Team #1 - MSBA - CARLSON SCHOOL OF MANAGEMENT

Clinical Note Intelligence with RAG

Helping Clinicians Find Relevant Answers Faster from Unstructured Notes

The Challenge

- Clinical notes are stored as unstructured text - long, fragmented, and inconsistent
- Clinicians cannot effectively search or compare notes across patients
- Critical questions (e.g., treatment efficacy, drug-symptom patterns) are slow to answer

Our Solution

- Retrieval-Augmented Generation (RAG) system combining structured and semantic search with LLM generation
- FAISS vector database retrieves the most relevant note segments using semantic similarity
- GPT-4 generates grounded, cited answers delivered via a Streamlit web interface

KEY FEATURES

Semantic Retrieval

Encodes each note chunk into vector embeddings; FAISS finds the top-k most clinically relevant segments and retrieves parent documents to preserve context

Grounded Generation

GPT-4 synthesizes answers only from retrieved sources. Every response includes note-level citations and confidence indicators to ensure transparency.

Population-Scale Insights

Surfaces treatment trends and drug-symptom relationships across large patient cohorts, enabling evidence-backed clinical decision-making at scale.

SYSTEM PIPELINE

Data Processing

Clinical notes chunked by section type (Assessment, Plan, etc.). Patient ID and admission metadata preserved for traceability and audit.

Vector Retrieval

Chunks encoded as dense embeddings. FAISS vector database performs semantic similarity search, returning the top-k relevant note segments.

GPT-4 Generation

Retrieved segments passed to GPT-4. Generates structured, cited responses grounded in source data. Answers are never produced without retrieved evidence.

SAMPLE INTERACTION

USER QUERY

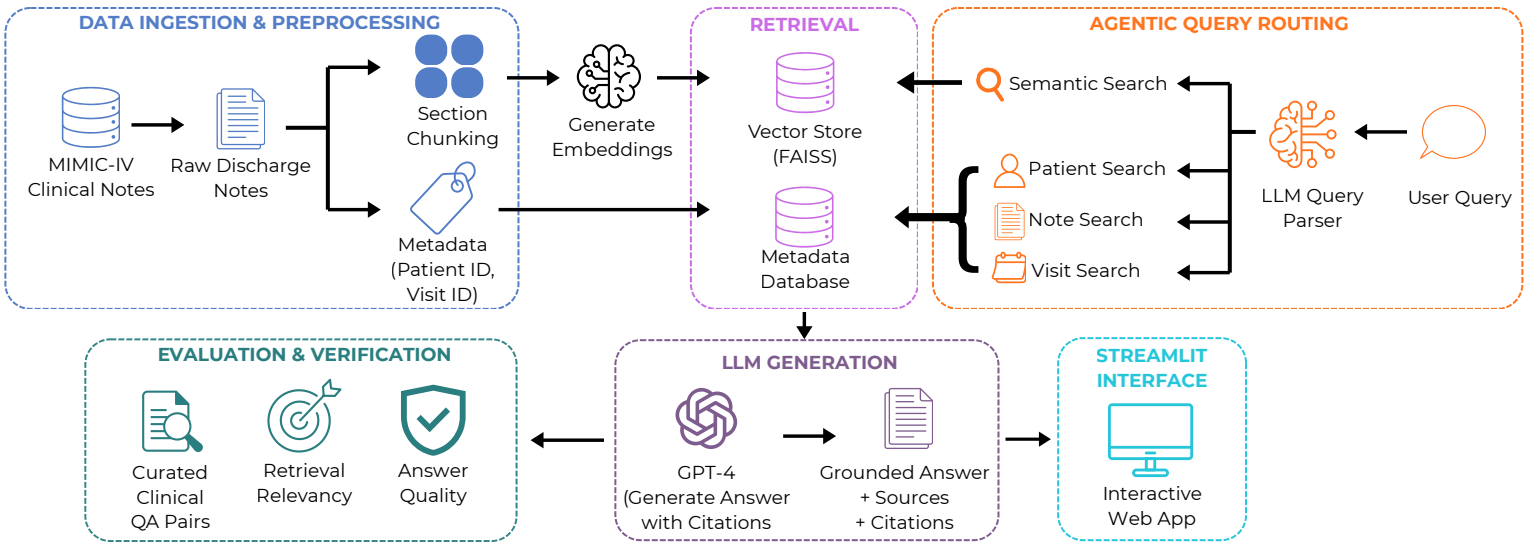
"What treatments worked for elderly heart failure patients?"

AI-GENERATED INSIGHT

"Diuretics and beta-blockers most common across 12 matched notes."

Sources: Note IDs 123, 456 ... | Confidence: High

METHODOLOGY



Standard RAG vs. Our RAG

Capability	Standard RAG	Our System
Query Handling	Embedding only	Intent-aware routing (using agentic AI)
Retrieval	Single-step searching	Structured + semantic search
Context	Uses chunks only	Chunk-to-document expansion
Grounding	Can hallucinate	Answers grounded in real clinical note evidence
Citations	Sometimes included	Always included
Confidence	None	Confidence from supporting notes

RESULTS & IMPACT

Faster Chart Review

Reduces hours of manual note review and retrieval into seconds.

Grounded Answers

Source citations with every response - reduces unsupported responses.

Scalable Analysis

Operates across large-scale longitudinal discharge-note datasets

Reduced Hallucination

Answers strictly bounded by retrieved evidence from patient clinical notes

TEAM 1


Ethan Armstrong
armst813@umn.edu


Ziqi Cao
cao00520@umn.edu


Ko-Jung Hsu
hsu00206@umn.edu


Cole Johnson
joh21802@umn.edu


Mashhood Khan
khan1200@umn.edu


Wenyu Zhong
zhong577@umn.edu