Hybrid Memory-Retrieval Model: Enhancing Trust in Medical Chatbots

Anonymous ACL submission

Abstract

Medical chatbots powered by large language models (LLMs) face two critical challenges: hallucination, where the model produces plausible but incorrect responses, and loss of context in multi-turn conversations. These issues undermine reliability and trust in healthcare settings. This paper introduces a hybrid memory-retrieval architecture designed to enhance factual grounding and conversational coherence. The system integrates a dual-retriever pipeline (BM25 and MedCPT) with long-term memory retrieval using ChromaDB. Retrieved documents and past interactions are fused via Reciprocal Rank Fusion and provided as input to a compact language model (Phi-2) for response generation. A fallback mechanism is employed when insufficient context is available to reduce hallucinated responses. Evaluation on the MedQuAD dataset demonstrates high semantic alignment (BERTScore F1 = 0.8644), improved fluency, and significantly faster response times compared to baseline retrieval-augmented models. These results support the effectiveness of combining structured memory with selective retrieval to develop more trustworthy medical dialogue systems.

1 Introduction

Large language models (LLMs) are increasingly adopted in medical chatbots to support symptom checking, deliver health information, and facilitate conversational interactions. Despite their growing use, two fundamental challenges limit their reliability in healthcare applications: hallucination, where the model produces confident yet incorrect information, and insufficient context retention across multi-turn conversations. These issues can result in misleading advice, decreased user trust, and unsafe interactions.

To mitigate these limitations, retrieval-augmented generation (RAG) techniques have been introduced to improve factual grounding, while memory-augmented systems aim to enhance personalization and context continuity. However, RAG systems are still susceptible to hallucination when retrieval is incomplete or misaligned, and memory-based approaches often face scalability and coherence constraints.

This work presents a hybrid architecture that integrates structured memory retrieval with selective document retrieval to enable safer, more context-aware medical dialogue. The system combines long-term memory via ChromaDB with a dual-retriever pipeline leveraging BM25 and MedCPT. Retrieved content is merged through Reciprocal Rank Fusion (RRF) and formatted into a token-limited prompt for a compact LLM (Phi-2). A fallback strategy is incorporated to reduce hallucination in cases of insufficient context. Evaluation on the publicly available MedQuAD dataset shows strong semantic alignment (BERTScore F1 = 0.8644), improved fluency, and substantially lower response latency compared to baseline RAG systems. These findings support the effectiveness of combining memory and retrieval for building more trustworthy and responsive medical chatbots.

1.1 Motivation

Given the limitations of current medical dialogue systems, this work explores a hybrid architecture that combines structured memory retention with retrieval-based factual grounding. The proposed system emphasizes long-term, user-specific context through ChromaDB-based memory retrieval while selectively employing an advanced retrieval pipeline for access to recent medical knowledge. This dual approach aims to mitigate hallucinations, maintain conversational continuity across turns, and ensure factual reliability without compromising on system efficiency or user trust.

1.2 Research Goals

This study is guided by the following research questions:

- To what extent can an advanced RAG pipeline reduce hallucinated outputs in medical chatbots?
- How effectively does ChromaDB-based memory retrieval improve multi-turn context retention?
- Can a hybrid memory-retrieval system enhance response reliability, fluency, and factual consistency in medical question-answering?

2 Background

Large language models (LLMs) such as Phi-2, a 2.7 billion-parameter transformer, generate responses by predicting the next token based on prior inputs and training data. While effective in generating fluent text, LLMs frequently suffer from hallucinations—plausible but factually incorrect outputs—which can be especially problematic in high-stakes domains like healthcare. These models also lack persistent memory, leading to context loss across multi-turn conversations. This may result in snowballing effects, where early inaccuracies propagate due to forgotten context.

To mitigate hallucinations, Retrieval-Augmented Generation (RAG) pipelines have been proposed. These systems augment the prompt with documents retrieved from external sources using information retrieval techniques. However, hallucinations can still arise if retrieval fails or if the language model inadequately integrates retrieved evidence. Furthermore, most RAG systems treat each user query independently and do not incorporate prior conversation history.

To improve retrieval, both lexical and semantic methods are employed. Lexical models, such as BM25 score documents based on token frequency and overlap, while semantic methods, such as MedCPT utilize dense biomedical embeddings to capture deeper contextual similarity. Combining these methods can enhance retrieval relevance.

In addition, memory-based retrieval has emerged as a strategy to address long-term context retention. Vector databases like ChromaDB can store conversation history (e.g., user queries and system responses) as embeddings. During inference, new queries are compared against this memory bank using similarity metrics such as cosine similarity, allowing the system to retrieve relevant prior interactions and produce more coherent, personalized responses.

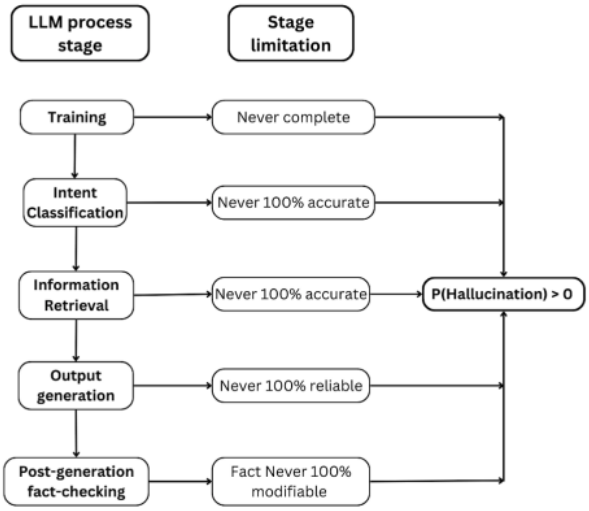

Figure 1: Limitations associated with each step of LLM leading to non-zero probability of hallucinations (Banerjee et al., 2024)

3 Related Work

Large language models (LLMs) have driven substantial progress in both open-domain and specialized question answering, including medical applications. While these models demonstrate strong generative capabilities, they remain vulnerable to hallucination, producing factually incorrect or unverifiable information with high confidence (Banerjee et al., 2024). LLMs also lack mechanisms for long-term context retention in multi-turn dialogues, which poses significant risks in clinical settings.

To address these issues, Retrieval-Augmented Generation (RAG) architectures have been developed to improve factual grounding by appending relevant documents to the model's input (Xiong et al., 2024). Cache-Augmented Generation (CAG) further extends this approach by integrating persistent memory modules that maintain historical context across sessions (Chan et al., 2024). Despite these advances, recent systematic evaluations indicate that hallucinations persist in many RAG-enhanced systems, particularly when retrieval is incomplete or misaligned with the user query (Bora and Cuayáhuitl, 2024).

Many real-world medical chatbots also continue to operate without reliable long-term memory or robust fallback mechanisms. They often fail to recall

prior user interactions, fabricate responses when retrieval fails, and struggle with latency or scalability during real-time usage. These ongoing limitations suggest a need for architectures that more effectively integrate document retrieval, memory retention, and response control mechanisms.

4 Methodology

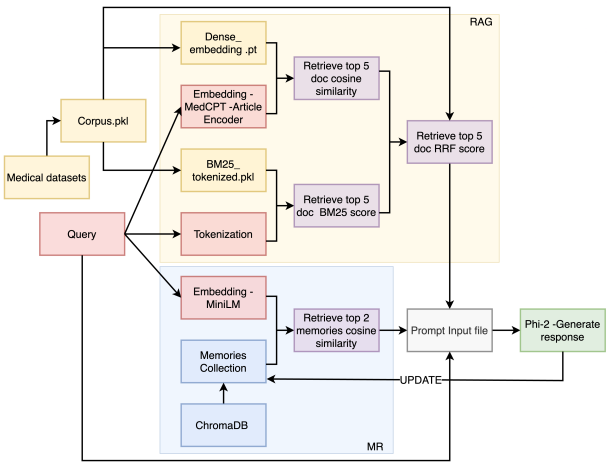

Figure 2: Architecture of our proposed method

4.1 Project Formulation

This work presents a medical chatbot architecture designed to reduce hallucinations and improve multi-turn context retention. The approach combines structured memory retrieval with selective document retrieval, enabling the system to fetch both user-specific past interactions and relevant external medical information in parallel. Retrieved contexts are assembled into a unified, focused prompt to support grounded and context-aware response generation.

4.2 Proposed Method

The system adopts a hybrid architecture that integrates long-term memory retrieval with real-time document search. Conversational memory is maintained using a vector database (e.g., ChromaDB), where past user interactions are stored as dense embeddings. At inference time, the current user query is embedded and compared with stored entries using cosine similarity, allowing retrieval of semantically similar memory segments.

For external knowledge retrieval, the system combines BM25—a lexical search model—with MedCPT, a dense retrieval model trained on biomedical literature. Outputs from both retrievers are merged using Reciprocal Rank Fusion (RRF), ensuring a balance between keyword matching and semantic relevance. The top-ranked documents and retrieved memory entries are combined into a token-limited prompt (maximum 1024 tokens), which is passed to a compact transformer-based language model (Phi-2, 2.7B parameters) for response generation. If the prompt lacks sufficient context, it includes fallback instructions directing the model to return a safe, conservative response.

5 Experiments

5.1 Advanced RAG Pipeline

To enhance factual grounding and mitigate hallucinations, a selective Retrieval-Augmented Generation (RAG) pipeline was developed to retrieve contextually relevant documents from a curated medical corpus. The corpus consists of 216,102 question-answer and passage samples derived from publicly available datasets, including MedQuAD (Ben Abacha and Demner-Fushman, 2019), MedMCQA (Pal et al., 2022), BioASQ Task B (Tsatsaronis et al., 2015), and a Kaggle-hosted medical QA dataset.

Two complementary retrieval methods are employed: BM25 for lexical matching and MedCPT for semantic similarity using dense biomedical embeddings. Retrieved documents from each method are independently ranked. The final selection is determined using Reciprocal Rank Fusion (RRF), which balances token-based and embedding-based relevance. In cases of tie scores, documents prioritized by BM25 are selected, while ensuring that at least one semantically relevant document from MedCPT is included. This retrieval strategy offers robust coverage across both exact-match and semantically aligned documents.

Performance is evaluated on the MedQuAD dataset using standard information retrieval metrics: Recall@5, Precision@5, and BERTScore F1.

5.2 Dataset

The document retrieval system is built upon a combined medical corpus containing 216,102 entries. All sources are publicly available and intended for academic and research purposes. The largest subset originates from MedMCQA (Pal et al., 2022), comprising 192,000 multiple-choice medical questions. MedQuAD (Ben Abacha and Demner-Fushman, 2019) contributes 17,236 question-answer pairs collected from NIH websites. BioASQ Task B (Tsatsaronis et al., 2015) provides 4,065 biomed-

ical factoid and list-based QA samples. Finally, a publicly shared dataset from Kaggle adds 801 entries related to symptoms and treatments. Each dataset entry includes textual content, metadata (e.g., title, category), and source references.

5.3 Memory Retrieval

To support multi-turn dialogue and personalized interaction, the system incorporates a memory retrieval mechanism using a vector database. Each user query and corresponding response is stored as a memory entry, encoded into dense embeddings via MiniLM (embedding dimension = 384). At inference time, the current user query is embedded and compared against stored entries using cosine similarity. If similarity exceeds a predefined threshold (0.4), the top two most similar entries are retrieved.

This memory module enables the chatbot to recall relevant prior interactions and maintain coherence across sessions. All new interactions are automatically appended to the memory store, and mechanisms for memory management (e.g., clearing past interactions) are included to support user control. The memory retrieval system is evaluated on the MedQuAD dataset using BERTScore F1 and Perplexity.

5.4 Medical Chatbot: Integration of Components

The end-to-end chatbot integrates both document retrieval and memory retrieval modules to generate informed responses. Upon receiving a user query, the system pre-processes the input and simultaneously performs memory-based and corpus-based retrieval. The retrieved content—including instructions, relevant documents, and prior conversation memory—is assembled into a structured prompt limited to 1024 tokens.

This composite prompt is passed to the Phi-2 language model for generation. In cases where neither document nor memory retrieval yields adequate context, a fallback instruction is included in the prompt to encourage a safe, conservative output. All chatbot interactions are persistently stored to enhance personalization and continuity in future sessions. System performance is assessed using BERTScore F1, Perplexity, and average response latency on the MedQuAD dataset.

5.5 Privacy and User Control

To support privacy-aware conversational AI, the system incorporates features that allow users to manage stored interaction data. All user queries and chatbot responses are encoded and stored as dense embeddings in a vector database for memory retrieval. During inference, relevant memory entries are retrieved and used to condition the model's response. The system includes mechanisms for users to review the memory content influencing their responses and to delete their stored memory entries at any time. This functionality ensures that user data is neither persistently retained nor used without consent. By offering transparent memory management, the system aligns with emerging best practices in responsible AI and privacy-centric chatbot design.

5.6 Implementation Details

The system was developed in Python using modular components for retrieval, memory, and generation. Lexical retrieval was implemented using BM25, while dense retrieval employed MedCPT-based embeddings for semantic similarity. Memory embeddings were generated using MiniLM and stored using a vector database client. Prompt construction modules integrated both memory and document retrieval results, constrained to a token limit. Response generation was performed using the Phi-2 language model via a standard transformer-based causal generation interface.

6 Results and Discussion

The complete hybrid system (incorporating the advanced RAG pipeline, memory retrieval, and Phi-2 for generation) was compared against two baselines: Mistral with RAG and fine-tuned Mistral with RAG (Bora and Cuayáhuitl, 2024). Results are summarized in Table 1.

Metric	**Mistral + RAG**	**FT Mistral + RAG**	**Hybrid Memory-RAG 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. Response Time (s)	78	150	**28**

Table 1: Comparison of the hybrid memory-retrieval system with baseline RAG-based models on 20 QA samples.

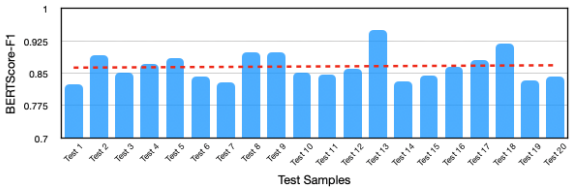

Figure 3: BERTScore F1 across 20 test QA pairs (MedQuAD) for the hybrid system.

The hybrid system demonstrates a substantial improvement in semantic alignment, as reflected by the BERTScore F1 metric, while offering significantly lower latency than fine-tuned RAG baselines. Although perplexity is higher, this is attributable to the conservative fallback strategy, which prioritizes safety in the absence of adequate context. ROUGE-L performance remains comparable across models.

To assess the impact of memory context on generation, Table 2 presents results for three scenarios: no memory, one retrieved memory, and two retrieved memories.

Memory Context	**BERTScore F1**	**Perplexity**
No Memory	0.8692	11.595
1 Memory	0.8520	10.230
2 Memories	0.8473	**8.690**

Table 2: Effect of memory context on semantic alignment and fluency.

While the addition of memory entries leads to a slight reduction in BERTScore F1, it significantly improves fluency, as indicated by decreased perplexity. This suggests that memory retrieval contributes to more coherent and contextually grounded multi-turn responses.

To isolate the retrieval component, Table 3 presents evaluation results of the RAG pipeline alone (excluding memory or generation):

Metric	**Value**
Recall@5	0.750
Precision@5	0.260
BERTScore F1	0.8654

Table 3: Evaluation of the selective RAG pipeline on document retrieval.

These results confirm that the retrieval pipeline successfully identifies relevant documents, with high recall and semantic alignment. Lower precision is expected due to variability in document formats and medical terminology.

7 Conclusion

A hybrid medical chatbot architecture was developed by integrating structured memory retrieval with selective document retrieval to enhance the factual accuracy and contextual relevance of multi-turn interactions. The system combines long-term memory stored in a vector database with a dual-retriever RAG pipeline (BM25 and MedCPT, fused via Reciprocal Rank Fusion). Prompt construction incorporates both sources of context and includes a fallback mechanism for safe response generation in low-retrieval scenarios. Empirical results indicate improved semantic alignment, response fluency, and lower latency compared to RAG-only baselines, providing a foundation for scalable, trust-enhancing medical dialogue systems.

8 Limitations

Several limitations affect the current implementation. The use of Phi-2, a lightweight transformer model, restricts expressive capacity and complex reasoning compared to larger-scale LLMs. While beneficial for latency and resource efficiency, this may limit utility in highly nuanced clinical contexts. Evaluation was conducted primarily on the MedQuAD dataset and a small sample of synthetic queries, limiting generalizability to broader or real-world user populations. Additionally, the memory retrieval module uses static similarity thresholds and lacks dynamic memory management, which may lead to inefficiencies or retrieval noise as stored data grows.

9 Future Work

Subsequent work may explore the integration of larger or domain-specialized language models to improve reasoning, fluency, and naturalness of responses. Expanding the medical corpus to incorporate real-time clinical guidelines, medical literature, and EMR-compatible content could further enhance retrieval relevance. Introducing adaptive memory management and re-ranking strategies may improve memory efficiency and relevance over time. Additional real-world testing and longitudinal evaluations are also necessary to assess robustness, usability, and trust under deployment conditions.

10 Ethical Considerations

The system is designed for medical question-answering and educational purposes only and is not intended to diagnose, treat, or manage medical conditions. It provides factual information sourced from publicly available medical datasets, such as MedMCQA, MedQuAD, BioASQ, and a publicly shared Kaggle dataset. No private or patient-identifiable data is included.

To minimize potential harm, a fallback mechanism is used to prevent hallucinated or speculative responses when relevant context is lacking. Users are assigned randomized session identifiers, and no personal identifying information is collected or stored. All interaction history is stored as embeddings solely to support contextual recall. Users retain full control over memory and may delete their stored history at any time.

The design reflects a privacy-preserving approach aligned with responsible AI practices. Future deployment in real-world or clinical contexts would necessitate further safeguards, including encryption, audit logs, user consent mechanisms, and compliance with regulatory standards such as HIPAA or GDPR. Ethical review and domain-specific oversight would also be essential before integration into any sensitive workflows.

References

Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. Llms will always hallucinate, and we need to live with this.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

A Bora and H Cuayáhuitl. 2024. Systematic analysis of retrieval-augmented generation-based llms for medical chatbot applications. *Machine Learning and Knowledge Extraction*, 6(4):2355–2374.

Brian J Chan, Chao-Ting Chen, Jui-Hung Cheng, and Hen-Hsen Huang. 2024. Don't do rag: When cache-augmented generation is all you need for knowledge tasks.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael Alvers, Dirk Weißenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine.