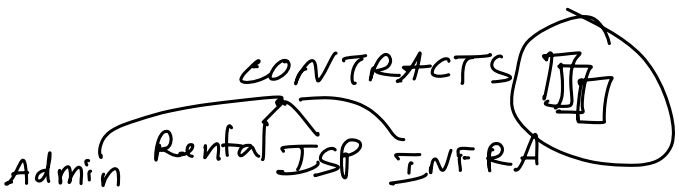
RAG (Retrieval-Augmented Generation)



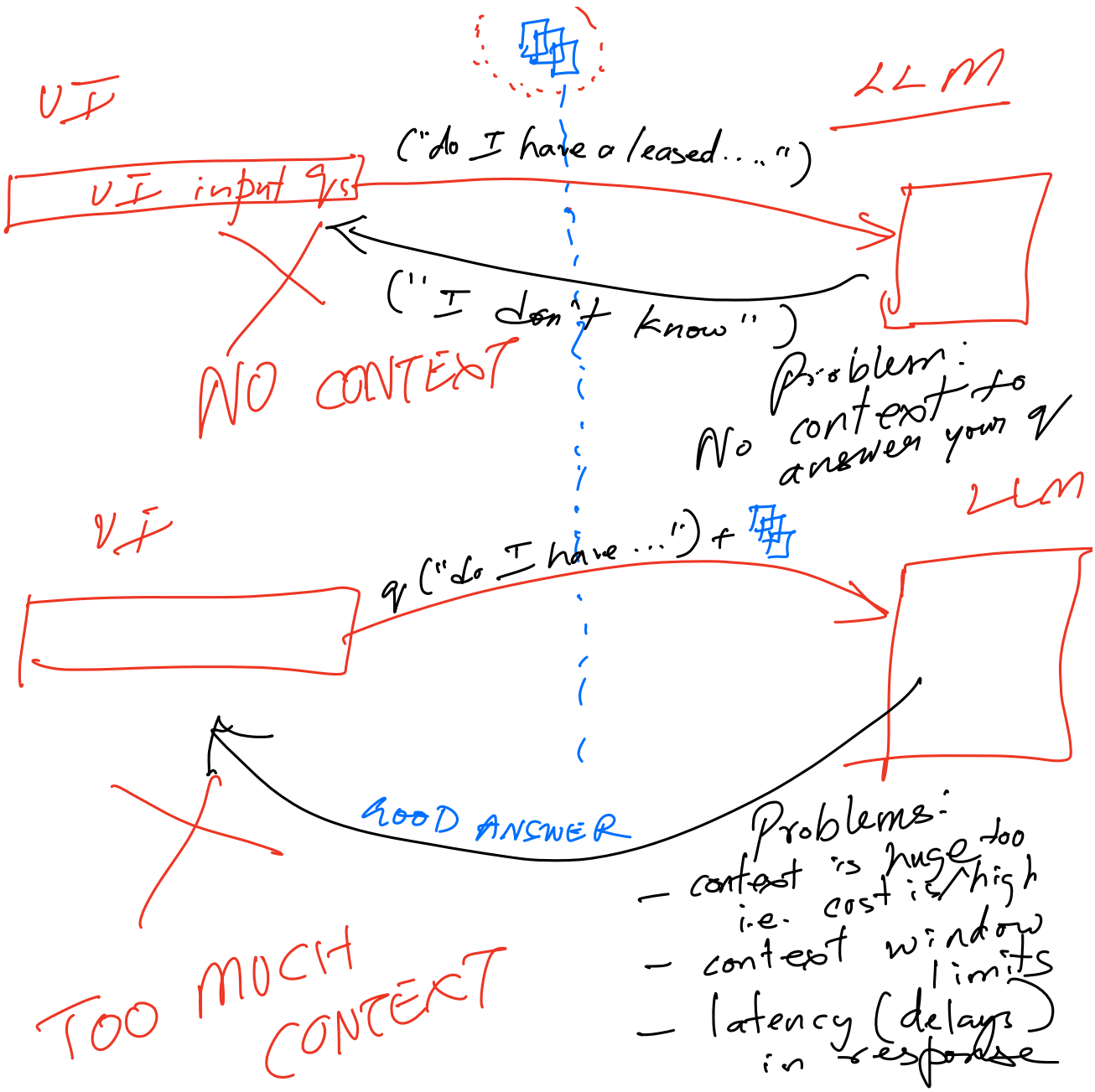
Implement a UI/chat i/f that allows for questions to be asked about the contracts.

For e.g. q: when does my lease on the Road No. 10 property end?

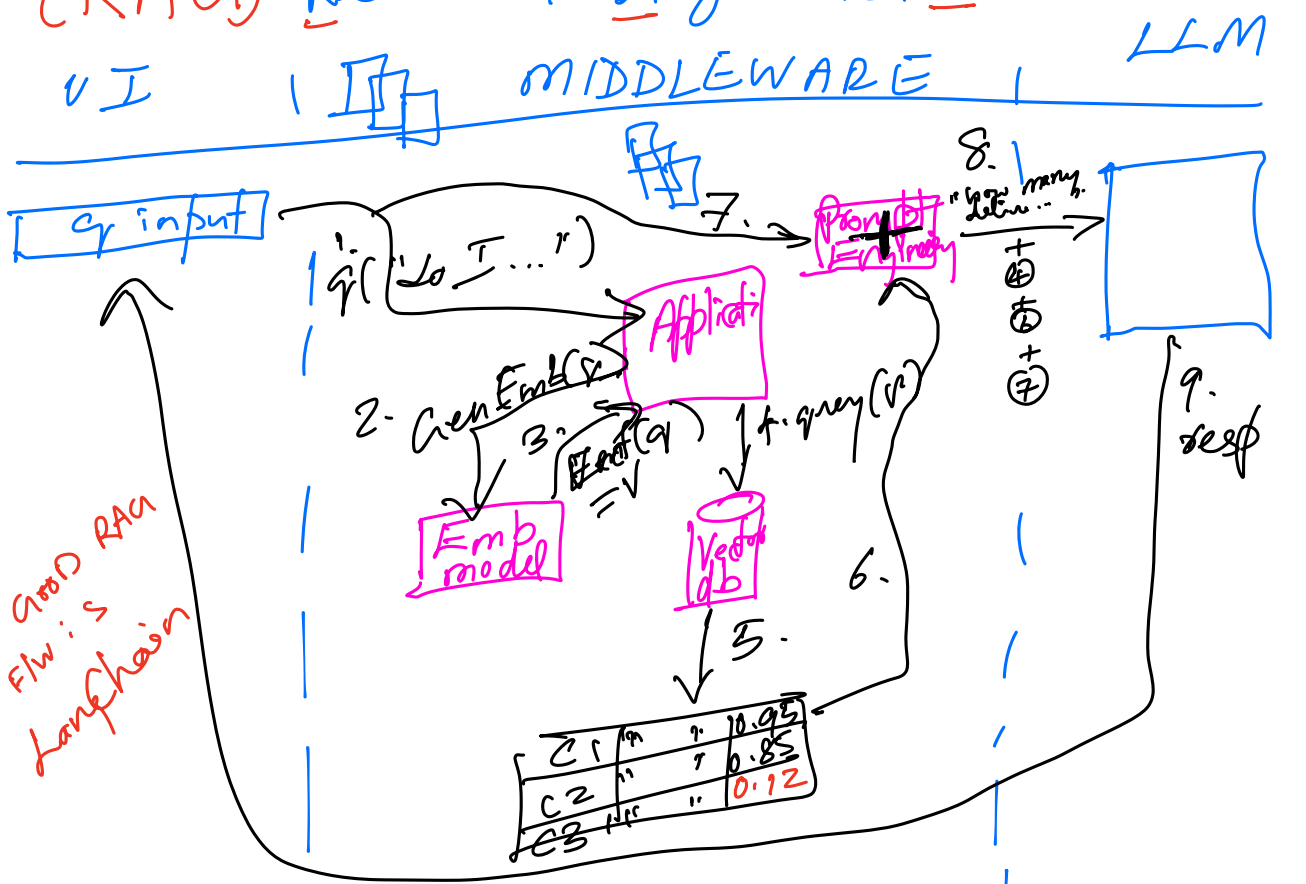
q: do I have a leased line in my JIO ISP service

q: what is the penalty for ending our catering contract?

q: summarize my maintenance contract of my sector-7 lease


UI
LLM
UI input q's
("do I have a leased...")
("I don't know")
NO CONTEXT
Problem:
No context to answer your q
UI
LLM
q ("do I have ...") +
GOOD ANSWER
Problems:
- context is huge too i.e. cost is high
- context window limits
- latency (delays) in response
TOO MUCH CONTEXT

(RAG) Retrieval-Augmented Generation



ONE-TIME PREP

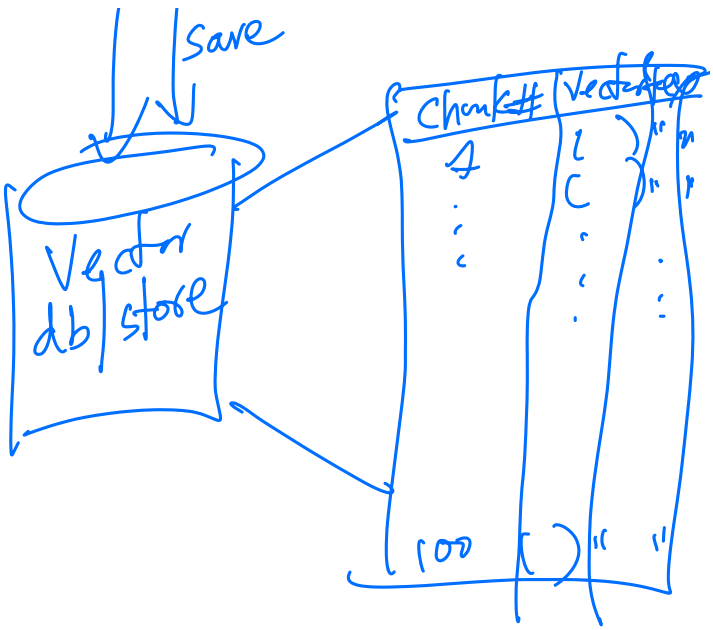
(1000 docs x 10,000 chars)

chunk it up into 1000 characters/chunk

(10 chunks x 10 docs = 100 chunks)

vectorize (gen embeddings)

Emb model lib



CONTROLS WITH YOU

- PROMPT ENGINEERING
- CHUNK SIZE
- RELEVANCE SCORE
- VECTOR DB SEARCH METHOD

Prompting

Retrieval

FILTERING

COSINE vs EUCLIDEAN

Q: are any of my contracts with a Maharashtra-based company?

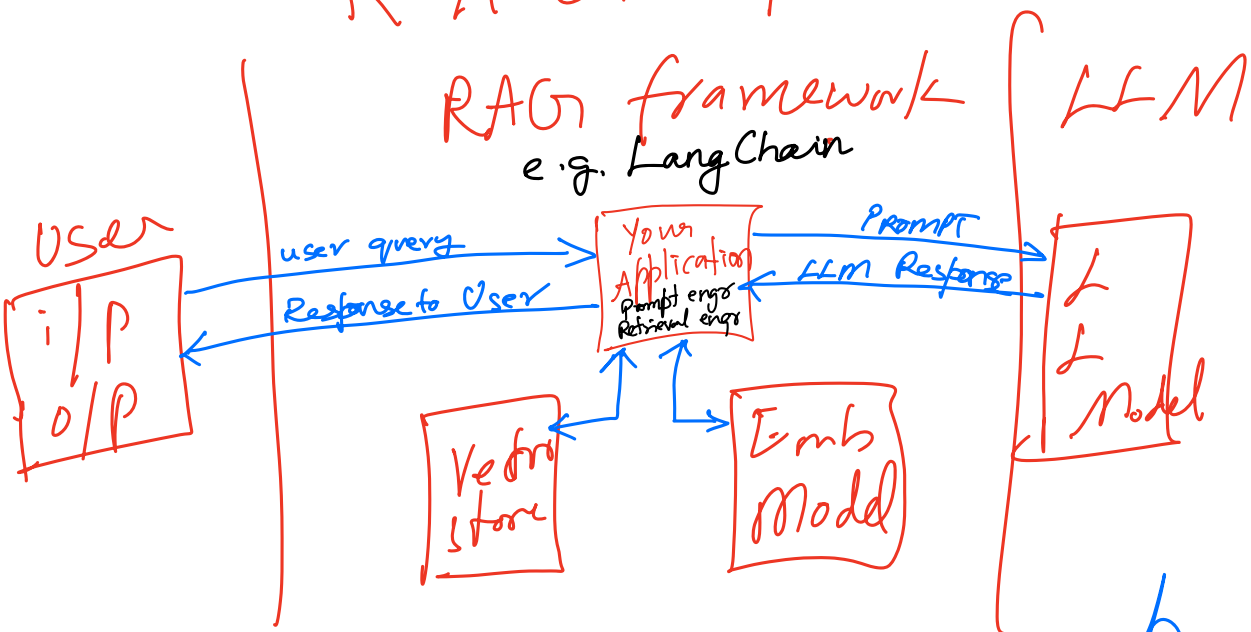
RAG m/w

UI

"are any of my ... Maharashtra..."
(use the below as context)
"82, ... New Delhi"
"75, ..., Mumbai"
"Catering Contract – ABC Holdings
37, ... Road, Pune, ..."

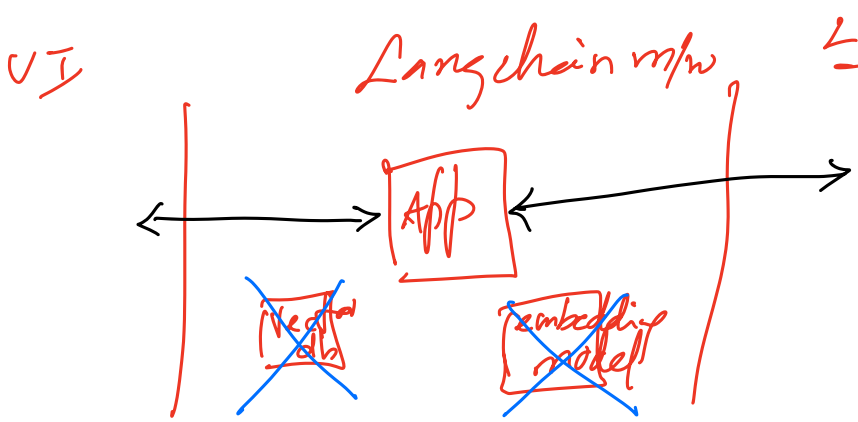
LLM

RAG m/ware

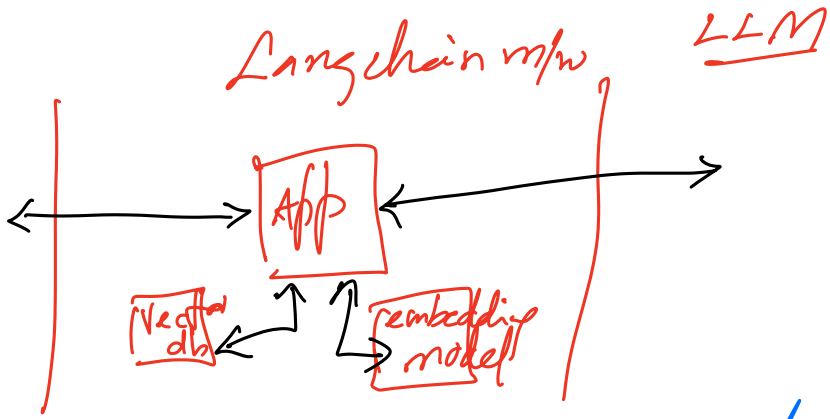


demos coming up

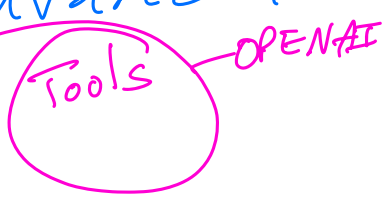
1) Embeddings demo (nothing to do with Langchain)
2) Langchain demo (direct with LLM)



3) Langchain RAV end-to-end demo



4) Lang Chain Advanced



1. Reqget_emp(,)

2. Get Projects(,)