Capstone Project Review

TONY WANG
PS&S

REALLY GREAT COMPANY

Contents

Overview

1

built the system that processes the timesheet data (convert into embedding space using mxbai-embed-large)

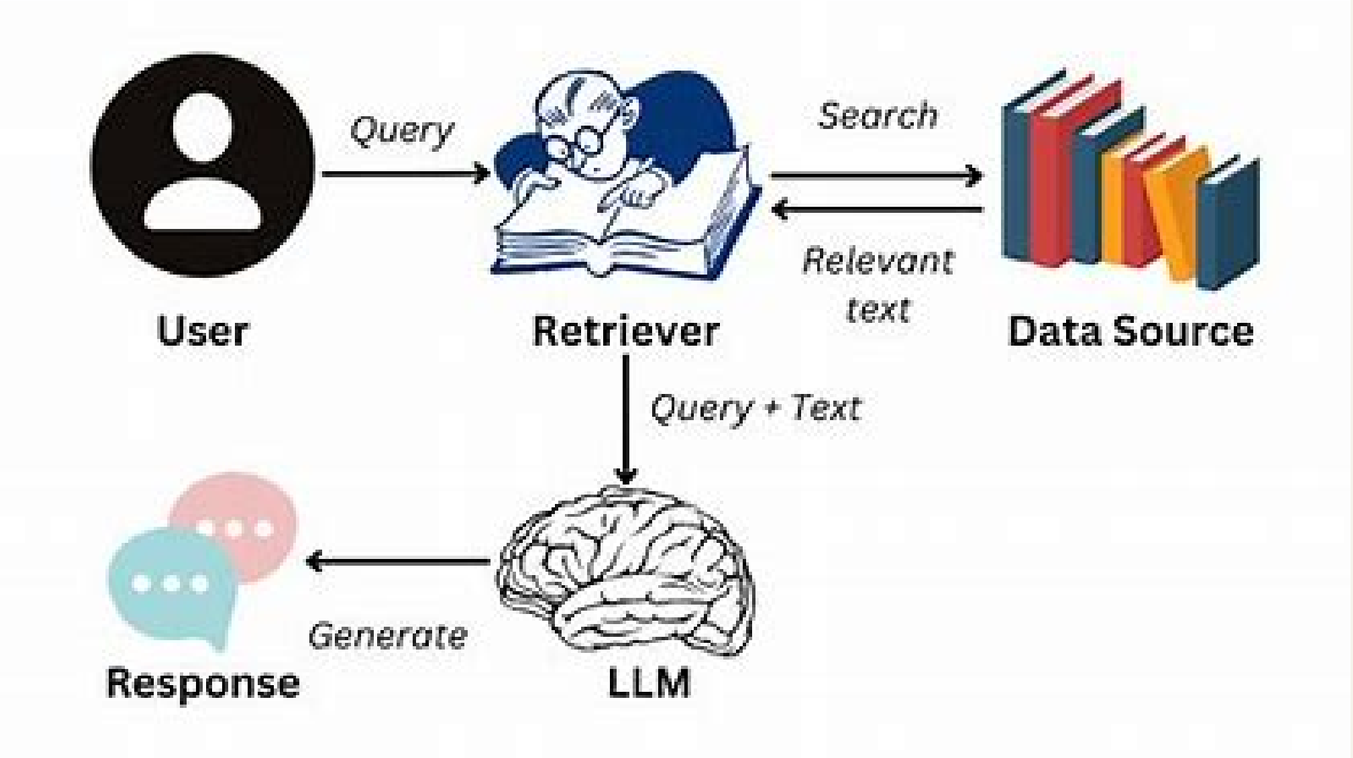
2

store embeddings in ChromaDB

3

retrieve relevant projects based on user queries and generates response using llama3.2

Basic RAG system



- given certain query, find top k most relevant projects
- eg: “background”
- Timesheet Comment: background research
  Project Code : 089900001
  Project Name : Cedar Holdings-Site Plan/Permitting”

REALLY GREAT COMPANY

Potential Issues

Bad Query

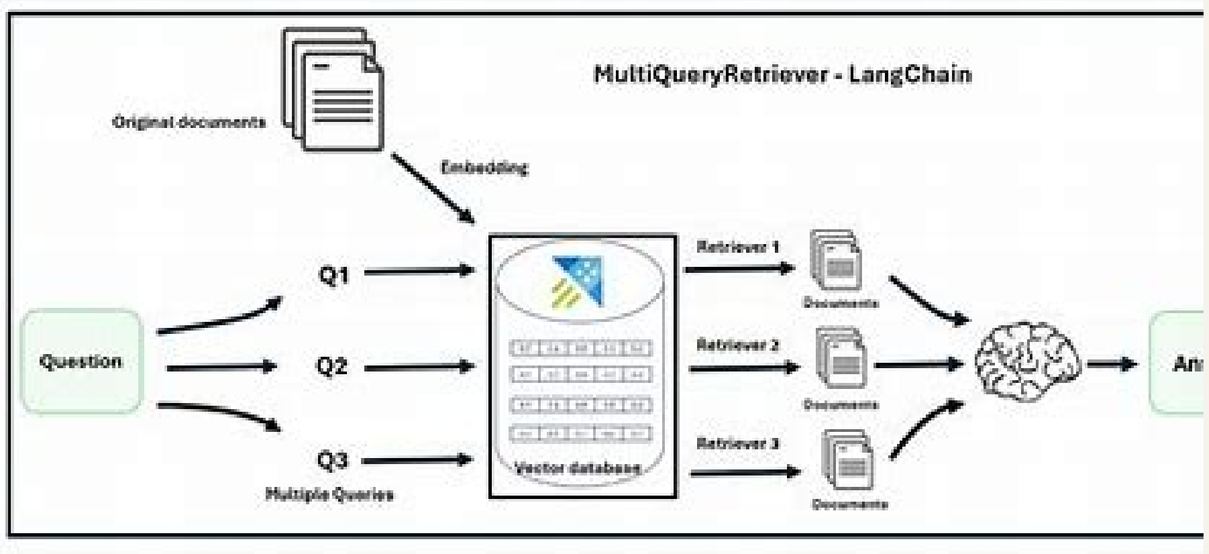
- Vague (sinigle word)
- Abbreviated (info as information)
- Typo (bacgoun as background)

Solutions- Query Refinement

Phase 1
Expand the query from different perspective

Phase 2
Reasoning and rewrite

Phase 3
Put refined query into embedding sapce



REALLY GREAT COMPANY

ChromaDB Fix

- Avoid duplicate embeddings when inserting

REALLY GREAT COMPANY

Enriched Timesheet Comments

- Using llama3.2 to make the original timesheet comments more detailed
- Using Json file to handle similar timesheet comments to speed up the enriched process

LLM as a Judge

REALLY GREAT COMPANY

- Using DeepSeek-r1:7b on local to evaluate the entire RAG system performance
- Evaluation prompt includes:
  - original query
  - refined query
  - retrieved documents
  - final llm answer
- Scoring Criteria:
  - Accuracy
  - Relevance
  - Clarity
  - Overall Performance

Demo Time