Industry Project – CMI Internship with Alumni

Mentors:
Sarvesh Bhandaokar
Ameya Kamat

Information Retrieval

Using LLMs

Anurag Dey
Shubhangi Sanyal

23rd April, 2024

Table of Contents

01 **Project Overview**

02 **Background**

03 **Analysis Pipeline**

04 **Conclusion**

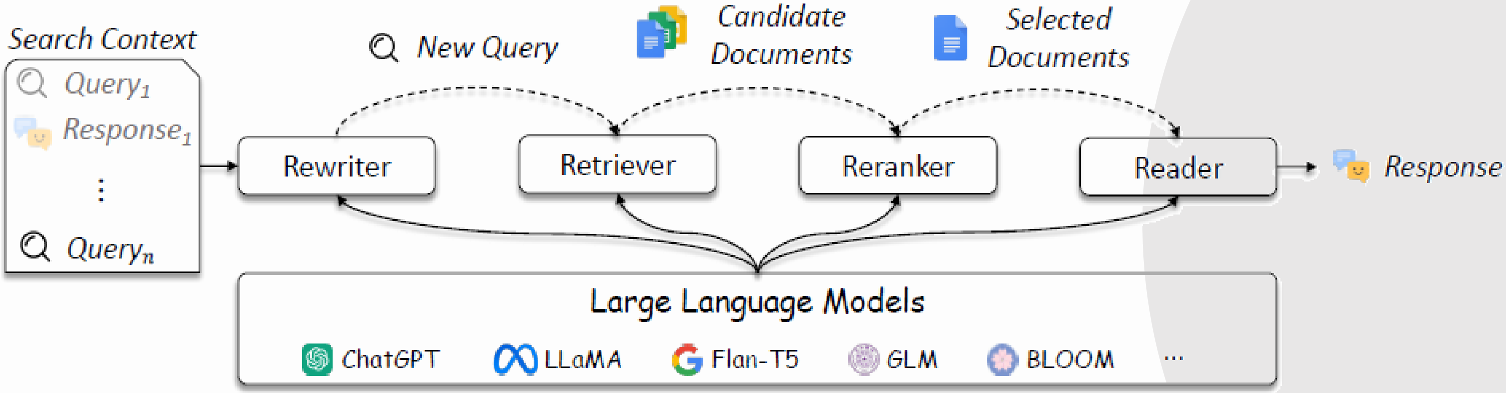
01 Project Overview

Building an advanced information retrieval engine with semantic search capabilities for custom data.

- Explore and implement Language Model-based approaches.
- Leveraging Pre-trained Language Models (LLMs), to enhance information retrieval systems.
- Optimizing the search, relevance, and contextual understanding of retrieved information.

Our focus is on Information Retrieval

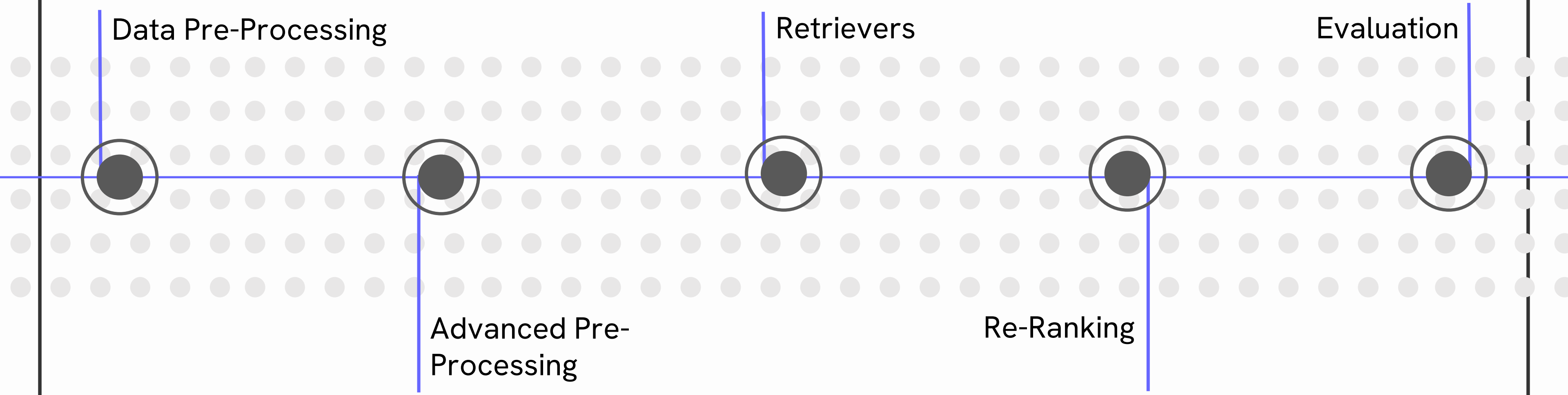
02 Background



Overview of existing studies that apply LLMs into information retrieval.

LLMs can be used in query rewriter, retriever, re-ranker, and reader.

03 Analysis Pipeline



Data Pre-Processing

Advanced Pre-Processing

Retrievers

Re-Ranking

Evaluation

Data Pre-processing

Generating transcripts from CMI lecture videos

- Tool used – Assembly AI

Generating queries and ground truths based on those transcripts

- Tool Used – GPT3.5 (with human evaluation)

Some trials and errors

- Dataset used: Resume data
- Extracting texts from pdfs with multiple columns
- Tool used: PyMuPDF
- Technique used: Bounding box with added heuristics

Dataset **discarded** because:

- Homogeneous nature
- Unfit for ranking documents

Advanced Pre-processing

Text Embeddings

`BAAI/bge-large-en`

- create semantic embeddings for textual data
- designed to capture the meanings of sentences or text in English

Text Splitting

`RecursiveCharacterTextSplitter`

- splits a text into smaller chunks recursively based on a specified length or other criteria
- useful for handling large text files or documents

Some other trials

`SentenceTransformerEmbeddings`

- State-of-the-art sentence, text and image embeddings

`Semantic Based Chunker`

- Splits Text based on semantic Similarity

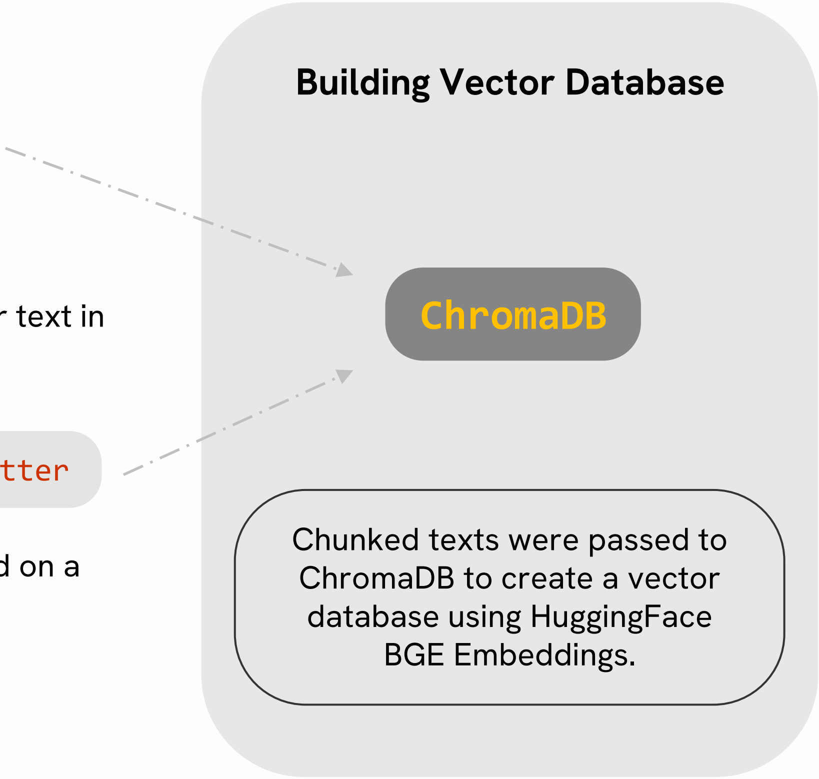
Advanced Pre-processing

Text Embeddings `BAAI/bge-large-en`

- create semantic embeddings for textual data
- designed to capture the meanings of sentences or text in English

Text Splitting `RecursiveCharacterTextSplitter`

- splits a text into smaller chunks recursively based on a specified length or other criteria
- useful for handling large text files or documents



Building Vector Database

Chunked texts were passed to ChromaDB to create a vector database using HuggingFace BGE Embeddings.

Experimenting with Retrievers

Multi-Query Retriever

Parent-Document Retriever

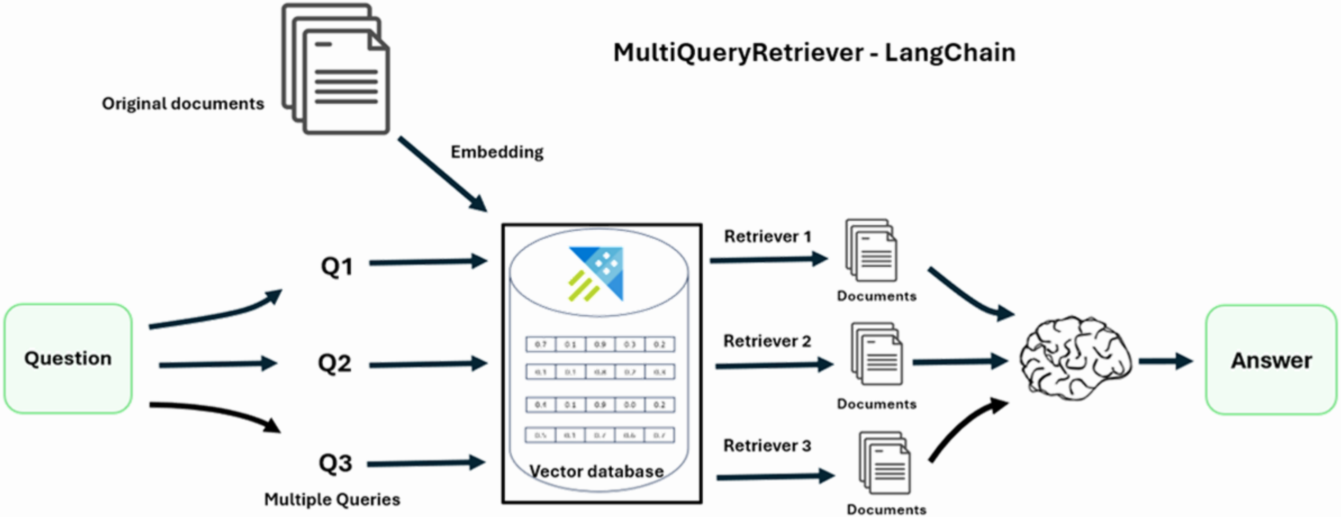
Merger Retriever

Experimenting with Retrievers

Multi-Query Retriever

Parent-Document Retriever

Merger Retriever

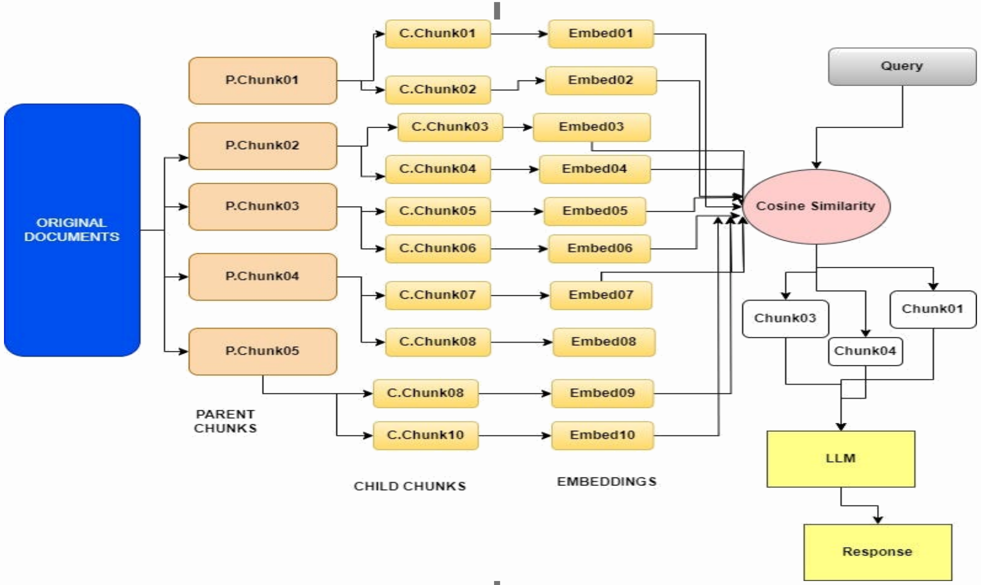


Experimenting with Retrievers

Multi-Query Retriever

Parent-Document Retriever

Merger Retriever



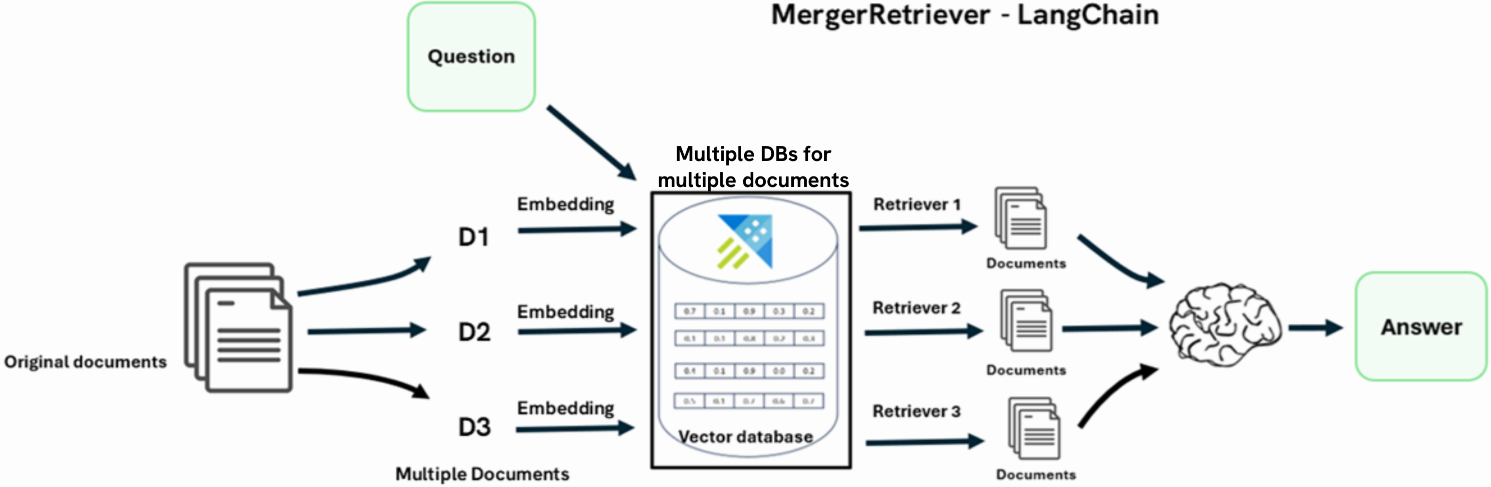
Experimenting with Retrievers

Multi-Query Retriever

Parent-Document Retriever

Merger Retriever

MergerRetriever - LangChain



Re-ranking

- **Filter** down the total number of documents into a fixed number
- Re-rank the records and get the most **relevant** items at the top and they can be sent to the LLM
- Offers a solution by finding those **records** that may not be within the top 3 results and put them into a smaller set of results that can be further fed into the LLM

Cohere Re-ranker:

- Cohere: Canadian startup that provides NLP models
- Given a query and a list of documents, Re-rank indexes the documents from most to least semantically relevant to the query.

Evaluation

<table>
<tr><td colspan="3"><b>Query:</b> What are some challenges associated with data collection?</td></tr>
<tr><td colspan="3"><b>Ground Truth:</b> Challenges include potential errors from manual data entry, variations in data collection methods leading to non-uniform data, and issues with standardizing data formats.</td></tr>
<tr><th>Multi-Query</th><th>Parent-Document</th><th>Merger</th></tr>
<tr><td>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?</td><td rowspan="2"><b>Document Rank:</b> 1, Document Index: 0
<b>Source Document:</b> {'source': '/content/audio_1.txt'}
<b>Document:</b> 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
<b>Relevance Score:</b> 0.95</td><td rowspan="2"><b>Document Rank:</b> 1, Document Index: 0
<b>Document:</b> 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
<b>Relevance Score:</b> 0.92</td></tr>
<tr><td><b>Document Rank:</b> 1, Document Index: 0
<b>Source Document:</b> {'source': '/content/audio_1.txt'}
<b>Document:</b> 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.......
<b>Relevance Score:</b> 0.98</td></tr>
</table>

Evaluation

<table>
<tr><td colspan="3"><b>Query:</b> What considerations should businesses keep in mind when applying market basket analysis to understand customer behavior?</td></tr>
<tr><td colspan="3"><b>Ground Truth:</b> 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.</td></tr>
<tr><th>Multi-Query</th><th>Parent-Document</th><th>Merger</th></tr>
<tr><td>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?</td><td rowspan="2"><b>Document Rank:</b> 1, Document Index: 0
<b>Source Document:</b> {'source': '/content/audio_2.txt'}
<b>Document:</b> 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....
<b>Relevance Score:</b> 0.96</td><td rowspan="2"><b>Document Rank:</b> 1, Document Index: 1
<b>Document:</b> 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,
<b>Relevance Score:</b> 0.94</td></tr>
<tr><td><b>Document Rank:</b> 1, Document Index: 4
<b>Source Document:</b> {'source': '/content/audio_2.txt'}
<b>Document:</b> 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.....
<b>Relevance Score:</b> 0.29</td></tr>
</table>

Inferences

Merger Retriever

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 really good results for long context responses

Multi-Query Retriever

Gives really good results for short answers

Increases diversity in answers, reduces bias

More at: Appendix

Results

A note on context precision and context relevancy

context precision

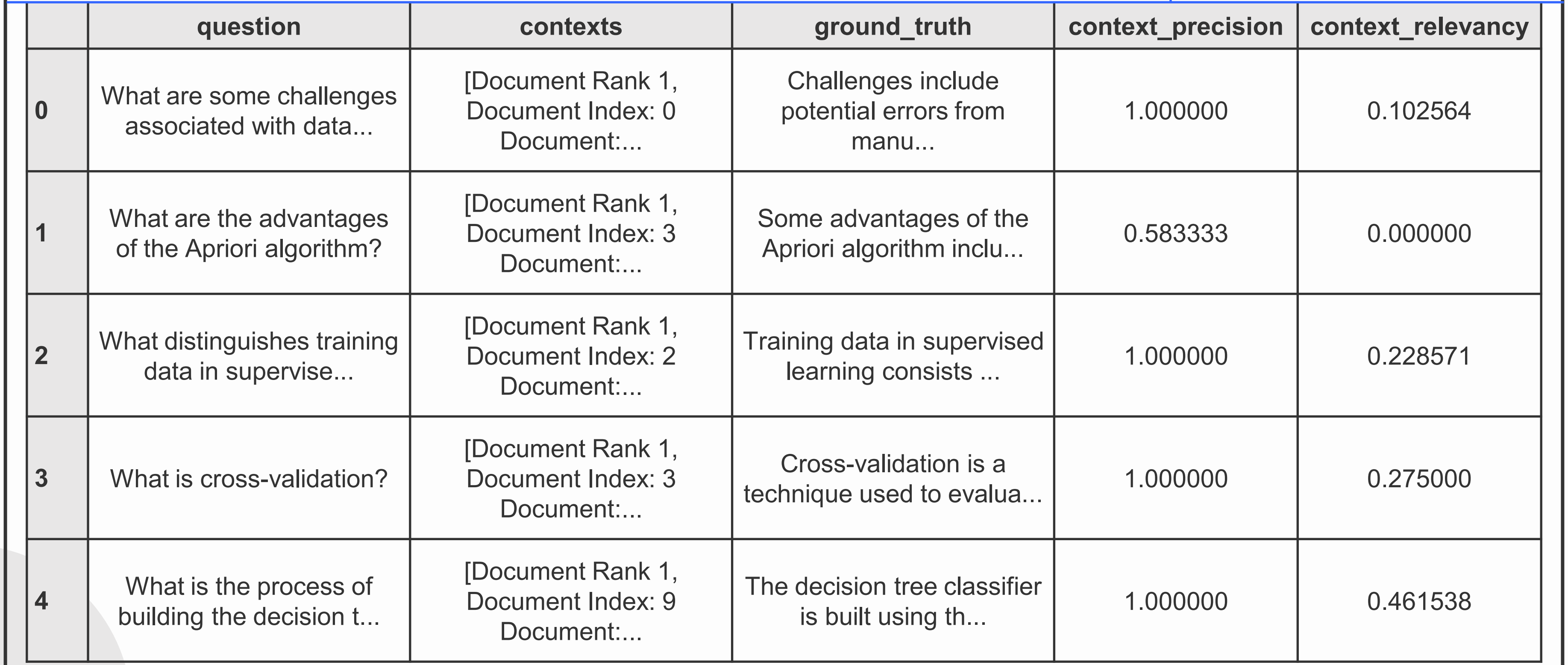
- Evaluates whether all of the ground-truth relevant items present in the contexts are ranked higher or not
- Ideally all the relevant chunks must appear at the top ranks

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

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

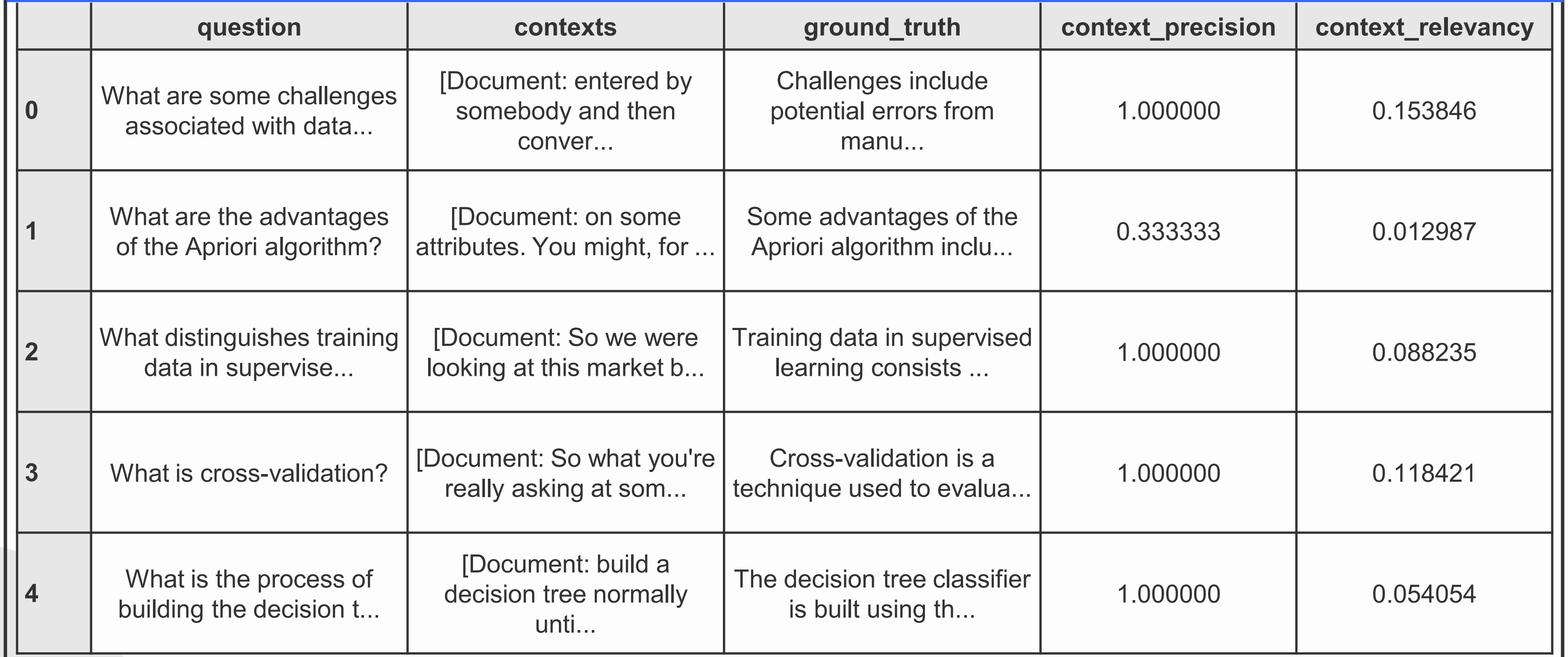
Results

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

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

Results

Inferences on context precision and context relevancy

context precision

Evaluates whether all of the ground-truth relevant items present in the contexts are ranked higher or not

context relevancy

retrieved context should exclusively contain essential information to address the provided query

Why do our Retrievers have

(a) **High context precision**

- Retrievers (with proper re-ranking) are able to extract the right information from the transcripts

(b) **Low context relevancy**

- Information extracted is in the form of raw text and has not been summarized by any LLM

04 Conclusion

Some **recommendations** for future work

Experimenting on 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 the retrievers

- Given a query, generate multiple queries.
- For each query, get relevant chunks based on child and parent splitter
- Display the unique union of responses

Thank You!

Created and Presented by:

Shubhangi Sanyal (shubhangi@cmi.ac.in)
Anurag Dey (anurag@cmi.ac.in)

Appendix

Context Relevancy

Estimate the value of $|S|$ by identifying sentences within the retrieved context that are relevant for answering the given question.

$$\text{context relevancy} = \frac{|S|}{|\text{Total number of sentences in retrieved context}|}$$

Context Precision

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K} (\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}}$$

$$\text{Precision@k} = \frac{\text{true positives@k}}{(\text{true positives@k} + \text{false positives@k})}$$

Where K is the total number of chunks in contexts and $v_k \in \{0,1\}$ is the relevance indicator at rank k.