Information Retrieval Using LLMs

Shubhangi Sanyal
Anurag Dey

Submitted to:
Sarvesh Bhandaokar,
Ameya Kamat

Industry Project Report
May 18, 2024

CHENNAI MATHEMATICAL INSTITUTE

Contents

1 Introduction **2**

2 Problem Statement **2**

3 Workflow **2**

4 Data Pre Processing **2**
4.1 Tools Used 2
4.2 Ground Truth Generation 2

5 Advanced Pre-Processing **3**
5.1 Text Splitting 3
5.2 Text Embeddings 3
5.3 Database 3

6 Experimenting With Retrievers **3**
6.1 Multi Query Retriever 3
6.2 Parent Document Retriever 4
6.3 Merger Retriever 4

7 Re-Ranking **4**
7.1 Re-Rankers 4

8 Evaluation **5**
8.1 Evaluation (Question — Context Relevance) 5
8.1.1 Methodology 5
8.1.2 Results 5
8.1.3 Inference 5
8.2 Evaluation (GroundTruth — Context Relevance) 6
8.2.1 Inference 6

9 Conclusion **6**

10 References **6**

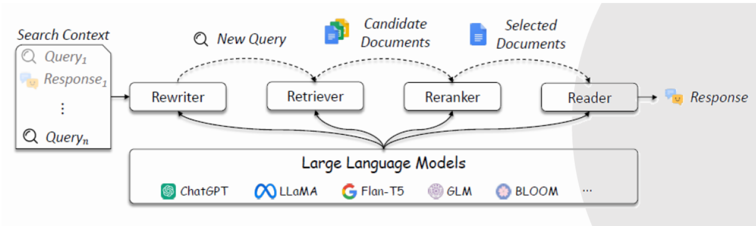
1 Introduction

Information Retrieval systems operate on extensive repositories. Hence, the efficiency of retrieval algorithms becomes of paramount importance. To improve the user experience, the retrieval performance is enhanced from both the upstream (query reformulation) and downstream (re-ranking and reading) perspectives.
The evolution of information retrieval (IR) has progressed from term-based methods and Boolean logic to the integration of neural models. Initially focused on keyword matching, IR transitioned to vector space models, allowing for the capture of semantic relationships. Further advancements included statistical language models — refining relevance estimation through contextual and probabilistic factors. Recently, large language models have been recognized as powerful tools exhibiting remarkable proficiency in language understanding and generation.

2 Problem Statement

This project aims to explore and implement Language Model-based approaches, specifically Leveraging Pre-trained Language Models (LLMs), to enhance information retrieval systems. Leveraging the power of advanced language models, such as GPT (Generative Pre-trained Transformer) and its variants, this project seeks to transform the traditional information retrieval process by optimizing the search, relevance, and contextual understanding of retrieved information.

3 Workflow



4 Data Pre Processing

Experimentation was done on Resume Data. The data was homogeneous in nature making the purpose of our project invalid. We moved to DMML lectures available on YouTube for further analysis.

4.1 Tools Used

- Assembly AI (API required)
- Whisper AI (Open Sourced)

4.2 Ground Truth Generation

- Generating Queries with ground truths based on transcripts. A total of 300 questions and answers were generated using GPT-3.5 Turbo along with human evaluation

The final text document is homogeneous and heterogeneous amongst themselves, leading to good retriever and re-ranking results

5 Advanced Pre-Processing

5.1 Text Splitting

Text splitting, often employed before embeddings, enhances data pre-processing and model training. Segmenting text prior to embedding extraction improves contextual understanding and downstream task performance."

- `RecursiveCharacterTextSplitting`: splits a text into smaller chunks recursively based on a specified length or other criteria. Useful for handling large text files or documents
- `Semantic Based Chunker` :splits text based on semantic similarity.

5.2 Text Embeddings

Embeddings in natural language processing encode words or phrases into dense vectors, capturing semantic relationships. These vectors enable algorithms to understand the contextual meanings of words, facilitating tasks like sentiment analysis and machine translation

- `BAAI/bge-large-en`: FlagEmbedding can map any text to a low-dimensional dense vector which can be used for tasks like retrieval, classification, clustering, or semantic search. And it also can be used in vector databases for LLMs.
- `SentenceTransformerEmbeddings`: BERT based embedding model that can be used for similar purposes.

5.3 Database

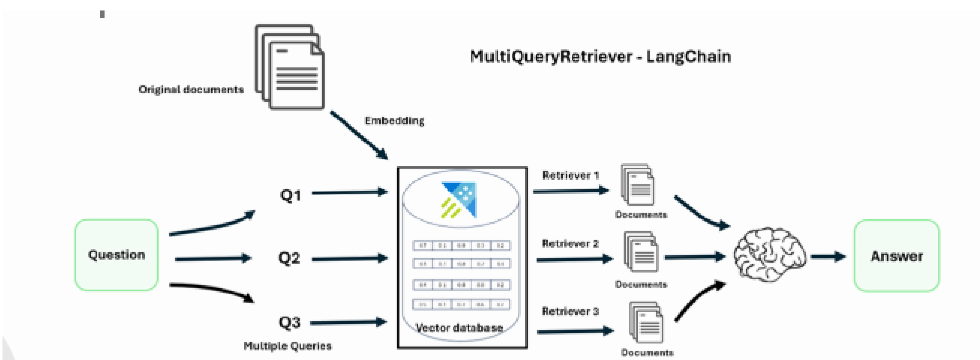
ChromaDB: ChromaDB is an open-source vector store that stores and retrieves vector embeddings and associated metadata for use by language models and semantic search engines. It's designed to manage and query collections of embeddings for tasks like semantic search and natural language processing.

---

6 Experimenting With Retrievers

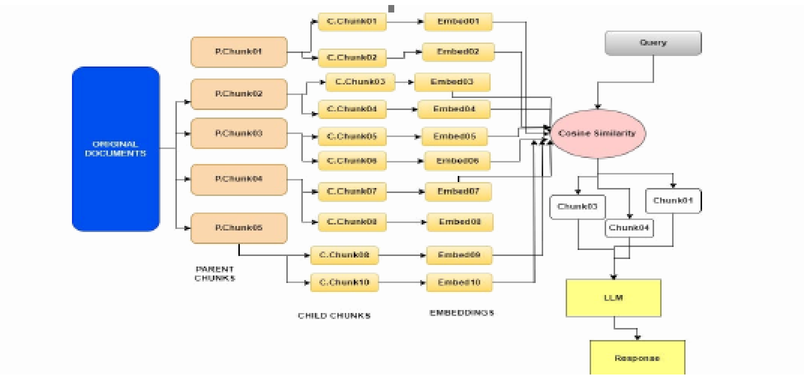
6.1 Multi Query Retriever

- The MultiQueryRetriever is a Python LangChain module.
- It automates prompt tuning by generating multiple queries from various perspectives based on a given user input query.
- For each query, it retrieves a set of relevant documents.
- The retriever combines all the queries to obtain a larger set of potentially relevant documents.



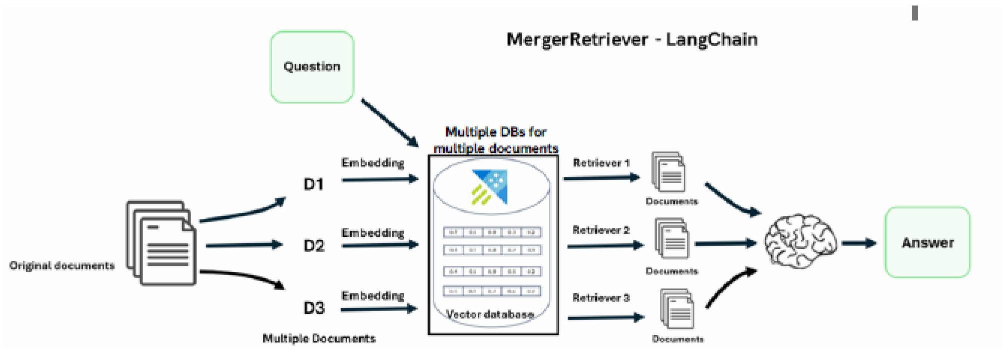
6.2 Parent Document Retriever

- PDRs are a type of multi-vector retrieval.
- Multi-vector retrieval is a retrieval method that allows the builder to embed alternative representations of their original documents.
- PDRs split and store small chunks of data to retain the context of each chunk.
- When retrieving, it first fetches the small chunks.
- It then looks up the parent IDs for those chunks and returns those larger documents.



6.3 Merger Retriever

- Lord of the Retrievers (LOTR), also known as MergerRetriever, takes a list of retrievers as input.
- It merges the results of their `get_relevant_documents()` methods into a single list.
- The merged results will be a list of documents that are relevant to the query and that have been ranked by the different retrievers.



---

7 Re-Ranking

Re-Ranking filters down the total number of documents into a fixed number. Re-ranker records and get the most **relevant** items at the top and they can be sent to the LLM.

7.1 Re-Rankers

- `Cross-Encoder Reranker`: A query and a possible document are passed simultaneously to a transformer network. The transformer network outputs a single score between 0 and 1 indicating how relevant the document is for the given query. The cross-encoder re-ranker reorders the top-N matching documents for a query.

- `Maximum Marginal Relevancy`: MMR calculates the similarity between a document and candidate keywords, as well as the similarity of already selected keyphrases and keywords. This results in a selection of keywords that maximize their within-diversity with respect to the document
- `Cohere`: API based reranker.Cohere Rerank is a semantic search technology that improves search results by ranking them based on semantic relevance, rather than just keywords. It's a component of Cohere's natural language processing (NLP) system, which uses a neural network to score candidates based on relevance, theme, style, and semantic similarity

---

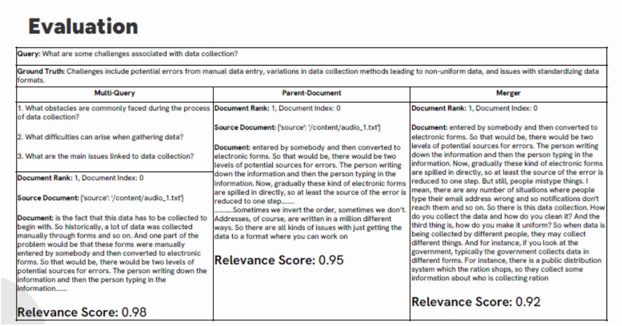
8 Evaluation

8.1 Evaluation (Question — Context Relevance)

8.1.1 Methodology

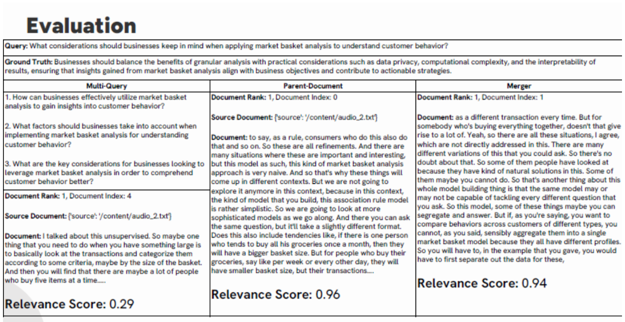
- We have selected 10 queries, 2 queries for each document.
- We retrieve 3 chunks from each document providing us a total of 15 items
- Use cohere re-ranking to evaluate the performance of Multi-Query, Parent-Document, and Merger Retriever (Relevance Score is provided)
- we also retrieve the source of the retrieved information as metadata

8.1.2 Results


Evaluation

Query: What are some challenges associated with data collection?		
Ground Truth: Challenges include potential errors from manual data entry, variations in data collection methods leading to non-uniform data, and issues with standardizing data formats.		
Multi-Query	**Parent-Document**	**Merger**
1. What obstacles are commonly faced during the process of data collection? 2. What difficulties can arise when gathering data? 3. What are the main issues linked to data collection? **Document Rank:** 1, Document Index: 0 **Source Document:** {'source': '/content/audio_1.txt'} **Document:** is the fact that this data has to be collected to begin with. So historically, a lot of data was collected manually through forms and so on. And one part of the problem would be that these forms were manually entered by somebody and then converted to electronic forms. So that would be, there would be two levels of potential sources for errors. The person writing down the information and then the person typing in the information....... **Relevance Score:** 0.98	**Document Rank:** 1, Document Index: 0 **Source Document:** {'source': '/content/audio_1.txt'} **Document:** entered by somebody and then converted to electronic forms. So that would be, there would be two levels of potential sources for errors. The person writing down the information and then the person typing in the information. Now, gradually these kind of electronic forms are spilled in directly, so at least the source of the error is reduced to one step........ ....Sometimes we invert the order, sometimes we don't. Addresses, of course, are written in a million different ways. So there are all kinds of issues with just getting the data to a format where you can work on **Relevance Score:** 0.95	**Document Rank:** 1, Document Index: 0 **Document:** entered by somebody and then converted to electronic forms. So that would be, there would be two levels of potential sources for errors. The person writing down the information and then the person typing in the information. Now, gradually these kind of electronic forms are spilled in directly, so at least the source of the error is reduced to one step. But still, people mistype things. I mean, there are any number of situations where people type their email address wrong and so notifications don't reach them and so on. So there is this data collection. How do you collect the data and how do you clean it? And the third thing is, how do you make it uniform? So when data is being collected by different people, they may collect different things. And for instance, if you look at the government, typically the government collects data in different forms. For instance, there is a public distribution system which the ration shops, so they collect some information about who is collecting ration **Relevance Score:** 0.92

(a) Figure 1


Evaluation

Query: What considerations should businesses keep in mind when applying market basket analysis to understand customer behavior?		
Ground Truth: Businesses should balance the benefits of granular analysis with practical considerations such as data privacy, computational complexity, and the interpretability of results, ensuring that insights gained from market basket analysis align with business objectives and contribute to actionable strategies.		
Multi-Query	**Parent-Document**	**Merger**
1. How can businesses effectively utilize market basket analysis to gain insights into customer behavior? 2. What factors should businesses take into account when implementing market basket analysis for understanding customer behavior? 3. What are the key considerations for businesses looking to leverage market basket analysis in order to comprehend customer behavior better? **Document Rank:** 1, Document Index: 4 **Source Document:** {'source': '/content/audio_2.txt'} **Document:** I talked about this unsupervised. So maybe one thing that you need to do when you have something large is to basically look at the transactions and categorize them according to some criteria, maybe by the size of the basket. And then you will find that there are maybe a lot of people who buy five items at a time..... **Relevance Score:** 0.29	**Document Rank:** 1, Document Index: 0 **Source Document:** {'source': '/content/audio_2.txt'} **Document:** to say, as a rule, consumers who do this also do that and so on. So these are all refinements. And there are many situations where these are important and interesting, but this model as such, this kind of market basket analysis approach is very naive. And so that's why these things will come up in different contexts. But we are not going to explore it anymore in this context, because in this context, the kind of model that you build, this association rule model is rather simplistic. So we are going to look at more sophisticated models as we go along. And there you can ask the same question, but it'll take a slightly different format. Does this also include tendencies like, if there is one person who tends to buy all his groceries once a month, then they will have a bigger basket size. But for people who buy their groceries, say like per week or every other day, they will have smaller basket size, but their transactions.... **Relevance Score:** 0.96	**Document Rank:** 1, Document Index: 1 **Document:** as a different transaction every time. But for somebody who's buying everything together, doesn't that give rise to a lot of. Yeah, so there are all these situations, I agree, which are not directly addressed in this. There are many different variations of this that you could ask. So there's no doubt about that. So some of them people have looked at because they have kind of natural solutions in this. Some of them maybe you cannot do. So that's another thing about this whole model building thing is that the same model may or may not be capable of tackling every different question that you ask. So this model, some of these things maybe you can segregate and answer. But if, as you're saying, you want to compare behaviors across customers of different types, you cannot, as you said, sensibly aggregate them into a single market basket model because they all have different profiles. So you will have to, in the example that you gave, you would have to first separate out the data for these, **Relevance Score:** 0.94

(b) Figure 2

8.1.3 Inference

- **MergerRetrievr**
  - Overall gives fairly good results
  - Average relevance score of top result across queries > 0.8
  - Increases diversity in answers, reduces bias
- **Parent Document Retriever**
  - Gives fairly good results for long context responses
- **Multi Query Retriever**
  - Gives fairly good results for short answers
  - Increases diversity in answers, reduces bias

8.2 Evaluation (GroundTruth — Context Relevance)

RAGAS evaluation:

Ragas is a framework that helps you evaluate your Retrieval Augmented Generation (RAG) pipelines. RAG denotes a class of LLM applications that use external data to augment the LLM's context.

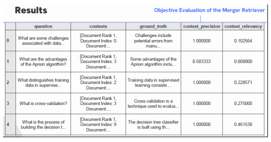
Context Precision:

Evaluates whether all of the ground-truth relevant items present in the contexts are ranked higher or not. Ideally, all the relevant chunks must appear at the top ranks.

Context Relevancy:

Gauges the relevancy of the retrieved context, calculated based on both the question and contexts. Ideally, the retrieved context should exclusively contain essential information to address the provided query.

Results

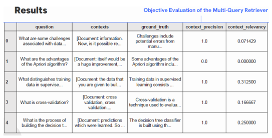


Results

Objective Evaluation of the Merger Retriever

	question	contexts	ground_truth	context_precision	context_relevancy
0	What are some challenges associated with data...	[Document Rank 1, Document Index: 0 Document:...	Challenges include potential errors from manu...	1.000000	0.102564
1	What are the advantages of the Apriori algorithm?	[Document Rank 1, Document Index: 3 Document:...	Some advantages of the Apriori algorithm inclu...	0.583333	0.000000
2	What distinguishes training data in supervise...	[Document Rank 1, Document Index: 2 Document:...	Training data in supervised learning consists ...	1.000000	0.228571
3	What is cross-validation?	[Document Rank 1, Document Index: 3 Document:...	Cross-validation is a technique used to evalua...	1.000000	0.275000
4	What is the process of building the decision t...	[Document Rank 1, Document Index: 9 Document:...	The decision tree classifier is built using th...	1.000000	0.461538

(a) Figure 1 -Merger Retriever

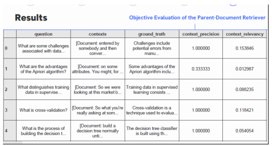


Results

Objective Evaluation of the Multi-Query Retriever

	question	contexts	ground_truth	context_precision	context_relevancy
0	What are some challenges associated with data...	[Document: information. Now, is it possible re...	Challenges include potential errors from manu...	1.0	0.071429
1	What are the advantages of the Apriori algorithm?	[Document: itself would be a huge improvement,...	Some advantages of the Apriori algorithm inclu...	0.0	0.000000
2	What distinguishes training data in supervise...	[Document: the data that you are given to buil...	Training data in supervised learning consists ...	1.0	0.312500
3	What is cross-validation?	[Document: cross validation, cross validation....	Cross-validation is a technique used to evalua...	1.0	0.166667
4	What is the process of building the decision t...	[Document: predictions which were learned. So ...	The decision tree classifier is built using th...	1.0	0.250000

(b) Figure 2- Multi Query Retriever



Results

Objective Evaluation of the Parent-Document Retriever

	question	contexts	ground_truth	context_precision	context_relevancy
0	What are some challenges associated with data...	[Document: entered by somebody and then conver...	Challenges include potential errors from manu...	1.000000	0.153846
1	What are the advantages of the Apriori algorithm?	[Document: on some attributes. You might, for ...	Some advantages of the Apriori algorithm inclu...	0.333333	0.012987
2	What distinguishes training data in supervise...	[Document: So we were looking at this market b...	Training data in supervised learning consists ...	1.000000	0.088235
3	What is cross-validation?	[Document: So what you're really asking at som...	Cross-validation is a technique used to evalua...	1.000000	0.118421
4	What is the process of building the decision t...	[Document: build a decision tree normally unti...	The decision tree classifier is built using th...	1.000000	0.054054

(c) Figure 3 -Parent Document Retreiever

8.2.1 Inference

- **High Context Precision**: Retrievers (with proper re-ranking) are able to extract the right information from the transcripts. Similar to analysis as done before
- **Low Context Relevancy**: Ground Truths are AI generated and there lies a high possibility of hallucination. We also know for a fact that ground truths are very precise in nature, opposite to the assumption that knowledge in documents is dispersed in nature

---

9 Conclusion

Some Recommendations for Future Work

- Experimenting with more structured and compact data
  - Might Lead to better context relevancy Multi Query might work better if the contextual answers are smaller
- Combining the features of Retriever
  - Given a query, generate multiple queries.
  - For each query, get relevant chunks based on child and parent splitter.
  - Display the unique union of responses

---

10 References

Click here to see all the code and report.