to relevant medical context.
- Memory context enhances fluency and coherence in responses, while fallback mechanisms ensure safety in the absence of reliable facts.

Reference
[1] Bora, A.; Cuayáhuitl, H. Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications. Mach. Learn. Knowl. Extr. 2024, 6, 2355-2374.
[2] Banerjee S, Agarwal A, Singla S; LLMs Will Always Hallucinate, and We Need to Live With This.