2025 Eighth International Women in Data Science Conference at Prince Sultan University (WiDS PSU) | 979-8-3315-2092-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/WIDS-PSU64963.2025.00030

C.A.L.I.X: Leveraging Hybrid RAG for Responsible AI in the Medical Field

Muddassir Khalidi
Artificial Intelligence and Data Analytics Lab
Prince Sultan University
Riyadh, Saudi Arabia
222110775@psu.edu.sa

Arwa Bawazir
Artificial Intelligence and Data Analytics Lab
Prince Sultan University
Riyadh, Saudi Arabia
arwaabdullah639@gmail.com

Tanzila Saba
Artificial Intelligence and Data Analytics Lab
Prince Sultan University
Riyadh, Saudi Arabia
tsaba@psu.edu.sa

Zainab Mariya Mohiuddin
Artificial Intelligence and Data Analytics Lab
Prince Sultan University
Riyadh, Saudi Arabia
zainab.mariya.mohiuddin@gmail.com

Saeed Lababidi
Artificial Intelligence and Data Analytics Lab
Prince Sultan University
Riyadh, Saudi Arabia
221111258@psu.edu.sa

Anees Ara
Artificial Intelligence and Data Analytics Lab
Prince Sultan University
Riyadh, Saudi Arabia
aara@psu.edu.sa

Asma Vaheed Khan
Artificial Intelligence and Data Analytics Lab
Prince Sultan University
Riyadh, Saudi Arabia
asmavaheedkhan4@gmail.com

Abdul Rahman Mamdouh
Artificial Intelligence and Data Analytics Lab
Prince Sultan University
Riyadh, Saudi Arabia
abdulrhmanstd@gmail.com

***Abstract*— Conversational AI memory systems work much like human memory. They record, maintain, and retrieve contextual information from earlier interactions, just as human memory draws on past experiences and learned knowledge to shape understanding and responses. However, individuals often face challenges with long-term recall of specific details from past events and short-term working memory, such as retaining lists, which can limit their ability to provide well-informed and nuanced responses. This research aims at implementing a simple Retrieve-Augment-Generate (RAG) system to implement a memory augmentation system with a particular focus on a use case in the medical field. This paper introduces C.A.L.I.X (Cognitive Archive for Learning and Information Exchange), an AI powered conversational memory assistant designed in alignment with Responsible AI principles. CA.L.I.X has been engineered to monitor conversations and extract data using voice interactions, ensuring that the information gathered is reliable. Future efforts will be focused on enhancing the accuracy of memory retrieval, streamlining the control of memory decay, and elevating the overall user experience.**

Keywords—RAG Application, Memory Management Systems, Conversational AI, Human Memory, AI Assistant, Memory Assistant, Healthcare Documentation

I. INTRODUCTION

As global connectivity continues to expand, conversational artificial intelligence (AI) has become a pivotal technology, transforming human-machine interactions and delivering seamless, intuitive solutions across diverse domains. Past implementations of such systems have focused on a more general replication of human memory by using different techniques of context storage. These techniques include creating taxonomies of contextual information [4], depending on the Ebbinghaus Forgetting Curve [3] for memory retention, using LLMs for semantic searching of relevant information [2], etc.

A. Problem Statement

The ability to efficiently manage and retain real-time information is essential in both industrial and personal settings, yet it comes with unique challenges and constraints. In industrial contexts, the focus is on optimizing storage efficiency and ensuring easy access to retrieved data. On a personal level, managing real-time information is hindered by limitations in memory retention, attention span, and recall accuracy. [7]. Studies on memory retention indicate that individuals tend to forget a substantial portion of information soon after hearing it. Within 15 minutes, approximately 50–60% of the information may still be retained, but recall declines rapidly over time. After an hour, retention can decrease by nearly half, leaving only 30–40% remembered accurately. Within 24 hours, this figure may drop to 20–30%, and after a week, only 10–20% of the information may be recalled without reinforcement [1]. These limitations in human memory emphasize the need for improved information management systems that enhance storage, retrieval, and support for memory retention, focus, and recall in daily life.

B. Objectives

- Develop a conversational memory assistant with a focus on a healthcare application, ensuring the system follows Responsible AI principles.
- Examine the issues users experience with current memory management solutions and evaluate the constraints of today's AI assistants.
- Categorize the different types of queries the system could face in a medical setting.
- Enhance the model's ability to mitigate and avoid hallucination.

C.A.L.I.X an AI-powered conversational memory assistant based on Responsible AI principles takes a simpler approach to implement a Hybrid RAG system which contributes to four aspects of Responsible AI [6]:

1) Accurate Responses: Responses which are factually correct according to the data stored.

2) Acknowledged Uncertainty: Responses when the model cannot provide an accurate response

979-8-3315-2092-2/25/$31.00 ©2025 IEEE
DOI 10.1109/WiDS-PSU64963.2025.00030

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 05:10:03 UTC from IEEE Xplore. Restrictions apply.

3) Corrective Clarification: Correcting the user's query if the query is posed in a way that could possibly lead to hallucination.

4) Verification: Allowing the user to verify the information retrieved by the model.

This paper details the architecture of C.A.L.I.X and the experimental results that highlight its effectiveness in handling memory-dependent tasks. Through these innovations, C.A.L.I.X aims to set a new standard for interactive, context-aware, and memory-capable conversational AI.

The rest of the paper is organized as follows: The section II includes a comprehensive literature review, Section III proposes CALIX the personal memory management system, section IV discusses a case study of CALIX in healthcare and section V shows implementation of the system and section VI Future directions and section VII concludes the paper.

II. LITERATURE REVIEW

Wearable systems such as Memoro [2] demonstrate potential in enhancing memory by leveraging large language models (LLMs) to facilitate real-time, context-aware recall through audio-based input. Memoro is designed with two distinct modes: Query Mode, which enables direct interactions, and Queryless Mode, which offers predictive assistance to minimize user interruptions during conversations.

Building on Memoro's work, OmniQuery, a state-of-the-art memory augmentation system, enhances personal question-answering by developing a taxonomy of contextual information to facilitate more accurate responses:

1) Atomic Context: Information directly obtainable from a single memory, such as time, location, people, or visual elements.

2) Composite Context: Context combining multiple atomic contexts, representing events or activities, eg., a "lab retreat" involving time, location and group activities.

3) Semantic Knowledge: High-level patterns or general knowledge inferred from multiple memories, such as "Jason goes to the gym 3-4 times a week.

Furthermore, they categorized questions into three different categories:

1) Direct Content Queries: Simple queries answered using explicit information in a single captured memory, e.g., "What is my driver's license number?"

2) Contextual Filters: Queries requiring retrieval based on specific contexts like time, location, or event, e.g., "All the photos in Hawaii."

3) Hybrid Queries: Complex queries combining content and context, requiring multi-hop reasoning, e.g., "Which meat did I order the last time I came to this Japanese BBQ restaurant?"

These implementations were general memory augmentation systems with innovative storage and retrieval techniques. Building on these systems, C.A.L.I.X was built with a focus on the medical field and consequently had to focus on simplicity while also taking particular care that the system aligned with Responsible AI principles, such as accountability, transparency and reliability.

C.A.L.I.X addressed an important limitation found in Memoro, i.e. verification of information. The method used to accomplish this was by allowing a way for the user to verify that the information stored and also the information being retrieved by the system.

Furthermore, the idea of categorizing the types of queries was highly useful to develop test cases to push C.A.L.I.X to the limits by testing it with five different categories of queries [8]:

1) Direct Recall: Require C.A.L.I.X to retrieve specific information from memory without external cues [5]. Example: What diagnosis was made for David during his visit?

2) Recognition-Based Recall: Present multiple options, and C.A.L.I.X must identify the correct one. Example: Was the condition we discussed as a possible cause for Mrs Brown headaches, tension headache, sinus infection, or migraine?

3) Contextual Recall: Require retrieval of information tied to specific contexts or scenarios [5]. Example: What type of foods did we identify as contributing to Mr Patel's stomach discomfort during our previous consultation?

4) Spatial Recall: Require retrieval of spatial information or locations. Example: Where on Mr Brown's neck did he report feeling the lump during his last visit?

5) Emotional Recall: Engage the recall of experiences or events tied to emotions. The context in the transcriptions does not take into account any vocal or facial attributes of emotions and only takes into account textual information. Example: How did Mr John feel when we discussed the potential need for further investigations, such as a biopsy, for the lump in his neck?

III. C.A.L.I.X: PROPOSED PERSONAL MEMORY MANAGEMENT SYSTEM

C.A.L.I.X (Cognitive Archive for Learning and Information Exchange) is a conversational memory augmentation system implemented with a focus on the medical industry. It includes two modules for interaction: one for storage and the other for retrieval. The storage module is implemented to store the user's consultations with audio-based functionality. Upon recording, the user is alerted of a document being created. This document can be used to verify the users' information. The document is created in accordance to the following prompt template:

```
You are a doctor listening to a consultation. Given below is a
transcription of a consultation with a patient. Return the
following information about the patient in JSON style text. If
any information is not provided add a "not provided" to it.
1. Date: {self.date}
2. Time: {self.time}
3. Name
4. Age (return as a string)
5. Presenting Complaint
6. History of Present Illness
7. Past Medical History
```


Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 05:10:03 UTC from IEEE Xplore. Restrictions apply.

8. Medication and Allergies
9. Family and Social History
10. Review of Systems - Assessment of other body systems
11. Physical Examination Findings
12. Assessment and Diagnosis
13. Treatment Plan
14. Advanced Decisions and Directives - Any advanced decisions to refuse treatment or directives discussed and agreed upon.

Furthermore, the retrieval module allows the user to record a question. C.A.L.I.X utilizes a vector store to retrieve relevant context and then uses an LLM to semantically search the context for an accurate response according to the following prompt template:

You are a doctor's personal memory assistant. Respond in natural language format as if you were speaking directly to the doctor. Only use the facts provided in the sources below, and do not make up information. If the answer isn't found in the sources, say, 'I don't know.' Maintain a conversational tone while sticking strictly to the provided information.
Facts: {context}
Question: {query}

A. Technical Overview

1) Storage and Retrieval Mechanism

C.A.L.I.X utilizes the Microsoft Azure AI Search integrating semantic search and vector-based similarity retrieval, enabling efficient handling of unstructured data. Key features include:

- **Vector Representation**: Content is encoded into 1536-dimensional vectors to support the embedding model being used for high-dimensional similarity search using cosine metrics.
- **Hybrid Search:** Combines BM25 for keyword-based ranking and HNSW (Hierarchical Navigable Small World) for approximate nearest-neighbor (ANN) retrieval.
- **Storage Configuration:** Metadata *(meta_data_storage_name)* and content (content) are indexed for efficient retrieval, with text vectors stored separately for similarity matching.
- **Scalability:** Supports fast and scalable search across large datasets, balancing precision with performance through configurable parameters *(e.g., efSearch, efConstruction).*

2) Audio Functionality

The system utilizes OpenAI's whisper-1 endpoint to convert speech to text before being passed into the embedding model to generate embeddings. When generating a response the system uses OpenAI's tts-1 endpoint to convert text to speech.

- **Storage Module:** The storage module needed to account for audio files which exceeded the limit, so the AudioSegment library from pydub was utilized to split the audio files before passing it to the whisper-1 endpoint.

3) Response Generation and Text

When the user's query is transcribed, the text is passed to the Azure AI endpoint and relevant context is retrieved. This context is passed to OpenAI's gpt-4o-mini endpoint and the LLM returns a response.

B. System Architecture

C.A.L.I.X is designed around a Retrieval-Augmented Generation (RAG) framework that integrates both speech-to-text (STT) and text-to-speech (TTS) capabilities for smooth user interaction [9]. When a user initiates a recording via the listening module, the STT component converts the spoken words into text. This text is then transformed into embeddings using an embedding model, and these embeddings are saved in a vector database. When retrieving information, the system again records the user's speech to generate embeddings, which are then compared against those in the database. The matching information is sent to the LLM to craft a response, which is ultimately converted back into speech through the TTS system and delivered to the user.

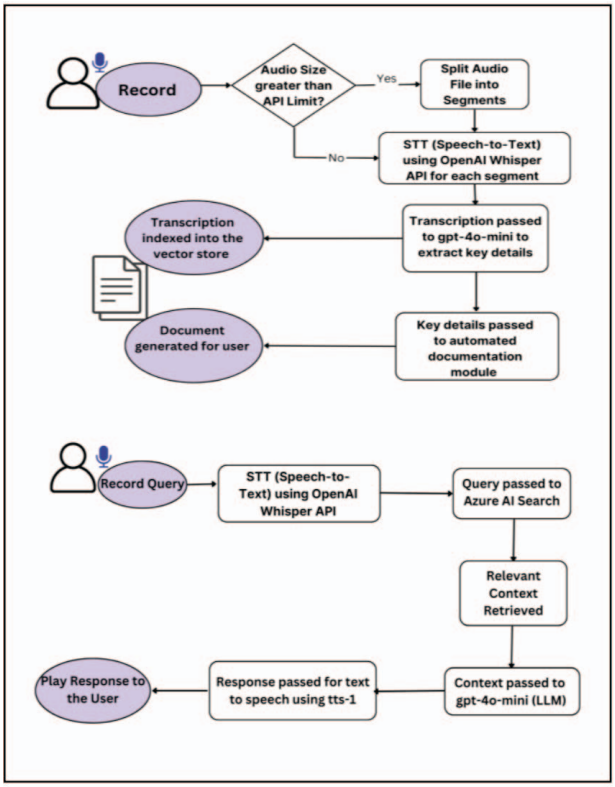

Fig. 1. Data Flow Diagram

Fig. 1. depicts the data flow diagram of Calix, illustrating its processes for recording and querying information. The top section details the audio recording flow, including segmenting large audio files, speech-to-text conversion via OpenAI Whisper API, indexing transcriptions into a vector store, and generating user documents with key detail extraction. The bottom section showcases the query process, starting from speech-to-text conversion, context retrieval using Azure AI Search,

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 05:10:03 UTC from IEEE Xplore. Restrictions apply.

response generation via gpt-4o-mini, and text-to-speech playback using tts-1

IV. Application of C.A.L.I.X in Healthcare: A Focused Approach

A. AI Revolutionizing Healthcare Documentation

The integration of Artificial Intelligence (AI) into healthcare is transforming clinical workflows by automating administrative tasks, reducing time burdens, and allowing physicians to focus more on patient care. Studies highlight that physicians spend an average of 15.5 hours per week on administrative tasks, including documentation, contributing significantly to physician burnout [5]. AI tools like C.A.L.I.X address this challenge by streamlining documentation processes while ensuring precision and efficiency.

B. How CALIX Enhances Healthcare Operations

a) Real-Time Interaction and Compliant Documentation: C.A.L.I.X actively listens during patient consultations, transcribing interactions into compliant documentation in real-time. This capability eliminates the need for physicians to take notes manually, allowing them to engage more effectively with their patients.

b) Efficient Data Retrieval: Doctors often face challenges recalling or locating specific patient records within vast databases [10]. With C.A.L.I.X, they no longer need to manually search or remember specific documents. Physicians can simply query C.A.L.I.X, and the system retrieves the relevant information instantly, saving time and reducing cognitive strain.

C. Benefits of C.A.L.I.X in Healthcare

a) Time Savings and Increased Efficiency: By automating tedious documentation processes, C.A.L.I.X aims at reducing hours for manual documentation by an average of ~6 hours for healthcare professionals. This allows doctors to allocate more time to patient care, which improves the quality of healthcare delivery and patient satisfaction.

b) Improved Accuracy and Compliance: With real-time transcription and automated storage in EHR systems, C.A.L.I.X ensures that documentation is both accurate and comprehensive. This reduces errors and enhances compliance with medical record-keeping standards.

D. Integration in Hospital Settings

a) Listening and Documentation: During consultations, C.A.L.I.X records and transcribes interactions into structured complaint documents. These documents are systematically stored in the hospital's EHR system, ensuring that all relevant data is captured without manual intervention.

b) Query-Based Retrieval: Physicians can interact with C.A.L.I.X via simple queries to retrieve patient information. This eliminates the need for time-consuming searches and allows for immediate access to critical data, streamlining clinical decision-making.

V. Performance Evaluation

C.A.L.I.X was tested upon a dataset of 15 different patient records. These patient records were generated after transcribing patient consultations. These consultations were generated by gpt-o1 while making sure they replicated actual consultations. The prompt template used to generate these consultations is as follows:

Write the transcript of a realistic medical consultation between a doctor which replicates a standard consultation. The consultation should include the name of the patient.
1. Name (Mandatory)
2. Age (return as a string) (optional)
3. Presenting Complaint (mandatory)
4. History of Present Illness (mandatory)
5. Past Medical History (optional)
6. Medication and Allergies (optional)
7. Family and Social History (optional)
8. Review of Systems - Assessment of other body systems (mandatory)
9. Physical Examination Findings (mandatory)
10. Assessment and Diagnosis (mandatory)
11. Treatment Plan (mandatory)
12. Advanced Decisions and Directives - Any advanced decisions to refuse treatment or directives discussed and agreed upon. (optional)

The transcripts must not include labels for Doctor, Patient for their conversation segments. It should be a plain transcription within one pair of inverted commas transcribed like the following:
"Good morning, Mr. Ahmed. How can I help you today? Good morning, Doctor. I've been having some chest pain on and off for the past week. Can you describe the pain for me? Is it sharp, dull, or something else? It's more of a tight, squeezing sensation. It usually lasts for about 10 minutes and sometimes happens when I'm walking or climbing stairs." Give me 15 different transcripts.

The model was tested for response time and accuracy of responses, with 50 test cases in total.

TABLE I. Response Time

Processing Time for Queries (seconds)		
Number of Tests	***Query Type***	***Average Time (seconds)***
10	*Direct Recall*	*9.467*
13	*Recognition-Based*	*6.965*
14	*Contextual Recall*	*8.128*
8	*Spatial Recall*	*6.116*
5	*Emotional Recall*	*8.25*
	Average Time	***7.785***

The response times were recorded for the amount of time between asking the query and generating the response. The response times for different types of recall tasks: Direct

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 05:10:03 UTC from IEEE Xplore. Restrictions apply.

Recall, Recognition-Based Recall, Contextual Recall, Spatial Recall, and Emotional Recall; were analyzed across five test iterations. The following observations were made:

1) Direct Recall: This query type exhibits the highest average processing time, indicating a higher computational demand for retrieving specific, unprompted information.

2) Recognition-Based Recall: The processing time for recognition-based tasks is comparatively lower, reflecting the system's efficiency in identifying the correct response from predefined options.

3) Contextual Recall: Contextual recall tasks require a moderate amount of processing time, as they depend on understanding the context or scenario tied to the query.

4) Spatial Recall: Spatial recall tasks demonstrate the lowest average processing time, suggesting efficient handling of spatial or locational queries.

5) Emotional Recall: Emotional recall tasks require similar processing time to contextual recall, likely due to the nuanced nature of retrieving emotionally tied information.

This analysis highlights the system's strengths and areas for potential improvement, providing a foundation for enhancing response efficiency in real-time applications.

TABLE II. QUESTION CATEGORIES

Distribution of Types of Questions		
	Regular	**Hallucination Inducive**
Direct Recall	5	5
Recognition-Based	5	9
Contextual Recall	5	8
Spacial Recall	5	3
Emotional Recall	5	0

TABLE III. ACCURACY

Accuracy for Different Query Categories					
	Accurate (A)	**Acknowledged Uncertainty (AU)**	**Corrective Clarification (CC)**	**Inaccurate (I)**	**Overall Success (A + CC)**
Direct Recall	50%	30%	10%	10%	60%
Recognition Based Recall	28.57%	50%	21.43%	0%	50%
Contextual Recall	30.76%	53.85%	15.38%	0%	46.14%
Spatial Recall	50%	37.5%	12.5%	0%	62.5%
Emotional Recall	100%	0%	0%	0%	100%

This section evaluates the system's ability to handle various query types, including both regular and hallucination-induced questions. Hallucination-induced questions are designed to test the model's capacity to detect and address inconsistencies. For example, Amir had a cough and was prescribed Fevadol and Basma had a headache was prescribed Panadol. The hallucination induced question would be "Which medicine was prescribed to Basma for her cough?"

While regular questions assess its ability to accurately retrieve information. The metrics analyzed include the proportion of accurate responses, acknowledged uncertainty, corrective clarifications, and inaccurate responses.

1) Model Response Categories

- Accurate: The model provides a correct response based on stored data. This occurs when the question is not hallucination-induced or when the model correctly handles a hallucination-induced question.
- Acknowledged Uncertainty: The model admits that it does not know the answer when it cannot provide an accurate response. Example: "I don't know."
- Corrective Clarification: The model corrects the user's query if it contains confusing or incorrect details that could lead to hallucination. Example: "Basma did not have a cough, she had a headache and was prescribed Panadol."
- Inaccurate: The model provides an incorrect response, which can happen with any type of question—hallucination-induced or regular.

2) Key Findings

Tasks like *Recognition-Based Recall* and *Contextual Recall* showed a significant reliance on *Acknowledged Uncertainty,* with 50% and 53.85% responses, respectively. This demonstrates the system's caution in ambiguous scenarios. This feature depicts success with efforts being concentrated on developing a system which followed Responsible AI. However, the system's ability to provide *Corrective Clarifications* is particularly important for hallucination-induced queries. While its use was limited (e.g., 21.43% for Recognition-Based Recall), it successfully avoided inaccuracies in all cases.

VI. FUTURE DEVELOPMENTS

1) Photos and Videos for Context: Incorporating photos and videos as part of C.A.L.I.X's memory storage will provide richer contextual data, complementing text and audio information [4]. By processing visual content, C.A.L.I.X can infer details like the environment, activities, and participants in an interaction. For example, integrating a photo of a medical report with a patient's consultation allows doctors to cross-reference notes with visual evidence. This feature will enhance C.A.L.I.X's ability to provide comprehensive and contextually enriched responses during retrieval, improving accuracy and decision-making.

*2) Queryless Mode:*The Queryless Mode allows C.A.L.I.X to proactively surface relevant information without requiring explicit prompts from users [2]. By analyzing ongoing tasks or consultations, C.A.L.I.X can predict the information a doctor might need and present it preemptively. For instance, during a follow-up consultation, C.A.L.I.X could automatically retrieve the patient's last visit

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 05:10:03 UTC from IEEE Xplore. Restrictions apply.

details. This feature will reduce cognitive load for doctors and further streamline workflows, ensuring timely access to critical data.

*3) Query Rewriting and Context Augmentation:*Query rewriting involves transforming user queries into more structured formats and enriching them with additional context for precise retrieval [4]. For example, a vague query like "What happened last week?" could be rewritten to include specific details such as time, location, and patient identifiers. This feature will make C.A.L.I.X more intuitive and capable of handling complex, unstructured user inputs while providing accurate and comprehensive results.

*4) Adding Atomic Details to Stored Informatio:*Atomic details include metadata like time, location, participants, and visual elements associated with a memory. Storing these granular details allows for more specific and targeted retrieval [4]. In a healthcare setting, this could mean linking a diagnosis with its time of recording, the doctor involved, and any associated scans or test results. This enhanced granularity will make C.A.L.I.X's retrieval more precise and contextually aware, reducing the need for manual data interpretation

5) Composite and Semantic Context Derivation: Composite context involves synthesizing atomic details into broader events or activities, while semantic context derives patterns or trends across multiple memories [4]. For instance, C.A.L.I.X could recognize a sequence of consultations as a treatment plan or infer a pattern of recurring symptoms. These higher-order insights will enable C.A.L.I.X to deliver more meaningful and actionable responses, helping doctors analyze patient history more effectively and make informed decisions.

Future developments for C.A.L.I.X will aim to enhance its ability to accurately recall and prioritize information while efficiently managing memory decay. One goal is to improve the algorithms that support context-sensitive data retrieval, ensuring that the most pertinent past interactions and details are easily accessible when needed. Continuous training of machine learning models will help the system focus on high-value memories, while less important information is gradually phased out to prevent overload and keep the knowledge base streamlined. Intelligent decay mechanisms will work to diminish the significance of outdated or unnecessary data, allowing C.A.L.I.X to adjust its recall strategies and retain only the most relevant information as user requirements change.

Moreover, there are promising prospects for extending C.A.L.I.X's functionality into sectors such as marketing and warehouse management. In the realm of marketing, the assistant could support professionals by managing detailed records of customer interactions, campaign performance, and market trends, thereby improving decision-making and customer engagement. Similarly, in warehouse management, C.A.L.I.X could transform inventory control and logistics through precise monitoring of stock levels, shipment timelines, and supply chain details, leading to enhanced operational efficiency and fewer errors. These innovations could significantly expand C.A.L.I.X's utility, making it a flexible tool adaptable for both personal and business applications.

VII. Conclusion

C.A.L.I.X (Cognitive Archive for Learning and Information Exchange) represents a significant advancement in the integration of memory augmentation systems with conversational AI, specifically tailored for the healthcare domain. By leveraging a hybrid Retrieve-Augment-Generate (RAG) system and adhering to Responsible AI principles, C.A.L.I.X provides accurate, contextually relevant, and verifiable responses while addressing critical challenges such as hallucination and data privacy. The system's innovative architecture incorporates a combination of semantic vector retrieval, advanced query processing, and real-time documentation, streamlining healthcare workflows and reducing cognitive strain on physicians. Through its design, C.A.L.I.X not only improves operational efficiency but also enhances patient-physician interactions by enabling healthcare professionals to focus on patient care instead of administrative tasks. Performance evaluations have demonstrated the system's capability to handle diverse query types with impressive accuracy and response times, setting a foundation for further refinement. Future developments, such as incorporating multimedia context, query-less operation, and enhanced memory decay mechanisms, promise to elevate the system's utility across broader domains. As conversational AI continues to evolve, systems like C.A.L.I.X highlight the transformative potential of domain-specific RAG implementations. By bridging the gap between human and AI memory capabilities, C.A.L.I.X underscores the role of responsible AI in shaping the future of healthcare and other critical sectors, paving the way for more efficient, empathetic, and context-aware AI solutions.

Acknowledgements

The authors would like to thank Dr.Anees Ara and Prof. Tanzila Saba immensely for their generosity in providing constructive feedback on the paper. This research is supported by The Artificial Intelligence & Data Analytics (AIDA) Lab, Prince Sultan University, Riyadh, Saudi Arabia.

References

[1] Indegene. (n.d.). Understanding the science behind learning retention [Online]. Available: https://www.indegene.com/what-we-think/reports/understanding-science-behind-learning-retention

[2] Katz, B., Lin, J., & Quick, R. (2023). Investigating the impact of neural language models on question answering. Massachusetts Institute of Technology. Retrieved October 31, 2024, from https://dspace.mit.edu/handle/1721.1/155181

[3] Schneegass, C., Wojcicki, Y., & Niforatos, E. (2021, May). Design for long-term memory augmentation in personal knowledge management applications. In 12th Augmented Human International Conference (pp. 1-5).

[4] Li, J. N., Zhang, Z. (J.), & Ma, J. (2024). OmniQuery: Contextually augmenting captured multimodal memory to enable personal question answering. Conference 17, July 2017, Washington, DC, USA. ACM

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 05:10:03 UTC from IEEE Xplore. Restrictions apply.

[5] National Jewish Health. (n.d.). Questions your doctor may ask. Retrieved January 26, 2025, from https://www.nationaljewish.org/patients-visitors/patient-info/prepare-for-your-appointment/questions-your-doctor-may-ask

[6] Microsoft, "Microsoft Responsible AI Standard: General Requirements," 2021. [Online]. Available: https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Responsible-AI-Standard-General-Requirements.pdf?culture=en-us&country=us

[7] Benteal, "Real-Time Management," 2021. [Online]. Available: https://www.benteal.com/real-time-management/

[8] N. Moer, "Retrieval," Cognitive Psychology, 2021. [Online]. Available: https://nmoer.pressbooks.pub/cognitivepsychology/chapter/retrieval/

[9] Y. Liu and A. Jones, "New Advancements in Machine Learning," arXiv preprint, arXiv:2406.03714, 2024. [Online]. Available: https://arxiv.org/abs/2406.03714

[10] AMA, "7 EHR Usability and Safety Challenges and How to Overcome Them," American Medical Association, 2021. [Online]. Available: https://www.ama-assn.org/practice-management/digital/7-ehr-usability-safety-challenges-and-how-overcome-them

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 05:10:03 UTC from IEEE Xplore. Restrictions apply.