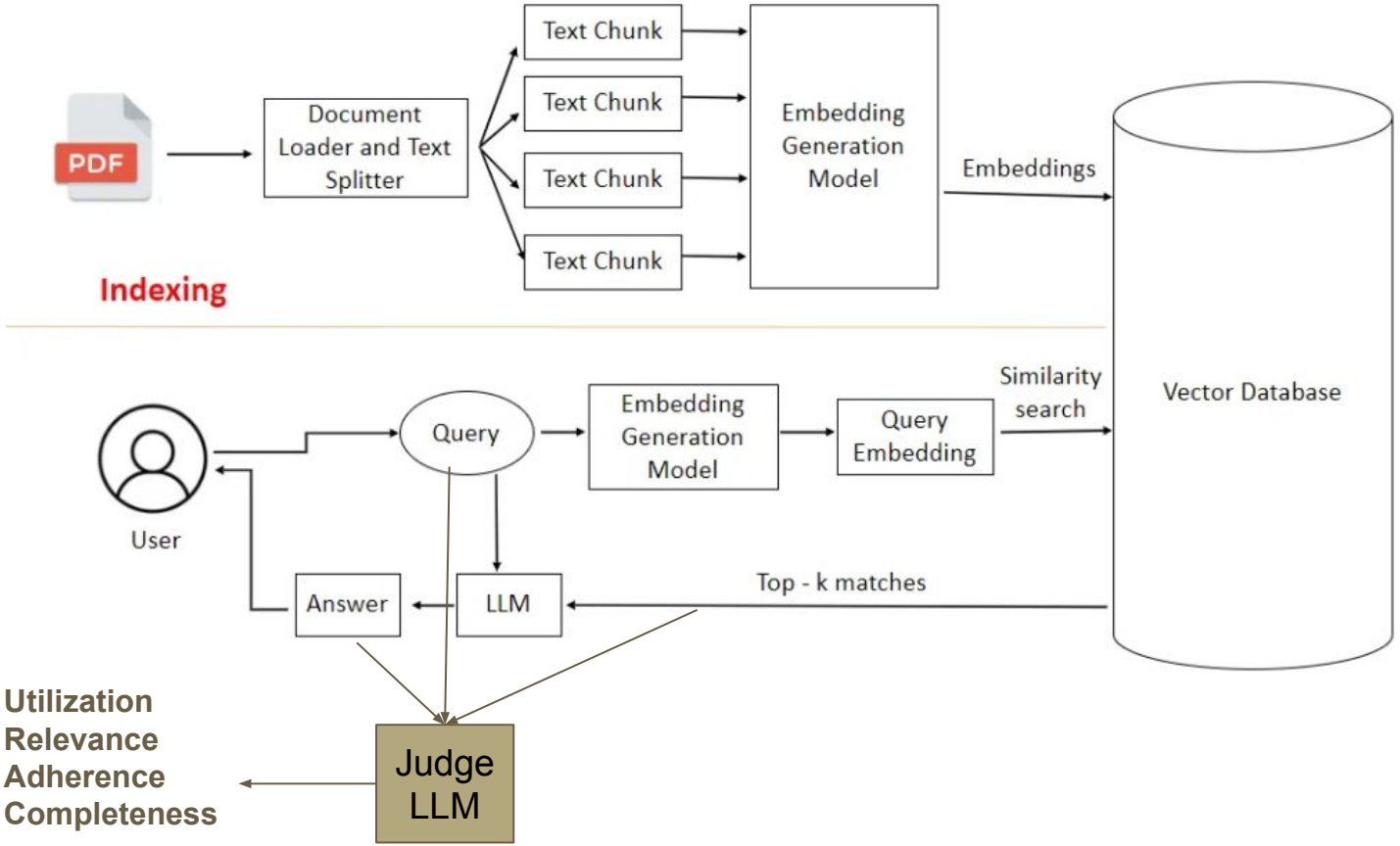
RAG Project

Group 25

Srinivas Gundu
Kaveri Pulibandla
Rahul Pelluri

Overview


PDF
Document Loader and Text Splitter
Text Chunk
Text Chunk
Text Chunk
Text Chunk
Embedding Generation Model
Embeddings
Vector Database
Indexing
User
Query
Embedding Generation Model
Query Embedding
Similarity search
Answer
LLM
Top - k matches
Utilization
Relevance
Adherence
Completeness
Judge LLM

Scope of Experiments

- Data chunking
- Embedding Model changes
- Using different Vector DBs
- Different Retrieval techniques
- Modifiying Judge LLM Prompt
- Gradio Implementation

Experiments Conducted

- Chunking strategies
  - Experiments
    - Sentence division
    - Recursive Character Text Splitter
    - Semantic Chunking
    - We stuck to recursive character text splitting. Semantic chunking took too much time and quickly exhausted our GPU credits.
- Embedding Model changes
  - Evaluated Models on MTEB with existing ratings for different db types
  - Experimented with small and long param models
  - Experimented with 11 models overall
- Vector DBs
  - Used FAISS and Chroma
  - Chroma offers more flexibility
  - Wrt Chroma, added metadata (generated from local LLM) info for better query
  - The process was GPU intensive and time taking but improved retrieval
- Retrieval Techniques
  - Hybrid, Vector retrieval mechanisms best combination was 0.1 ratio

Results

Finance/Legal

								Hybrid			Hybrid 25pc			Vector			MetadataSearch		
id	**gpt3_ad**	**gpt3_co**	**gpt35_u**	**relevan**	**utilizati**	**complet**	**Adhere**	utilizatio	complete	relevanc	utilizatio	complete	relevanc	utilizatio	complete	relevanc	utilizatio	complete	relevance
finqa_66	null	null	null	0.0588	0.0588	1	1	0.03455	1	0.02827	0.0911	1	0.15602	0.15393	1	0.14136	0.26154	1	0.54136
finqa_63	1	0.3333	0.1111	0.1111	0.1111	1	1	0.97059	0.33333	0.84454	NA	NA	NA	NA	NA	NA	NA	NA	NA
finqa_70	1	0.04	0.04	0.04	0.04	1	1	0.28558	1	0.06437	0.28105	0.75	0.11786	0.11242	1	0.11786	0.271	1	0.67664
finqa_70	1	0.2	0.05	0.05	0.05	1	1	0.09524	1	0.04762	0.08163	1	0.08163	0.33673	1	0.33673	NA	NA	NA

Embedding Model: “**multilingual-e5_chroma** Vector DB: “**Chroma**” Groq LLM: “**llama3-8b-8192**“

CovidQA Medical

							CovidQA		
id	gpt3_adherence	gpt3_context_relevance	gpt35_utilization	relevance_score	utilization_score	completeness_score	relevance_score	utilization_score	completeness_score
677	1	0.269231	0.115385	0.269231	0.076923	0.285714	0.125	1	1

Embedding Model: “pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb” Vector DB: “**Chroma**” Groq LLM: “**llama3-8b-8192**“

Hagrid General Knowledge

id	gpt3_adherence	gpt3_context_relevance	gpt35_utilization	relevance_score	utilization_score	completeness_score	relevance_score	utilization_score	completeness_score
hagrid_534_0	1	0.222222	0.111111	0.222222	0.111111	0.5	1	1	1

Embedding Model: “sentence-transformers/all-MiniLM-L6-v2” Vector DB: “**FAISS**” Groq LLM: “**llama3-8b-8192**“ Judge Model: deepseek/deepseek-r1:free

Observations

- Domain-specific embeddings (BioBERT) significantly improved retrieval relevance.
- Sentence-level metadata helped in accurate sentence mapping for judge LLM.
- Biomedical LLMs (or high-context generalist models like LLaMA3) performed well with guided prompts.
- Metadata search on Chroma gave better results but the GPU usage became very high and time taking.
- Query decomposition took time. But once implemented, it increased the number of retrieved documents and affected the final KPIs.
- Query decomposition is a good idea but takes more overall inference time.

Best Practices for Performance Improvement

- Use multiproc for chunking the data
- Chunk size less than 1000 for best results (As per research papers)
- ChromaDB Langchain wrapper does not support some functionality (progress bar, metadata incorporation).
  - Use Native Chroma library for creating the DB and Langchain for wrapper to query later.
  - Get summarization words for each chunk from a 7B LLM and feed as metadata. Query on the metadata as well as vector retrieval for best results.
  - Run a local LLM on Collab GPU for metatadata creation. (Used Mistral 7B)
- Tweaked the judge prompt for more straightforward results.
- Cache the results for repeated queries. More work needed.

Gradio Implementation

RAG System with Groq

Configuration | Query

Knowledge Domain

Customer Support

Vector Database

FAISS

Embedding Model (auto-selected)

sentence-transformers/all-MiniLM-L6-v2

Retriever LLM

llama3-8b-8192

Judge LLM

llama3-70b-8192

Chunk Size

500

Chunk Overlap

100

Top K Documents

5

Initialize System

Stop Initialization

Initialization Status

System initialized with Customer Support, FAISS, embedding: sentence-transformers/all-MiniLM-L6-v2

Chunk Information

20868 chunks created

Configuration Query

Your question

How to install Packaging utility? How to install Packaging utility?

Temperature 0.3

0.1 1

Max Tokens 512

128 2048

Submit

Answer

ling the installer package (pu.offering.disk.platform_version.zip) and running the install command.2. Using Installation Manager to install Packaging Utility from the Packaging Utility repository on www.ibm.com (if Installation Manager is already installed).Note that the fix pack package (pu.update_version.zip

Retrieved Documents

Document 1:
The installer package for Packaging Utility is pu.offering.disk.platform_version.zip where platform indicates the operating system and version indicates the version of Packaging Utility. The installer package contains files for only one platform. Using this package, you can install Packaging Utility and Installation Manager by running the install command. You can also add this package as a repository in Installation Manager and use the package to update Packaging Utility.

Document 2:
The installer package for Packaging Utility is pu.offering.disk.platform_version.zip where platform indicates the operating system and version indicates the version of Packaging Utility. The installer package contains files for only one platform. Using this package, you can install Packaging Utility and Installation Manager by running the install command. You can also add this package as a repository in Installation Manager and use the package to update Packaging Utility.

Document 3:
Back to top

Evaluation Results


```
{
 "relevance_explanation": "The response provides two methods to install Packaging Utility, which aligns with the question's intent.",
 "all_relevant_sentence_keys": [
  "0_0.",
  "0_2.",
  "1_0.",
  "1_2.",
  "2_0.",
  "2_1.",
  "3_0.",
  "3_1."
```


Metrics


```
 "context_relevance": 0.8,
 "context_utilization": 1.25,
 "completeness": 1.0,
 "adherence": 1.0,
 "explanation": "Relevant:8, Utilized:10, Supported:3/3"
}
```

Overall Results

After using metadata based search overall search quality shoots up

Dataset	Relevance	Utilization	Completion	Adherence
cuad	0.9	1	1	1
finqa	0.67	0.62	1	1
tatqa	1	1	1	1
covidqa	0.125	1	1	1
pubmedqa	1	1	1	1
hotpotqa	1	1	1	1
nsmarco	1	1	1	1
expertoa	1	1	1	1
hagrid	1	1	1	1
techqa	0.8	1	1	1
enaul	1	1	1	1
delucionqa	1	1	1	1

Overall Results

Dataset	RMSE	AUCROC
FINANCE	0.11	0.72
LEGAL	0.2	0.76
GENERAL KNOWLEDGE	0.15	0.89
CUSTOMER SUPPORT	0.12	0.91
BIOMEDICAL RESEARCH	0.14	0.87

TechQA

id	gpt3_adherence	gpt3_context_relevance	gpt35_utilization	relevance_score	utilization_score	completeness_score	relevance_score	utilization_score	completeness_score
techqa_DEV_Q243	0	0.027344	0.023438	0.011719	0.011719	1	1	1	1
techqa_DEV_Q253	null	null	null	0.00266	0.00133	0.5	1	1	1
techqa_DEV_Q008	1	0.020906	0.020906	0.062718	0.062718	1	1	1	1
techqa_DEV_Q266	1	0.107477	0.042056	0.070093	0.046729	0.533333	0.8	1.25	1

Embedding Model: “sentence-transformers/all-MiniLM-L6-v2” Vector DB: “**FAISS**” Groq LLM: “**llama3-8b-8192**“ Judge Model: llama3-70b-8192

RGB Dataset

Work Ongoing - Observations so far

- Code uploaded in github is buggy.
- Tried to run a LLAMA based quantised LLM on Collab for judge model but requires a Pro connection.
- Currently completed Evaluation and Fact evaluation checks for a small Qwen 2B LLM.
- Currently using a Mistral 7B model as a Judge using Huggingface Transformers.
- Will update the tables soon with comparisons of multiple LLMs.

Thank You