CareTrail

A Multi-Agent Healthcare Memory & Retrieval System using Qdrant

Convolve 4.0 - A Pan-IIT AI/ML Hackathon

Track :

Multi-Agent Intelligent Systems (MAS) Track – Qdrant

Submitted by

Team Size: Individual
Team Members: Rahul Kumar
Institution: University of Engineering and Management, Jaipur
Email: kadvasararahul@gmail.com

Date: 22-01-2026

Abstract

CareTrail is a multi-agent AI system designed to address the challenge of fragmented healthcare information and the lack of continuity in AI-assisted healthcare support. In real-world settings, patients' medical reports, symptom descriptions, and prior clinical notes are often distributed across time and sources, resulting in loss of context, repeated interactions, and inconsistent guidance. Conventional AI assistants operate in a stateless manner and are therefore unable to retain or reason over long-term medical history, limiting their effectiveness in healthcare scenarios that require contextual awareness.

To overcome these limitations, CareTrail employs **Qdrant as a persistent vector database** to store and manage long-term healthcare memory, and **LLaMA-3.1 (via Groq Cloud)** for grounded and explainable reasoning. The system follows a **retrieval-first, multi-agent architecture**, where dedicated agents handle data ingestion, semantic retrieval, and reasoning as separate responsibilities. Medical information is embedded and stored in Qdrant, enabling semantic similarity search and memory reuse across interactions. When a user submits a query, only the most relevant historical context is retrieved and provided to the reasoning agent, ensuring that responses are strictly grounded in stored evidence.

This design significantly reduces hallucination risks and improves transparency by enforcing traceable, evidence-based outputs. CareTrail is intentionally positioned as an informational support system and does not provide medical diagnosis or treatment recommendations. This report presents a detailed description of the system architecture, multi-agent workflow, role of vector memory, retrieval strategy, ethical and safety considerations, and alignment with the evaluation criteria of the Qdrant Multi-Agent Intelligence Systems track.

1. Introduction

Healthcare interactions are inherently **longitudinal** in nature. Patients engage with multiple healthcare providers over extended periods, generating diverse forms of medical data such as laboratory reports, prescriptions, diagnostic summaries, symptom descriptions, and clinical notes. These data points accumulate over time and are essential for informed decision-making and continuity of care. However, in practice, this information is often fragmented across different systems, documents, or conversations, making it difficult to maintain a coherent and accessible medical history.

Recent advances in artificial intelligence have led to the adoption of conversational AI systems for healthcare-related assistance, including symptom tracking, report explanation, and general medical information. While these systems can provide useful short-term support, they typically operate in a **stateless** manner. Each interaction is treated independently, with no persistent memory of prior conversations or medical records. As a result, users are frequently required to repeat information, and the system may produce responses that lack historical context, leading to inconsistent or incomplete guidance.

The absence of long-term memory in AI assistants presents significant challenges in healthcare scenarios, where prior information is often crucial for understanding current conditions. Moreover, generating responses without access to relevant historical context increases the risk of hallucinated or misleading outputs, which can be particularly harmful in sensitive domains such as healthcare.

CareTrail addresses these challenges by demonstrating how **multi-agent AI systems combined with long-term vector memory** can enable continuity, transparency, and safety in healthcare assistance. By leveraging a persistent vector database for memory storage and separating ingestion, retrieval, and reasoning into specialized agents, CareTrail provides context-aware, evidence-based responses while maintaining clear boundaries between information support and medical decision-making. This approach highlights a scalable and responsible direction for building AI systems that interact with complex, evolving real-world data.

2. Problem Statement

The effective use of artificial intelligence in healthcare requires systems that can reason over information accumulated across time. However, most existing AI assistants are designed as stateless systems, treating each interaction as an isolated event. This design choice introduces several fundamental challenges when applied to healthcare-related use cases.

First, there is a **loss of historical medical context**. Medical data such as laboratory results, symptom progression, and prior clinical notes are intrinsically temporal and interdependent. When AI systems fail to retain this historical context, they are unable to provide responses that reflect the full medical narrative of a patient.

Second, current AI assistants exhibit an **inability to remember prior interactions**. Users are often required to repeatedly provide the same information across multiple sessions, leading to inefficiency and frustration. More importantly, the lack of memory prevents the system from identifying patterns or changes in health conditions over time.

Third, the **risk of hallucinated responses** is significantly higher in stateless systems. Without access to verified historical data, AI models may generate responses based on incomplete or assumed context. In healthcare settings, such hallucinations can result in misleading or potentially harmful information.

Finally, there is a **lack of traceability in AI-generated outputs**. Many AI systems do not provide clear explanations of which data influenced a particular response. This opacity makes it difficult for users to assess the reliability of the output and undermines trust in the system.

Due to these limitations, stateless AI systems are fundamentally unsuitable for real-world healthcare continuity, where access to past data, transparency of reasoning, and reliability of outputs are essential for meaningful and responsible assistance.

3. Proposed Solution

CareTrail proposes a **multi-agent architecture with shared long-term memory** to address the limitations of stateless AI systems in healthcare contexts. The central idea is to decouple memory, retrieval, and reasoning into specialized agents that interact through a persistent vector database. This design enables continuity of context, transparency, and safer AI behavior.

At the core of the system, **medical information is embedded and stored in Qdrant** as vector representations. Inputs such as medical reports, symptom descriptions, and user notes are processed by a dedicated ingestion agent, which converts unstructured text into semantic embeddings and stores them along with relevant metadata (e.g., type of record and timestamp). This allows the system to retain healthcare information beyond a single interaction or session.

When a user submits a query, **relevant historical context is retrieved using semantic similarity search**. Instead of relying on keyword matching, the retrieval agent embeds the user query and performs similarity search over stored vectors in Qdrant. This ensures that the retrieved information is contextually relevant, even when the wording of the query differs from the original stored data.

Crucially, **reasoning is performed only over retrieved evidence**. The reasoning agent, powered by a large language model, receives only the subset of historical data returned by the retrieval agent. It does not rely on external knowledge or assumptions. This retrieval-first approach enforces evidence grounding, significantly reducing hallucination risks and improving the reliability of responses.

This architectural design ensures several key properties. First, it provides **persistent memory beyond a single prompt**, enabling the system to recall and reuse relevant information across interactions. Second, it produces **evidence-grounded responses**, as all outputs are derived from stored and retrieved data rather than inferred context. Finally, it enables a **clear separation of responsibilities across agents**, improving system interpretability, modularity, and scalability.

Through this approach, CareTrail demonstrates how multi-agent systems combined with vector memory can deliver context-aware, transparent, and responsible AI assistance in healthcare-related applications.

4. System Architecture

CareTrail follows a **retrieval-first, multi-agent architecture** in which multiple specialized agents collaborate through a shared long-term memory implemented using Qdrant. The system is designed to ensure persistence of information, clear separation of responsibilities, and evidence-based reasoning.Figure 4.1 illustrates the overall architecture and information flow of the system.

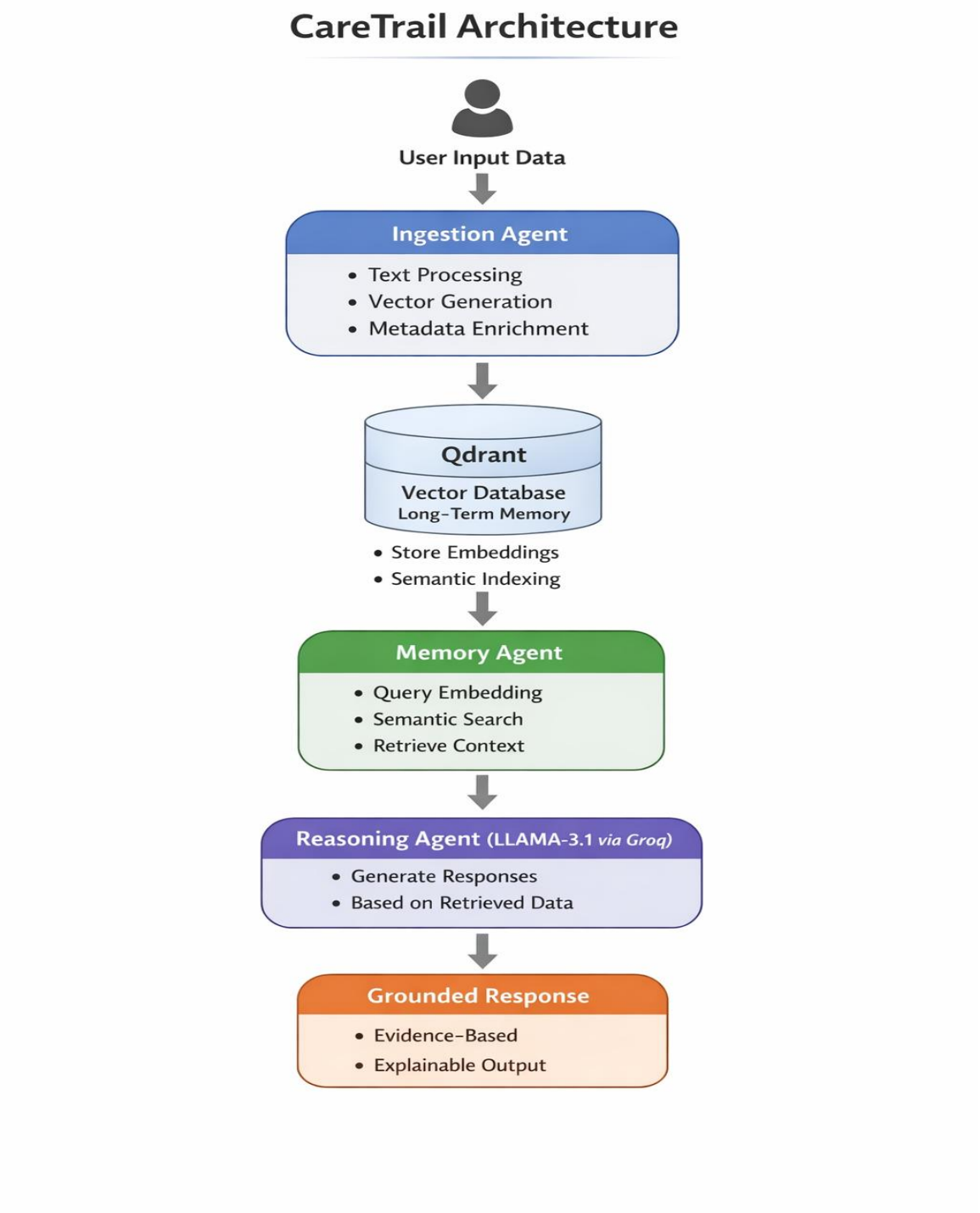

Figure 4.1: CareTrail System Architecture Flowchart

The architecture consists of three core agents: the **Ingestion Agent**, the **Memory (Retrieval) Agent**, and the **Reasoning Agent**, coordinated through **Qdrant**, which serves as the shared long-term memory layer enabling persistent contextual awareness.

4.1. Ingestion Agent

The **Ingestion Agent** processes and stores incoming medical information such as reports, symptoms, and notes. Its key responsibilities include:

- Accepting unstructured medical text
- Splitting inputs into semantic chunks
- Generating vector embeddings with metadata
- Storing embeddings in Qdrant

4.2. Memory (Retrieval) Agent

The **Memory (Retrieval) Agent** retrieves relevant historical information when a user submits a query. Its primary functions are:

- Converting user queries into embeddings
- Performing semantic similarity search in Qdrant
- Retrieving the most relevant stored records

This agent enforces a strict **retrieval-before-reasoning** pipeline.

4.3. Reasoning Agent

The **Reasoning Agent**, powered by **LLaMA-3.1 via Groq Cloud**, generates user-facing responses. It operates under the following constraints:

- Uses only retrieved memory as context
- Performs reasoning exclusively on retrieved evidence
- Produces grounded and explainable outputs

4.4. End-to-End Flow Summary

The system follows this sequence:

1. User provides medical data or submits a query
2. Data is ingested and stored in Qdrant
3. Relevant history is retrieved
4. Evidence-based reasoning is performed
5. A grounded response is returned

5. Technology Stack

CareTrail is implemented using a carefully selected set of technologies that support long-term memory, semantic retrieval, and evidence-based reasoning. Qdrant is used as the core vector database to store and retrieve healthcare information across interactions, while LLaMA-3.1 accessed via Groq Cloud performs reasoning strictly over retrieved context. Sentence Transformers (MiniLM) generate dense embeddings for both medical data and user queries, enabling effective semantic similarity search. The overall system logic is implemented in Python with a command-line interface to support interactive ingestion and querying, and a local Docker-based setup is used to ensure reproducibility and ease of deployment, as shown in Table 5.1.

Table 5.1: Technology Stack Used in CareTrail

Component	Technology Used	Description
Vector Database	Qdrant	Stores long-term healthcare memory as vector embeddings and enables semantic retrieval
Large Language Model	LLaMA-3.1 (via Groq Cloud)	Performs grounded reasoning using retrieved evidence only
Embedding Model	Sentence Transformers (MiniLM)	Converts medical text and queries into dense vector representations
Backend	Python	Implements multi-agent logic, retrieval pipeline, and system orchestration
User Interface	Command Line Interface (CLI)	Provides simple interactive input for ingestion and querying
Deployment	Docker (Local Setup)	Runs Qdrant locally for reproducible and lightweight deployment

6. Limitations & Future Work

10.1. Limitations

Despite its effectiveness as a proof-of-concept system, CareTrail has several limitations. First, the system has **no real clinical validation**, as it has not been tested with healthcare professionals or real patient data. Consequently, its outputs should be interpreted strictly as informational support rather than medically authoritative guidance.

Second, the current implementation is **limited to textual medical data**. While this allows effective handling of reports and symptom descriptions, it does not yet support medical images, scanned documents, or other multimodal data commonly used in healthcare settings.

Finally, the system currently includes **no authentication or access control mechanisms**. This limits its suitability for real-world deployment, where secure user identification and access management are essential to protect sensitive healthcare information.

10.2. Future Work

Several extensions can be explored to enhance the system's capabilities. One important direction is **multimodal ingestion**, enabling the processing of medical images, scanned reports, or other non-textual data using appropriate embedding models.

Another area of future work is the implementation of **secure data storage and encryption** to protect sensitive medical information and ensure compliance with data privacy requirements. Additionally, **user-specific memory segmentation** can be introduced to prevent cross-user data leakage and to support personalized healthcare memory.

Finally, integrating CareTrail with **existing healthcare systems or electronic health record (EHR) platforms** could improve data consistency and practical usability, enabling seamless continuity of care across real-world clinical workflows.

7. Conclusion

CareTrail demonstrates how **multi-agent AI systems combined with vector databases** can effectively address real-world societal challenges, particularly in the domain of healthcare information continuity. By integrating **Qdrant's semantic retrieval capabilities** with **LLaMA-based reasoning**, the system enables long-term memory, contextual awareness, and evidence-based response generation across interactions.

The retrieval-first, multi-agent architecture ensures that reasoning is grounded in stored information, improving transparency and reducing the risk of hallucinated outputs. Through clear separation of responsibilities across agents, CareTrail also highlights the importance of modularity and explainability in the design of responsible AI systems.

Overall, this project illustrates how persistent memory, semantic retrieval, and controlled reasoning can be combined to build AI assistants that are safer, more reliable, and better suited for real-world applications where context and continuity are essential.