

Research Article

Hybrid Memory-Retrieval Model: Enhancing Trust in Medical Chatbots

Sagarika Singh¹

1. Rochester Institute of Technology, United States

Large language model-based medical chatbots face two major challenges: hallucination, where models generate plausible but incorrect information, and context loss in multi-turn conversations. These issues reduce trust and safety in healthcare applications. This work presents a hybrid memory-retrieval architecture that enhances factual accuracy and conversational continuity. The system combines a dual-retriever pipeline using BM25 and MedCPT with long-term memory retrieval via ChromaDB. Retrieved documents and past interactions are fused using Reciprocal Rank Fusion and passed to a compact language model (Phi-2) for response generation. When relevant context is missing, the model defaults to fallback instructions to avoid hallucinated outputs. Evaluation on the MedQuAD dataset shows strong semantic alignment (BERTScore F1 = 0.8644), improved fluency, and significantly faster response times compared to baseline retrieval-augmented models. This approach demonstrates the effectiveness of integrating structured memory with selective retrieval to build more trustworthy and reliable medical chatbots.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

I. Introduction

Large language models (LLMs) are increasingly used in medical chatbots to provide symptom checking, health information, and conversational support. However, two major limitations hinder their trustworthiness in healthcare: hallucinations, where the model generates plausible but incorrect information, and poor context retention across multi-turn conversations. These shortcomings can lead to misleading advice, reduced user trust, and unsafe outcomes. To address these issues, recent work has explored retrieval-augmented generation (RAG) to enhance factual grounding and memory-augmented systems to improve personalization. However, existing RAG-based systems continue to hallucinate when

retrieval fails or documents are irrelevant, and most memory-based methods struggle with long-term coherence and scalability.

In this work, we propose a hybrid architecture that integrates structured memory retrieval and selective document retrieval to support safer, context-aware medical conversations. Our system combines ChromaDB-based long-term memory storage with a dual-retriever pipeline using BM25 and MedCPT. The retrieved results are fused via Reciprocal Rank Fusion (RRF) and passed to a compact language model, Phi-2, through a token-limited prompt. A fallback mechanism is included to prevent hallucination when context is insufficient. We evaluate our system on the MedQuAD dataset, showing strong semantic alignment (BERTScore F1 = 0.8644), improved fluency, and significantly lower response latency compared to baseline RAG models. This paper demonstrates that combining memory and retrieval enables more reliable and responsive medical chatbot systems.

A. Motivation

Building on recent limitations, we propose a hybrid architecture that combines structured memory retention and retrieval-based factual grounding. Our system prioritizes long-term, user-specific context via ChromaDB-based memory retrieval while selectively using an advanced RAG pipeline for updated information. This approach addresses key challenges in medical chatbots—hallucination control, multi-turn continuity, and factual reliability—without compromising speed or user trust.

B. Research Goals

1. How effectively can the advanced RAG pipeline reduce hallucinations?
2. How well does ChromaDB-based memory improve context retention?
3. Can our hybrid model enhance the reliability, fluency, and factual consistency of chatbot responses?

II. Background

Large language models (LLMs) like Phi-2, a 2.7 billion-parameter transformer developed by Microsoft, generate text by predicting the next token based on previous input and training data. While effective, these models often suffer from hallucinations, confidently producing factually incorrect or unverifiable information. This issue is especially problematic in healthcare, where accuracy is critical. LLMs also lack persistent memory, which leads to context loss in multi-turn interactions. This can cause snowballing, where early mistakes are reinforced across a conversation due to forgotten context. To reduce

hallucination, Retrieval-Augmented Generation (RAG) pipelines retrieve external documents relevant to the query and add them to the model's prompt. However, hallucinations can still occur if retrieval fails or the model does not integrate the documents effectively. Most RAG systems also treat each query independently, without remembering user history. To enhance retrieval, both lexical and semantic methods are used. BM25 retrieves documents based on token frequency, while MedCPT uses dense biomedical embeddings for semantic search. These methods can be combined to improve relevance. To address context loss, memory retrieval is used to bring back relevant past interactions. In our system, ChromaDB, an open-source vector database, stores user-query-response pairs as embeddings. Similar entries are retrieved during new queries using cosine similarity, enabling coherent, personalized responses across sessions.

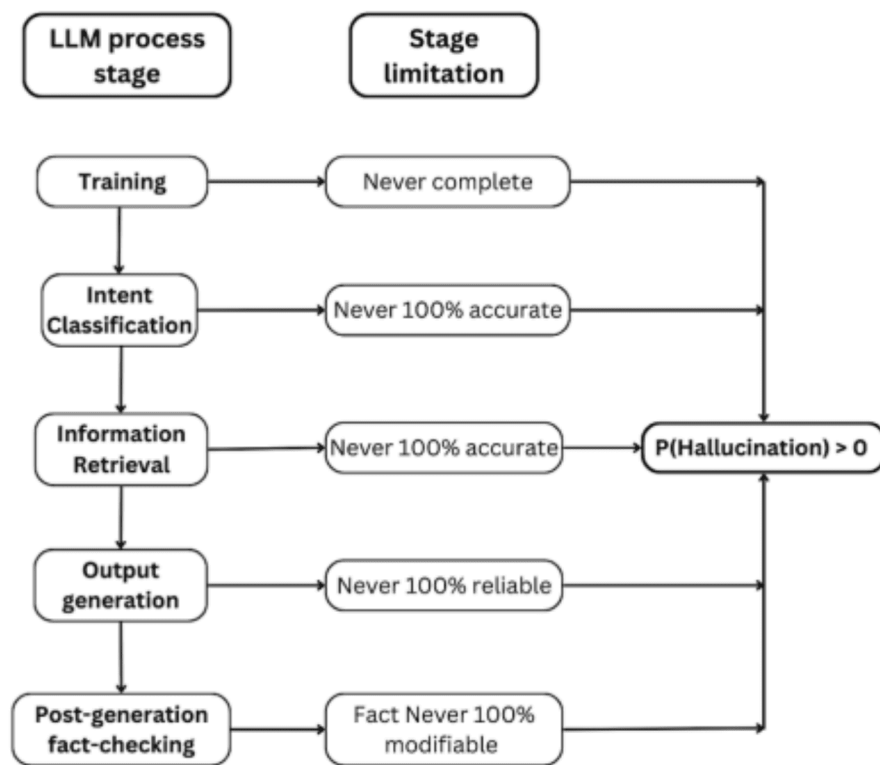


Figure 1. Limitations associated with each step of LLM leading to non-zero probability of hallucinations^[1]

III. Related Work

Large language models (LLMs) have significantly advanced open-domain and domain-specific question answering, enabling conversational agents to respond to complex queries in areas such as healthcare. Despite these capabilities, LLMs remain prone to hallucination, generating factually incorrect or unverifiable content with high confidence^[1]. They also struggle to retain long-term context across multi-turn interactions, which is particularly concerning in clinical settings where misinformation can have severe consequences. To address these challenges, Retrieval-Augmented Generation (RAG) architectures retrieve relevant documents and append them to the model's input to improve factual grounding^[2]. Cache-Augmented Generation (CAG) techniques extend this idea by incorporating structured memory modules that persist across sessions^[3]. However, recent systematic evaluations have shown that even advanced RAG-based systems frequently hallucinate, especially when retrieval fails or retrieved content is only loosely relevant^[4]. Moreover, most deployed medical chatbots still lack robust long-term memory and fallback mechanisms. They often forget prior user interactions, generate fabricated responses in low-retrieval scenarios, and exhibit latency or scalability issues in real-time use. These limitations highlight the need for architectures that tightly integrate retrieval, memory, and response safety.

IV. Methodology

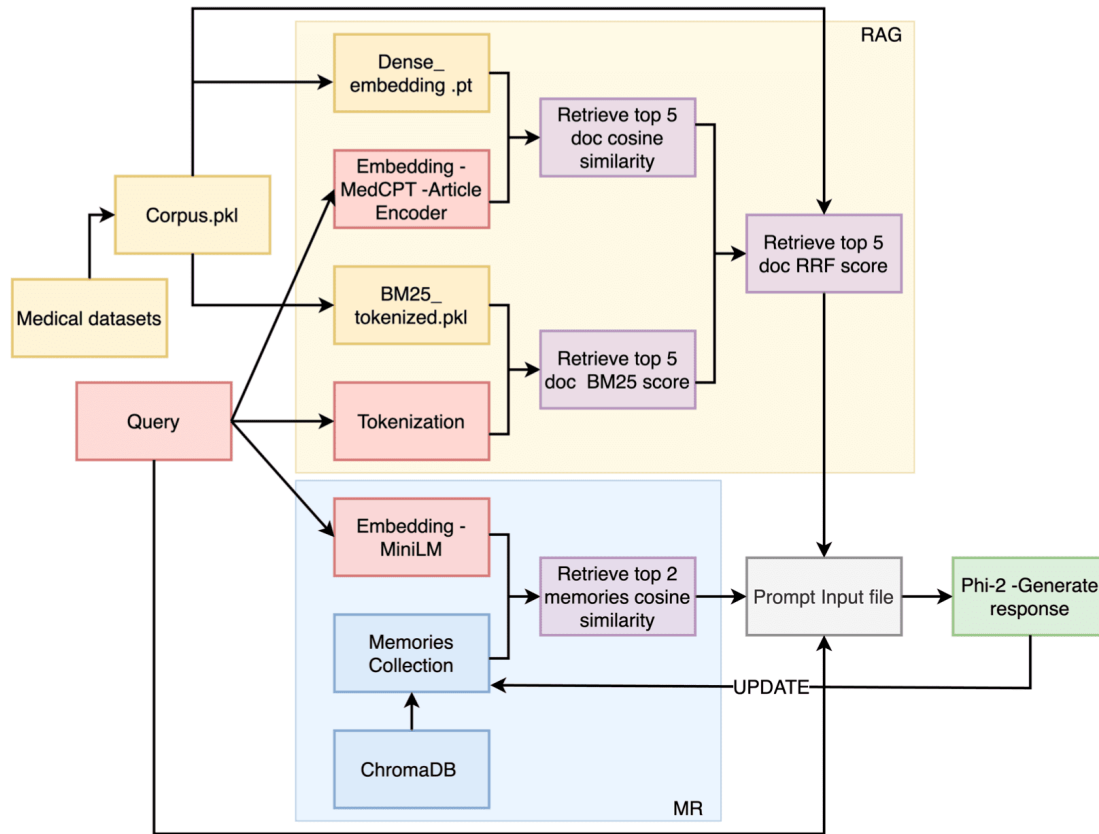


Figure 2. Architecture of our proposed method

A. Project Formulation

This project proposes a medical chatbot designed to mitigate hallucinations and improve multi-turn context retention. The core idea is to combine structured memory retrieval with selective document retrieval. The system retrieves relevant user-specific memories and external medical documents in parallel, integrating both into a focused prompt to support grounded and context-aware response generation.

B. Proposed Method

We implement a hybrid architecture that leverages both long-term conversational memory and real-time document retrieval. ChromaDB, a vector database, is used to store user interactions as dense embeddings.

At inference time, the current user query is embedded and compared with stored memories using cosine similarity, enabling the retrieval of relevant past interactions. For external document retrieval, we combine BM25—a lexical search model—with MedCPT, a semantic search model trained on biomedical literature. Results from both retrievers are merged using Reciprocal Rank Fusion (RRF) to balance keyword matching and semantic relevance. The final prompt is constructed by combining top-ranked documents and retrieved memory entries, constrained to 1024 tokens. This prompt is passed to Phi-2, a 2.7 billion parameter transformer-based language model. If insufficient context is available, the prompt includes a fallback instruction that directs the model to return a safe, conservative response.

C. Mathematical Formulas

Our system relies on three primary mathematical formulations to retrieve and rank relevant memories and documents:

Cosine Similarity is used to compare dense vector embeddings between the user’s query and stored memories or MedCPT document embeddings; given two vectors \vec{A} and \vec{B} :

$$\text{cosine_sim}(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|}$$

BM25 Scoring is used to retrieve top-matching documents from the corpus based on lexical word overlap between the query q and documents d :

$$\text{BM25}(q, d) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{f(t) \cdot (k_1 + 1)}{f(t) + k_1 \cdot (1 - b + b \cdot \frac{L_d}{L_{\text{avg}}})}$$

RRF combines the BM25 and MedCPT ranked retrieval lists into a final top-5 document selection by balancing lexical and semantic relevance; given k , which is a tuning constant, often set as 60, and $r_i(d)$ is the rank of document d :

$$\text{RRF}(d) = \sum \frac{1}{k + r(d)}$$

V. Experiments

A. Advanced RAG pipeline

To improve factual grounding and reduce hallucinations, we implement a selective Retrieval-Augmented Generation (RAG) pipeline that retrieves relevant documents from a curated medical corpus. The corpus

includes 216,102 samples sourced from MedQuAD^[5], MedMCQA^[6], BioASQ Task B^[7], and a Kaggle dataset. Each document entry contains metadata fields such as doc_id, text, title, source, and category. For retrieval, we use two complementary methods: BM25 for lexical matching and MedCPT for semantic similarity retrieval. The BM25-tokenized corpus is stored in `bm25_tokenized.pkl`, and documents are retrieved based on BM25 scoring. Dense embeddings are generated using the MedCPT/Article_Encoder model^[8] and saved in `dense_embeddings.pt`, allowing retrieval via cosine similarity. After retrieving the top 5 documents from each method, we apply RRF to obtain the final top 5 ranked documents. Documents retrieved by both methods are prioritized; in cases of equal RRF scores, BM25-retrieved documents are preferred, while ensuring that at least one MedCPT-retrieved document is included. This balanced retrieval strategy provides robust coverage of both exact-match and semantically relevant medical content. The performance of this pipeline is evaluated on the MedQuAD dataset using Recall@5, Precision@5, and BERTScore F1.

B. Dataset

To support document retrieval and grounding in medically accurate information, we constructed a corpus from four publicly available datasets. The combined dataset contains a total of 216,102 medical question-answer or passage entries. All datasets are accessible for academic or research use. The largest portion of the corpus comes from MedMCQA^[6] with 192,000 questions. MedQuAD^[5] with 17,236 entries. BioASQ Task B^[7] provides 4,065 samples. Finally, a Kaggle-sourced symptom and treatment dataset adds approximately 801 entries.

C. Memory Retrieval

To support multi-turn conversations and personalized interaction, we implement a memory retrieval system using ChromaDB. Each user query and corresponding chatbot response pair is stored as a memory entry, with embeddings generated using the MiniLM model (dimension = 384). At runtime, the user's current query is embedded and compared against stored memories using cosine similarity. If the similarity exceeds a threshold of 0.4, the top 2 most similar memory entries are retrieved. All new conversations are appended to the memory store, and users can delete their stored memory anytime by typing "clear." The performance of the memory retrieval system is evaluated on the MedQuAD dataset using BERTScore F1 and Perplexity metrics.

D. Medical Chatbot – Integrating all Components

During interaction, the chatbot first pre-processes the input query and simultaneously retrieves relevant context from both the memory module and the document corpus. The retrieved results are assembled into a structured prompt containing instructions, user query, extracted documents, and extracted memories, constrained to a 1024-token limit. This final prompt is passed to the Phi-2 model, which generates the response. If neither memory nor document retrieval provides sufficient context, the prompt includes fallback instructions, prompting the model to return a safe, conservative response. All interactions are stored for future memory retrieval, continuously improving personalization over time. The complete system performance is evaluated on the MedQuAD dataset using BERTScore F1, Perplexity, and average time taken per response.

E. Privacy and User Control

To promote user trust and transparency, the system incorporates mechanisms that allow users to manage their stored conversational data. All prior interactions—comprising user queries and corresponding chatbot responses—are stored as dense embeddings in ChromaDB for memory retrieval. During subsequent interactions, users can view which memory entries were retrieved and used to generate a response. Additionally, users are provided with a simple "clear" command that deletes all stored memory entries associated with their session. This design empowers users to retain control over their data and ensures that the system does not accumulate or use personal information without consent. By offering user-level control over memory retention, the system aligns with emerging best practices in privacy-aware conversational AI.

F. Implementation Details

The chatbot system was implemented in Python and developed within a Jupyter Notebook environment for iterative testing and demonstration. Retrieval components utilize the rank_bm25 package for lexical search and the MedCPT dense retriever model via Hugging Face's transformers and sentence-transformers libraries. Memory retrieval is powered by ChromaDB, an open-source vector database, using the ChromaDB Python client. Query and memory embeddings were generated using MiniLM (for memory) and MedCPT (for documents). Prompt construction and response generation were handled by the Phi-2 language model, integrated using Hugging Face's AutoModelForCausalLM. All components were executed on a local CPU/GPU setup without deployment to a web interface. While the current system

runs in a notebook environment, it is modular and can be extended into a production-ready pipeline with minor architectural changes.

VI. Results and Discussion

Our fully integrated system (Advanced RAG + Memory + Phi-2) against two strong baselines: Mistral with RAG and fine-tuned Mistral with RAG^[4]:

Metrics	Mistral with RAG	FT Mistral with RAG	Our Hybrid System
Dataset	Meadow-MedQA	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

Table 1. Performance comparison of our hybrid memory-retrieval system with baseline RAG-based models (20 QA samples)

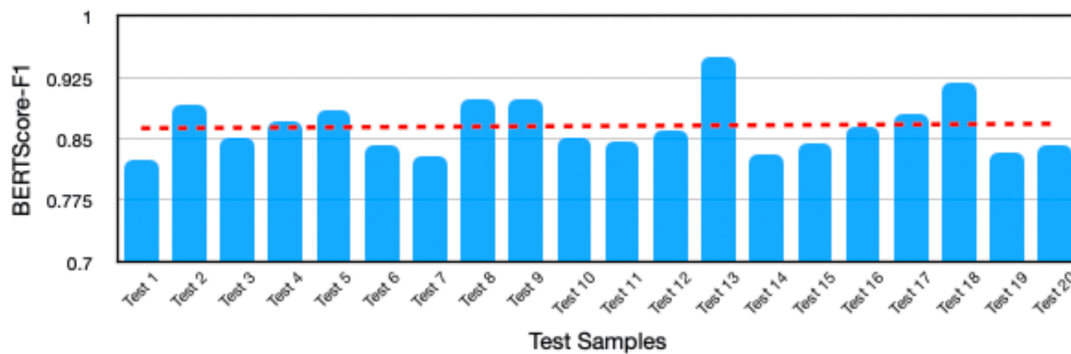


Figure 3. BERTScore F1 across 20 test Q/A (MedQuAD) from our hybrid system

Our system outperforms both baselines in BERTScore F1, indicating strong semantic alignment and low hallucination tendencies. Although perplexity is higher than the fine-tuned baseline, it is an acceptable trade-off given our fallback handling and factual grounding. As shown in Table 2, memory context

improves fluency, with perplexity decreasing as more memories are retrieved. However, since many queries were first-time questions, the memory module's full impact was limited during testing. Despite this, the system maintains strong relevance and is significantly faster, making it suitable for deployment. The impact of memory context (0, 1, and 2 retrieved memories) on generation quality:

Memory Context	BERTScore F1	Perplexity
None	0.8692	11.595
1 Memory	0.8520	10.23
2 Memory	0.84727	8.69

Table II. Effect of memory context on semantic quality and fluency of the chatbot's response

Adding 1 or 2 memory entries slightly reduced semantic alignment. However, perplexity improved significantly, suggesting memory improves fluency and coherence in multi-turn responses.

The Advanced RAG system alone, to benchmark pure retrieval quality:

Metric	Value
Recall@5	0.7500
Precision@5	0.2600
BERTScore F1	0.8654

Table III. Selective RAG system evaluation using retrieval and semantic alignment metrics.

The testing indicates strong coverage of relevant documents. Low precision is expected due to high document variability. Despite that, BERTScore F1 shows that RAG still retrieves semantically aligned content effectively.

A. End-to-End Chatbot Execution Demo

We present a series of screenshots demonstrating the complete execution flow of the chatbot, from user input and retrieval to final response generation and memory storage.

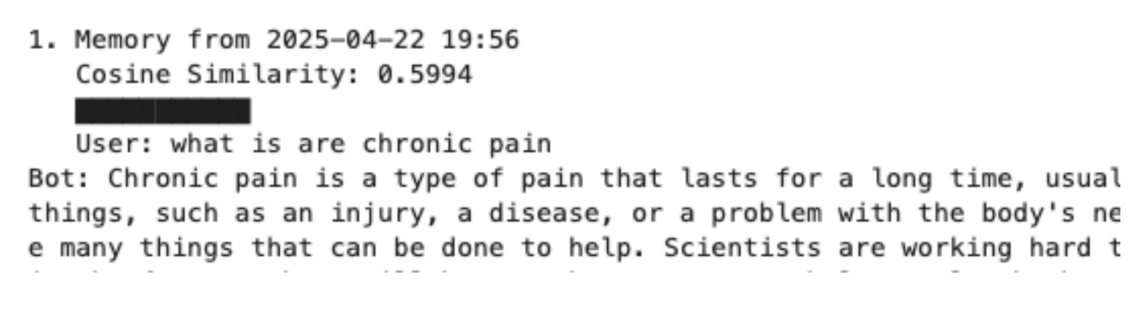


Figure 4. Memory Retrieval via ChromaDB

Doc ID	BM25 Rank	MedCPT Rank	RRF Score	Source
18	1	2	0.032522	both
895	2	-	0.016129	bm25
7983	3	-	0.015873	bm25
173656	4	-	0.015625	bm25
214605	5	-	0.015385	bm25
19346	-	1	0.016393	medcpt
86347	-	3	0.015873	medcpt
81193	-	4	0.015625	medcpt
73097	-	5	0.015385	medcpt

Figure 5. Documents extracted through BM25 and MedCPT system

Rank	Doc ID	RRF Score	BM25 Rank	MedCPT Rank
1	18	0.032522	1	2
2	895	0.016129	2	–
3	7983	0.015873	3	–
4	173656	0.015625	4	–
5	19346	0.016393	–	1

Figure 6. Documents finalized through RRF


 Assistant: (35.95s): Chronic pain is a type of pain that lasts for a long time, by many different things, such as an injury, a disease, or a problem with the body with, but there are many things that can be done to help. Scientists are working hard to help people who believe that in the future, there will be even better ways to help people who

Figure 7. Chatbot's response based on prompt file

VII. Conclusion

This work presents a hybrid chatbot architecture that integrates structured memory retrieval with selective document retrieval to improve the factual accuracy and contextual consistency of medical conversations. By combining ChromaDB-based long-term memory with a dual-retriever RAG pipeline (BM25 and MedCPT, fused via RRF), the system constructs personalized and evidence-grounded prompts. A fallback mechanism ensures response safety when retrieval is insufficient. Experimental results demonstrate high semantic alignment, reduced latency, and improved fluency in multi-turn scenarios, establishing a scalable foundation for trustworthy medical dialogue systems.

VIII. Contributions

This project introduces a novel hybrid chatbot architecture that unifies structured memory retrieval and dual-mode document retrieval into a single response generation framework. The system integrates ChromaDB-based long-term memory to retain and recall past user interactions, enabling personalization and multi-turn coherence. To ground responses in reliable medical content, it employs a dual-retriever strategy—combining BM25 for lexical relevance and MedCPT for semantic alignment—merged through

Reciprocal Rank Fusion (RRF). A key contribution is the inclusion of a fallback prompting mechanism that ensures safe, non-hallucinatory responses when memory or document retrieval yields insufficient context. We also curated a multi-source medical corpus to support robust and diverse retrieval. Collectively, these innovations result in a chatbot that is not only semantically aligned and fluently responsive but also significantly faster and more trustworthy than baseline RAG-only systems, offering a viable path toward scalable, memory-aware, medically grounded AI assistants.

IX. Limitations

While the system demonstrates promising performance, several limitations must be acknowledged. First, the use of Phi-2, a lightweight transformer model, constrains the expressive and reasoning capabilities of the chatbot compared to larger-scale LLMs. This trade-off, though beneficial for latency and efficiency, may limit complex clinical reasoning or nuanced understanding. Second, the system was evaluated primarily on the MedQuAD dataset and a limited sample of synthetic queries, which may not fully reflect the diversity of real-world medical interactions. Finally, the memory module relies on fixed thresholds for similarity and does not yet support dynamic memory pruning, potentially leading to inefficiencies as memory grows.

X. Future Work

Future research will aim to enhance the system's generalizability, robustness, and usability in real-world medical environments. Incorporating larger or domain-adapted language models could improve reasoning and fluency while maintaining factual accuracy. Further, the medical corpus can be expanded to include real-time clinical updates, research articles, and EMR-compatible content, enhancing retrieval breadth.

XI. Ethical Considerations

This system is intended solely for medical question-answering and informational purposes, not for clinical diagnosis, triage, or treatment recommendations. It is designed to provide users with factual information about medical topics, conditions, and treatments sourced from public medical datasets. It should not be used as a substitute for consultation with licensed healthcare professionals. To mitigate risks of misinformation, the system includes a fallback mechanism that instructs the model to return

conservative or non-committal responses when relevant context is unavailable. This helps reduce the likelihood of hallucinated or misleading outputs.

All data used to build and evaluate the system—MedMCQA, MedQuAD, BioASQ, and a Kaggle dataset—are publicly available and contain no patient-identifiable or confidential data. The only data stored during chatbot interaction are user queries and chatbot responses, which are embedded and saved in ChromaDB to support memory-based retrieval. To preserve user privacy, users are assigned random, non-identifying user IDs, and no account or identity linkage is maintained. Users have full control over their stored interactions and may issue a "clear" command at any time to delete all memory entries associated with their session. These design decisions aim to promote transparency, trust, and ethical use of AI in sensitive domains like healthcare.

Future deployment in clinical or production environments would require additional safeguards, such as encryption at rest, access controls, and compliance with data protection standards such as HIPAA or GDPR.

XII. Reproducibility

To promote transparency and support reproducibility, the complete codebase, datasets, preprocessing scripts, and evaluation notebooks are publicly available on GitHub: [GitHubRepository](#).

Statements and Declarations

Acknowledgments

I would like to express my sincere gratitude to Prof. Zhiqiang Tao at Rochester Institute of Technology for their invaluable guidance, feedback, and support throughout the course of this project. Their insights were instrumental in shaping the direction of this work.

References

1. ^a ^bBanerjee S, Agarwal A, Singla S (2024). "LLMs Will Always Hallucinate, And We Need To Live With This." arXiv. <https://arxiv.org/abs/2409.05746>.
2. ^ΔXiong G, Jin Q, Lu Z, Zhang A (2024). "Benchmarking Retrieval-Augmented Generation For Medicine." arXiv. <https://arxiv.org/abs/2402.13178>.

3. ^aChan BJ, Chen CT, Cheng JH, Huang HH (2024). "Don't Do RAG: When Cache-Augmented Generation Is All You Need For Knowledge Tasks." *arXiv*. <https://arxiv.org/abs/2412.15605>.
4. ^a, ^bBora A, Cuayáhuitl H (2024). "Systematic Analysis Of Retrieval-Augmented Generation-Based LLMs For Medical Chatbot Applications." *Mach Learn Knowl Extr*. 6(4):2355–2374.
5. ^a, ^bAbacha AB, Demner-Fushman D (2019). "A Question-Entailment Approach To Question Answering." *BM C Bioinform*. 20(1):511:1–511:23. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3119-4>.
6. ^a, ^bPal A, Umapathi LK, Sankarasubbu M (2022). "MedMCQA: A Large-Scale Multi-Subject Multi-Choice Dataset For Medical Domain Question Answering." In: Flores G, Chen GH, Pollard T, Ho JC, Naumann T, editors. *Proceedings Of The Conference On Health, Inference, And Learning*. PMLR. pp. 248–260. (Proceedings of Machine Learning Research; vol. 174). <https://proceedings.mlr.press/v174/pal22a.html>.
7. ^a, ^bTsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, Weissenborn D, Krithara A, Petridis S, Polychronopoulos D, Almirantis Y, Pavlopoulos J, Baskiotis N, Gallinari P, Artières T, Ngomo ACN, Heino N, Gaussier E, Barrio-Alvers L, Schroeder M, Androutsopoulos I, Paliouras G (2015). "An Overview Of The BIOASQ Large-Scale Biomedical Semantic Indexing And Question Answering Competition." *BMC Bioinform*. 16:138. doi:10.1186/s12859-015-0564-6.
8. ^aJin Q, Kim W, Chen Q, Comeau DC, Yeganova L, Lu Z (2023). "MedCPT: Contrastive Pre-Trained Transformers With Large-Scale PubMed Search Logs For Zero-Shot Biomedical Information Retrieval." *Bioinformatics*. 39(11):btad651.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.