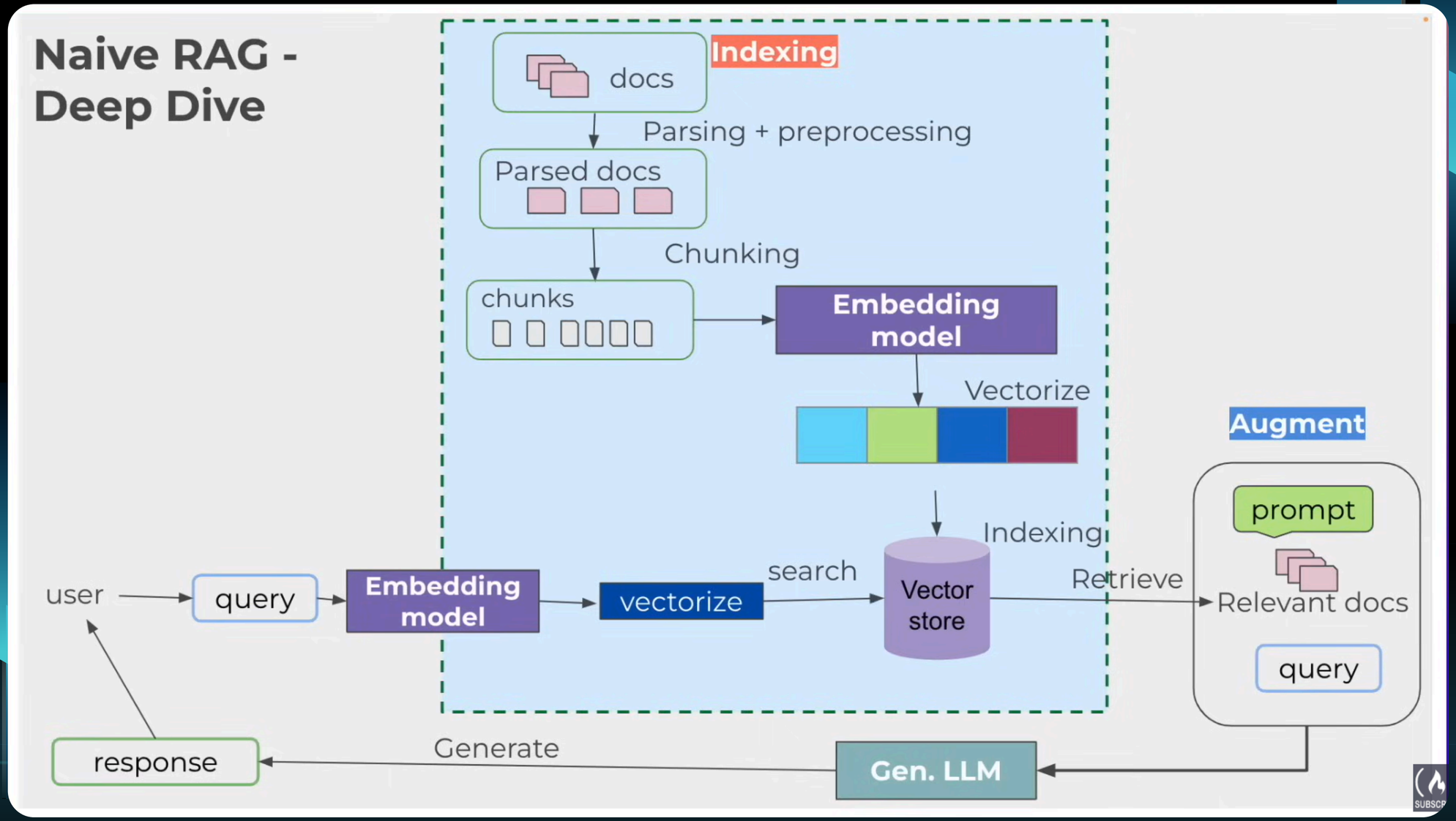
AI Club Mini Project

Rag Based LLMs

What is RAG?


Naive RAG - Deep Dive
Indexing
docs
Parsing + preprocessing
Parsed docs
Chunking
chunks
Embedding model
Vectorize
Augment
prompt
Indexing
search
Retrieve
user
query
Embedding model
vectorize
Vector store
Relevant docs
query
response
Generate
Gen. LLM

GLoVe Embedding

To perform any task with a model, we need to encode input as numbers.
In NLP, we need word embedding.
It helps make vector representation of words to capture their semantic relationships and contextual information. So vectors closely related in meaning or adjacent in common phrases are closer to each other.
GLoVe is a global **log bilinear regression model** that combines **global matrix factorization** and **local context window methods.** It formally defines the mathematical structure (such as co-occurrence probabilities) that leads to meaningful semantic similarities in word embeddings.

$$w_i^T \cdot \tilde{w}_j + b_i + \tilde{b}_j \approx \log(X_{ij})$$

BM25

BM25 is a ranking algorithm and is an extension of TFIDF
It gives a better estimate of relative document relevance especially when TF is high by modifying the below linearity

$$\frac{\partial(TFIDF)}{\partial(TF)} = IDF$$

We modify TF and IDF and estimate score as

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

$$\text{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

Sentiment Analysis Task

We uses a RNN based model to perform classification of SMS messages into Spam or not Spam

- First, we tokenise the sentences in the dataset.
- Then, we embed the sentences by concatenating the GLoVe embeddings of the tokens in the sentence.
- We convert them to pytorch tensors to send into the RNN in sequential order.
- The output of the RNN is a probability of it being spam.

PROJECTS

Rules Quiz Assist for Comp-teams

- My project reads the given set of PDFs, and uses libraries like langchain._google_genai to perform vector Embedding (GoogleGenerativeAIEmbeddings)
- I have used Facebook AI Similarity Search (faiss) library to identify the closest similar chunks (through Euclidian distance) to be fed in as context
- The contex is fed into Gemini 2.5
- The prompt is initiated with a pre defined start phrase and is continued by the user’s query

Dune Wiki ChatBot

1. We extract all the links from the web, using a simple web crawling algorithm.
2. We scrape the useful data from the URLs, and split into chunks.
3. These chunks are then sent to an embedding model (all-MiniLM-L6-v2) which transforms sentences to vectors.
4. The user inputs a query. The query itself is also transform to a vector.
5. Using FAISS (Similarity Search) algorithm, we find the top k vectors closest to our query vector.
6. The data corresponding to these vectors becomes our “context” which is then fed into an LLM (Gemini 1.5 Pro) along with our query.

FINANCIAL CHATBOT

1. We start with financial data exported from Screener.in in Excel format, with each sheet representing a different company.

2. From these sheets, we extract all key financial tables:
   Ratios, Quarterly Results, Balance Sheet, P&L, Cash Flow, and mini-tables.

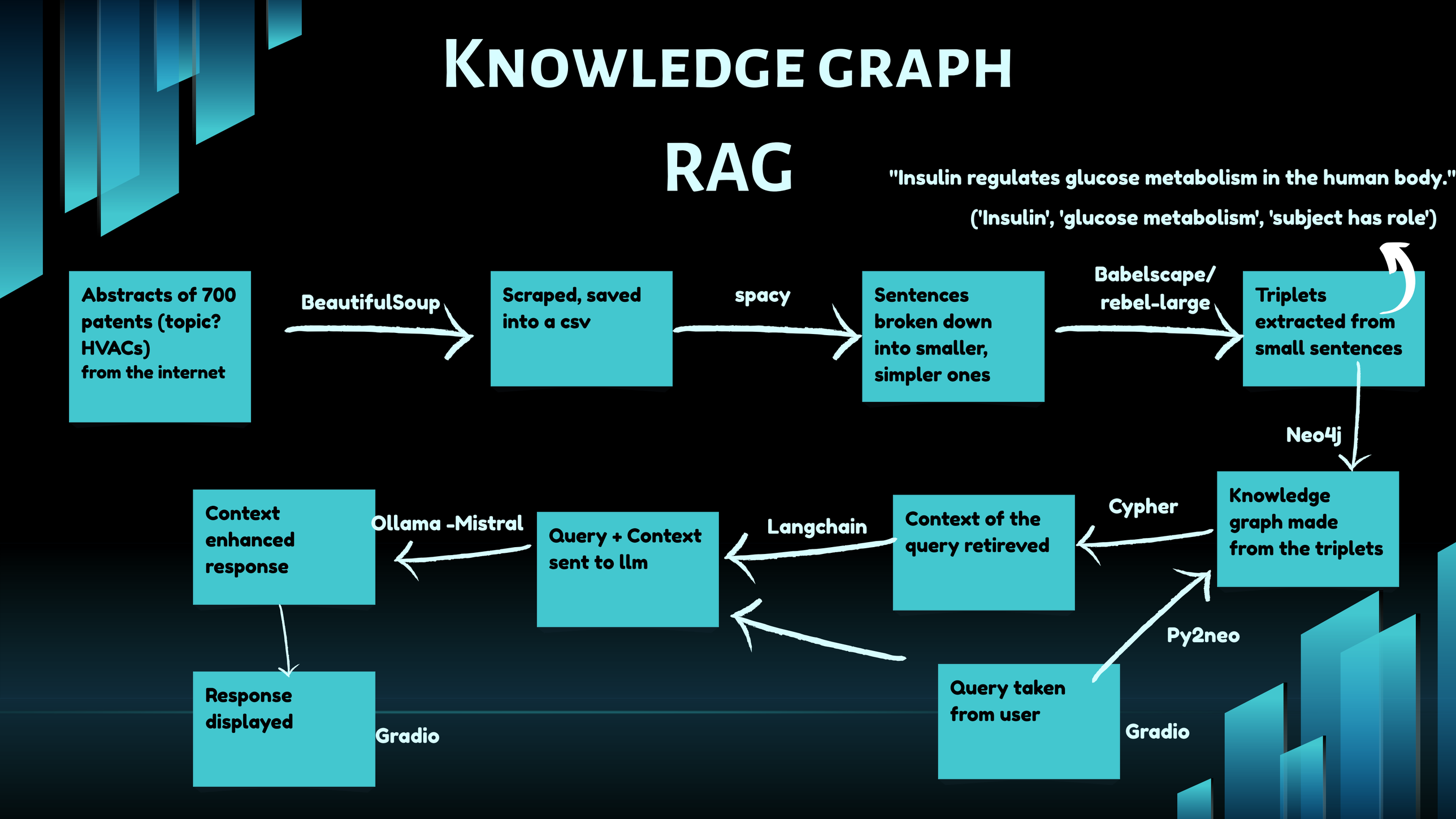
3. Each table is converted into clean text-based “chunks” and passed through a financial-domain embedding model (Fin-MPNET-Base).

4. When the user enters a query, it is first passed to a local LLM (Mistral via Ollama) which intelligently splits it into one or more subqueries
based on the structure of the financial data.

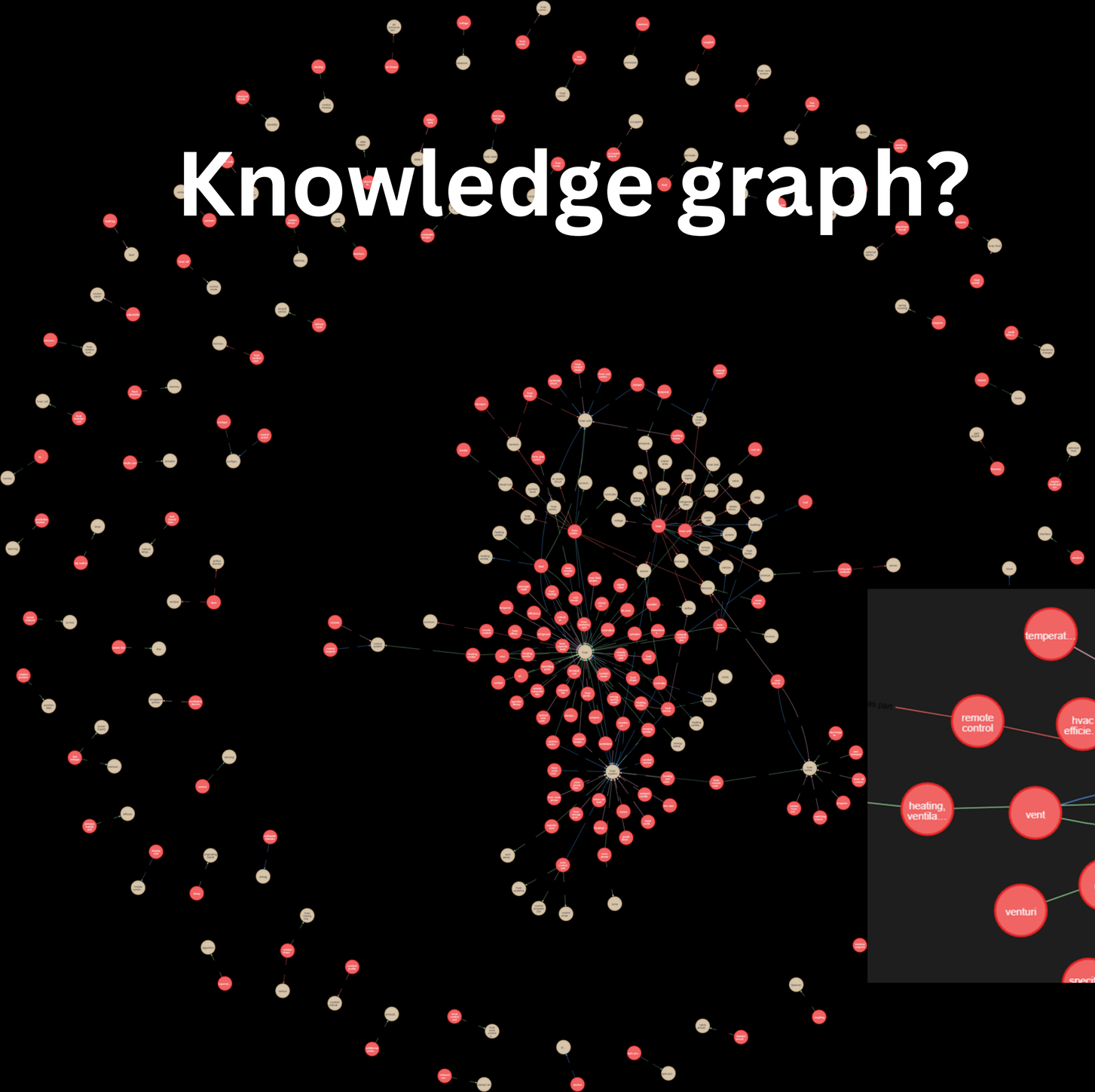
5. Each subquery is embedded into a vector, and Qdrant (vector database) is used
to retrieve the top-k most relevant chunks for each.

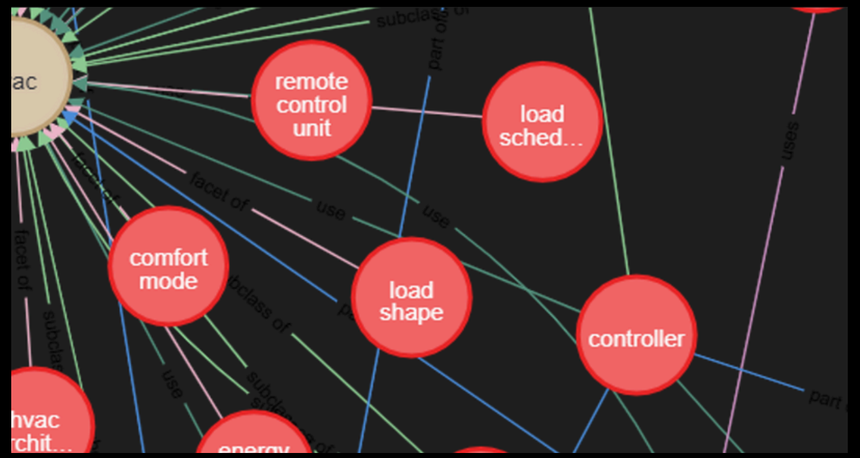
6. All retrieved chunks are merged to form the “context,” which, along with the original user query, is fed into the LLM to generate a grounded answer.

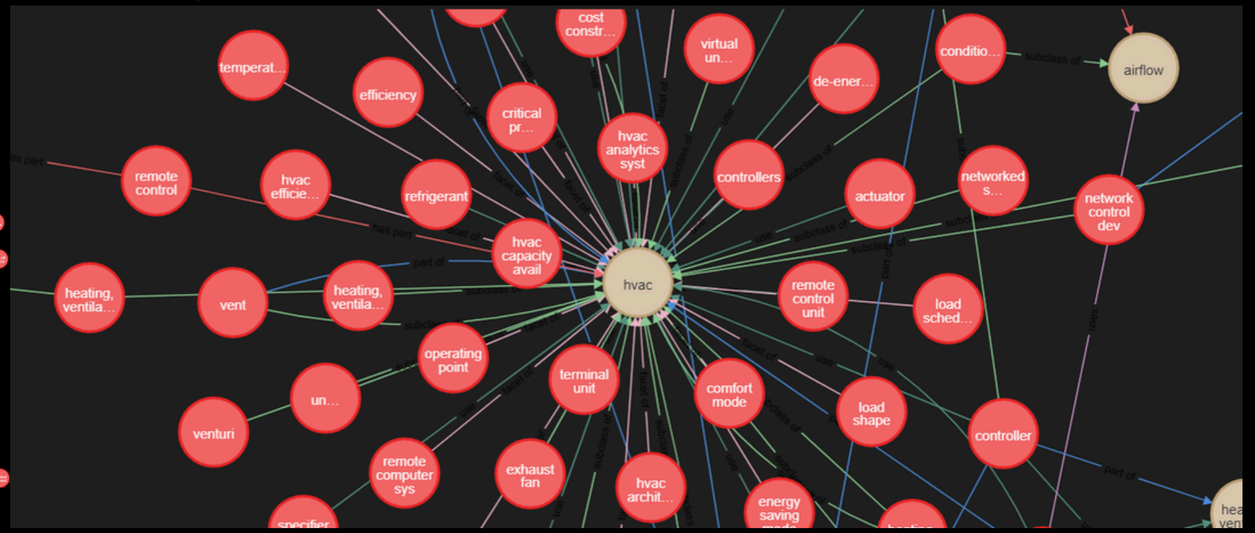
Knowledge graph RAG



Knowledge graph?







Neo4j + Ollama RAG Chatbot

Ask any question. The bot searches your Neo4j knowledge graph for relevant facts, then generates an answer using a local CPU-friendly LLM.

Ask a question about your knowledge graph

air temperature control

Clear

Submit

Answer with context

🤖 Answer: The term "air temperature control" can be found in the context provided as a specific type, which is called "discharge air temperature control." This term falls under the broader category of HVAC (Heating, Ventilating, and Air Conditioning) controllers. Therefore, air temperature control refers to any mechanism designed to regulate or manage the temperature of the air being distributed within an HVAC system.

📚 Context Used:
discharge air temperature control --subclass of--> hvac controller
discharge air temperature control --subclass of--> hvac controller

Thank You

GitHub Repos

- Aishwarya - https://github.com/Aishwarya-926/KG_RAG_HVAC_AI_Club_DC_Project
- Arnav -
- Avisha - https://github.com/velcroapple/Dune_Rag_ChatBot
- Hemanth - https://github.com/Hemanth-2706/My-Projects/tree/main/1)%20Gemini%20Chatbot
- Manish -
- Soham -