RAG-QABot

Part 1: RAG Model-based QA Bot:

- Located in the RAGModel folder.
- Implements a Retrieval-Augmented Generation pipeline in RAGModel-QABot.ipynb.

Part 2: Interactive QA Bot Interface:

- Located in the QA-Bot folder.
- Provides an interactive user interface using Streamlit for uploading financial PDFs and querying the extracted data.

Example Queries and Their Output

Future Improvements

Part 1: RAG Model QA Bot

Approach

The financial data is first loaded and preprocessed to extract meaningful information, such as tabular data from P&L statements. This preprocessing ensures that the data is structured and ready for embedding. Using a transformer-based Sentence Transformer model, the data is encoded into dense vector representations, capturing the semantic meaning of the financial content. These embeddings are then indexed in Pinecone, a high-performance vector database, allowing efficient storage and retrieval. Queries entered by users are similarly encoded into vectors and matched against the indexed data in Pinecone. The most relevant results are retrieved based on similarity scores, providing the context needed for generating accurate answers to user queries.

Usage Instructions

- Open RAGModel-QABot.ipynb in google collab.
- Follow the steps to:
  - Upload the pdf file containing P&L data
  - Run all the parts.
  - Test the model by running example queries.

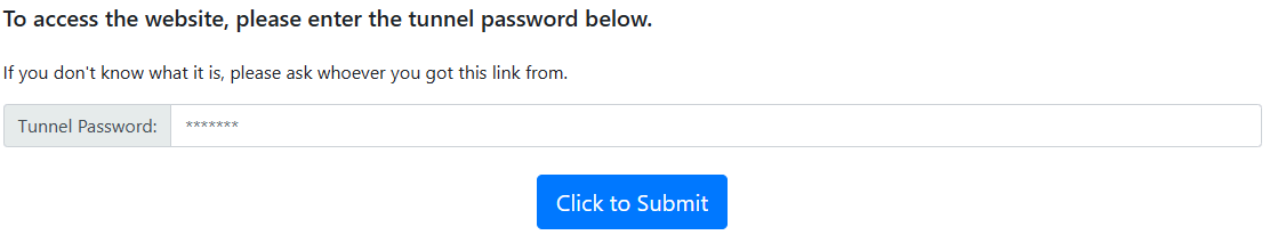
Part 2: Interactive QA Bot Interface

Approach

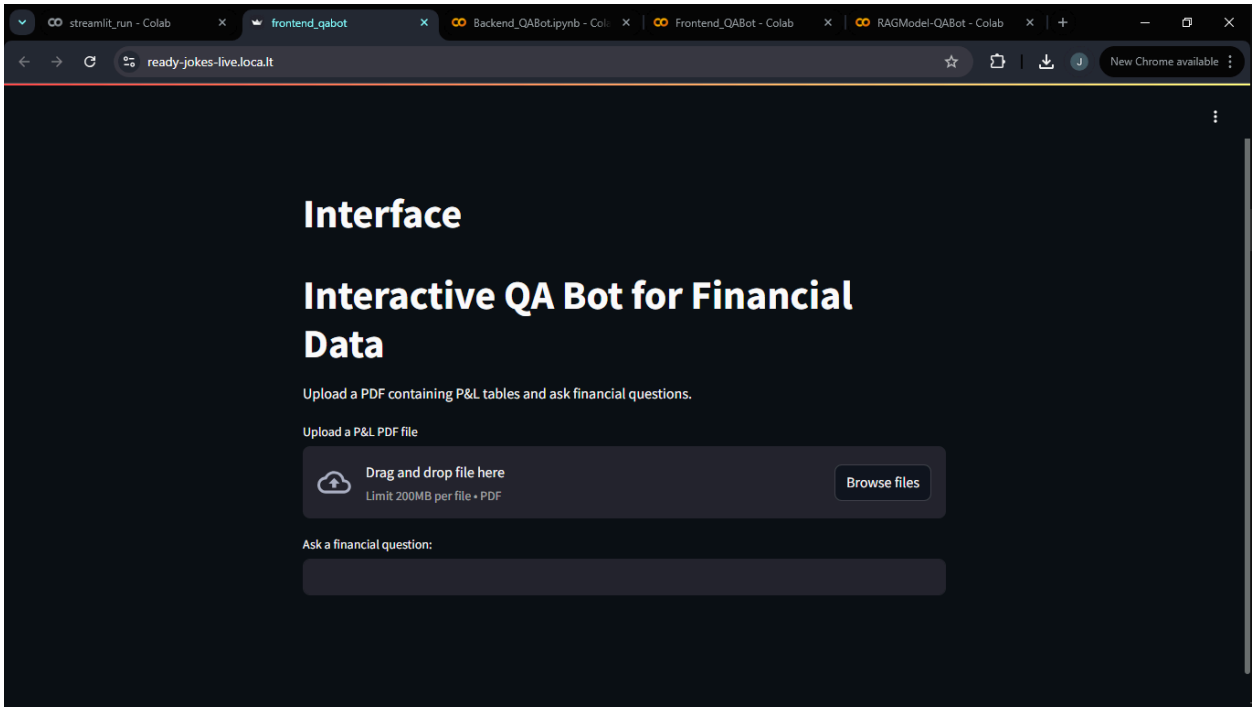
Frontend:
Built with Streamlit for uploading financial PDFs and querying.
Displays the extracted P&L data in a table format.
Backend:
Extracts data from uploaded PDFs (e.g., P&L tables) using tabula and pdfplumber.
Embeds data using Sentence Transformers and stores it in Pinecone.
Queries Pinecone and retrieves the most relevant rows from the document.
Deployment:
Deployed locally using Streamlit and exposed via localtunnel.

Usage Instructions

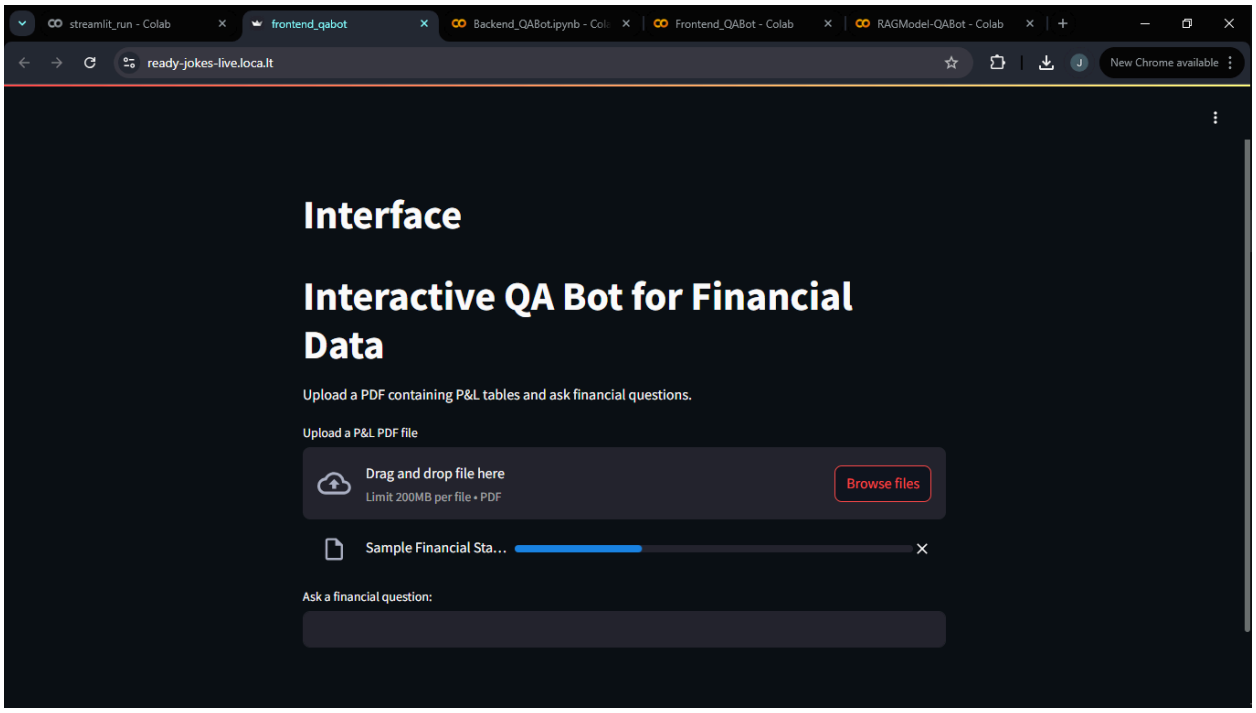
- Open streamlit_run.ipynb in Google Collab.
- Upload frontend_qabot.py and backend_qabot.py
- Install the libraries
- Run the command- **!wget -q -O - ipv4.icanhazip.com**
- copy the IP address
- Run the command- **!streamlit run frontend_qabot.py & npx localtunnel --port 8501**
- Click on *'your URL'*
- Enter the copied IP address



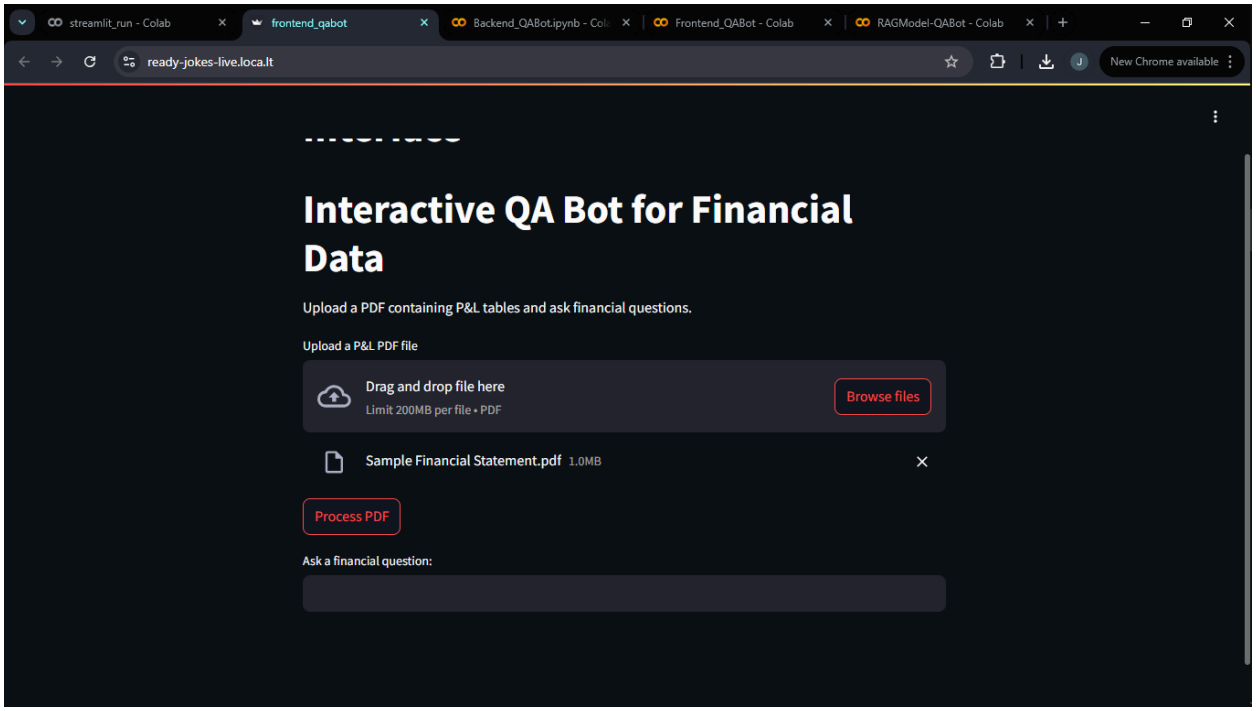
- Hence you will get the Interface of the QA BOT



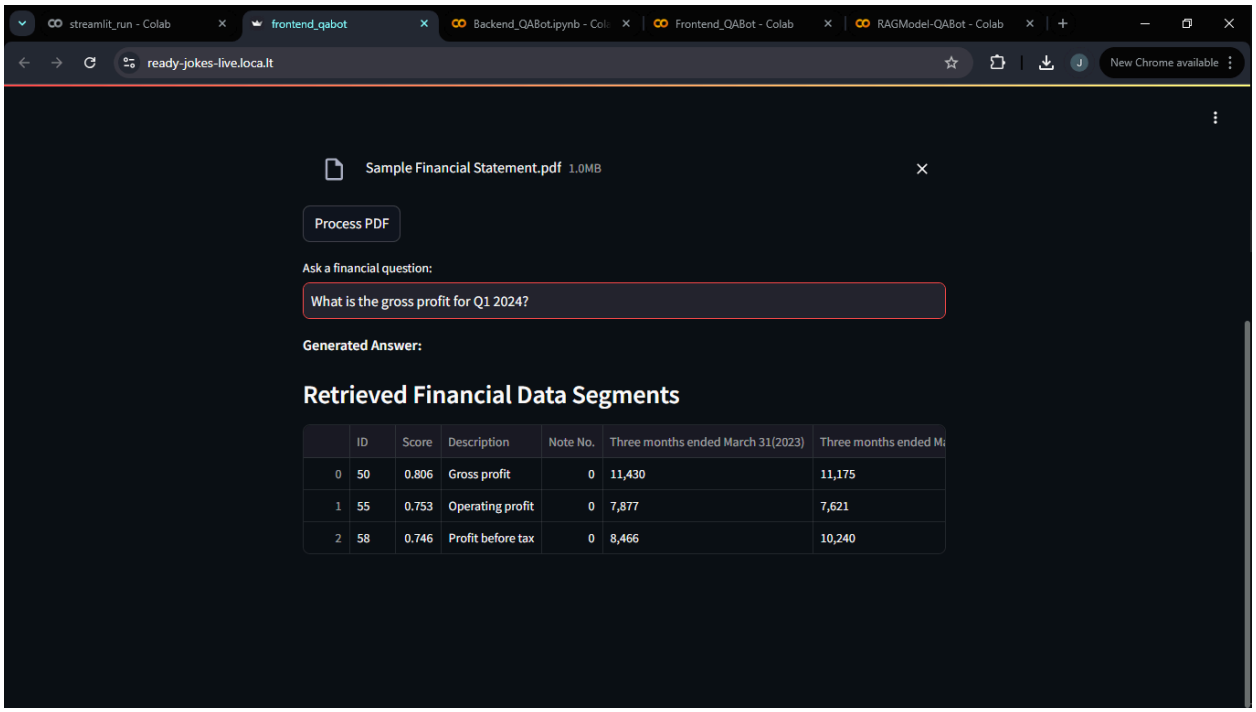
- Upload the PDF file containing P&L data



- Click on the ‘Process PDF’ button



- Enter your financial Queries



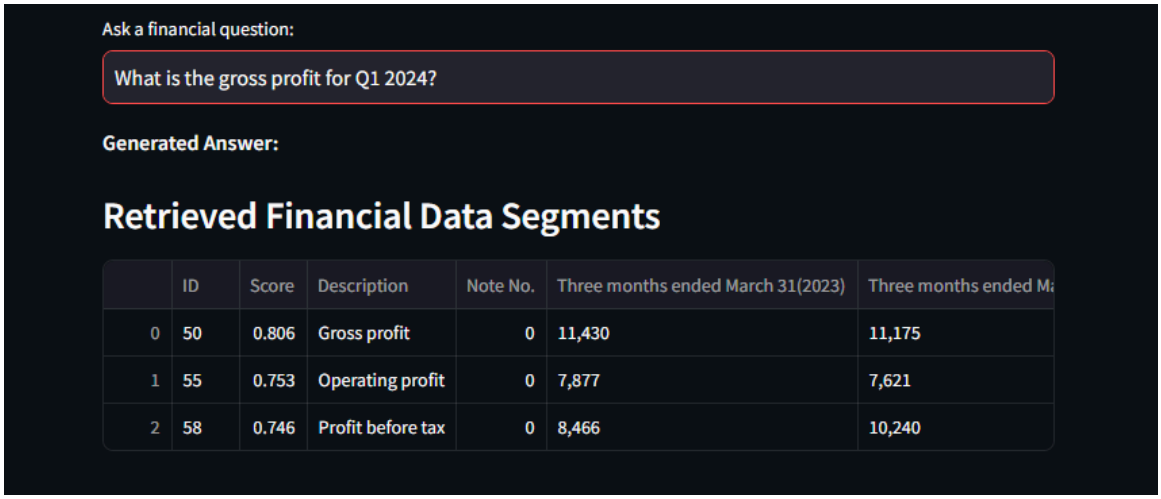
NOTE:
The table displays the top 3 results for the financial query given to the bot.

Overall Workflow

- Upload a PDF containing a P&L table.
- Click "Process PDF" to preprocess and index the data in Pinecone.
- Enter financial queries (e.g., "What is the gross profit?") in the input box.
- View the retrieved answers alongside the relevant P&L segments in an interactive table.

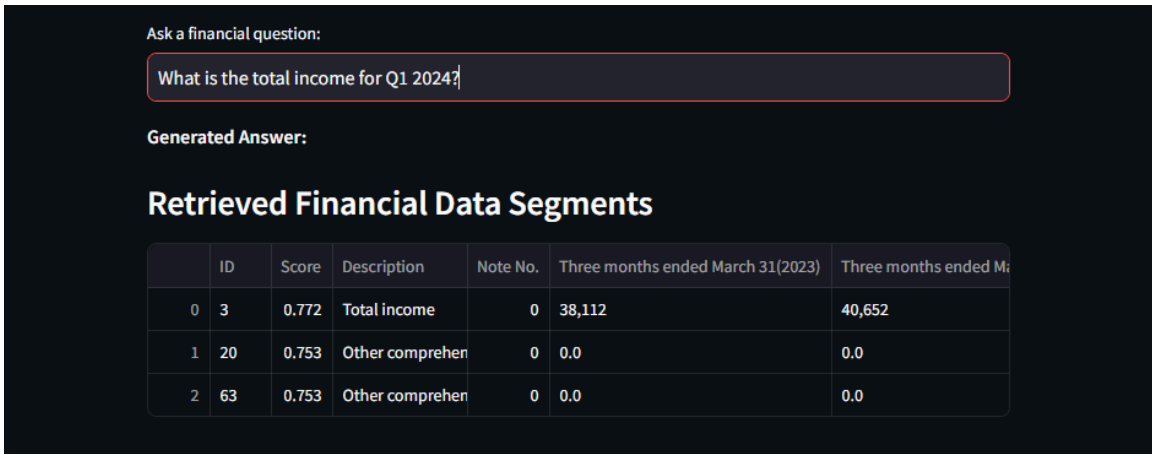
Example Queries and their Output

Query 1: What is the gross profit for Q1 2024?
The first most result scores 0.806 followed by 0.75 and 0.74



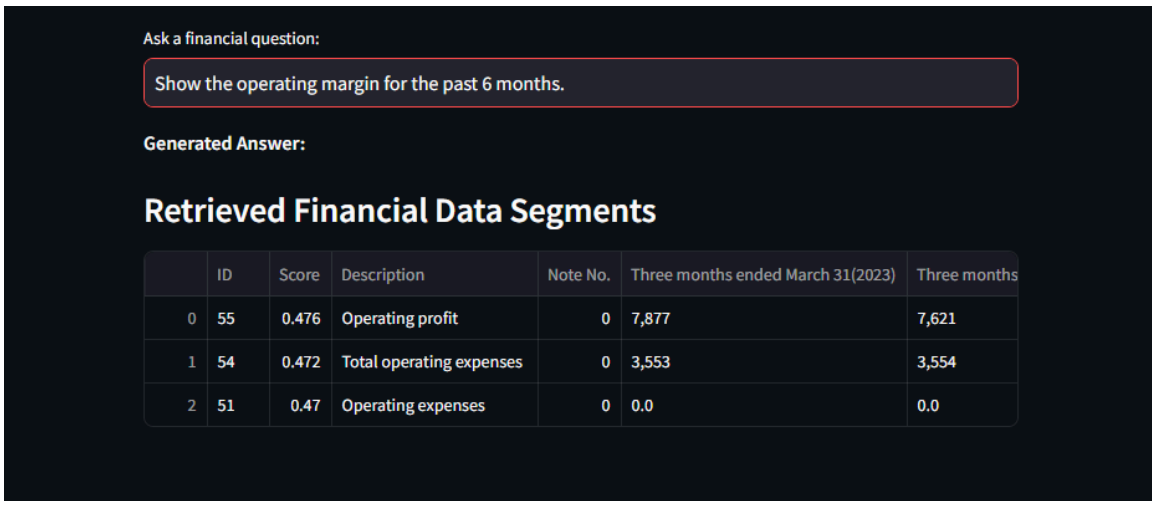
	ID	Score	Description	Note No.	Three months ended March 31(2023)	Three months ended Ma
0	50	0.806	Gross profit	0	11,430	11,175
1	55	0.753	Operating profit	0	7,877	7,621
2	58	0.746	Profit before tax	0	8,466	10,240

Query 2: What is the total for Q1 2024?
The first most result scores 0.772 followed by 0.75 and 0.75



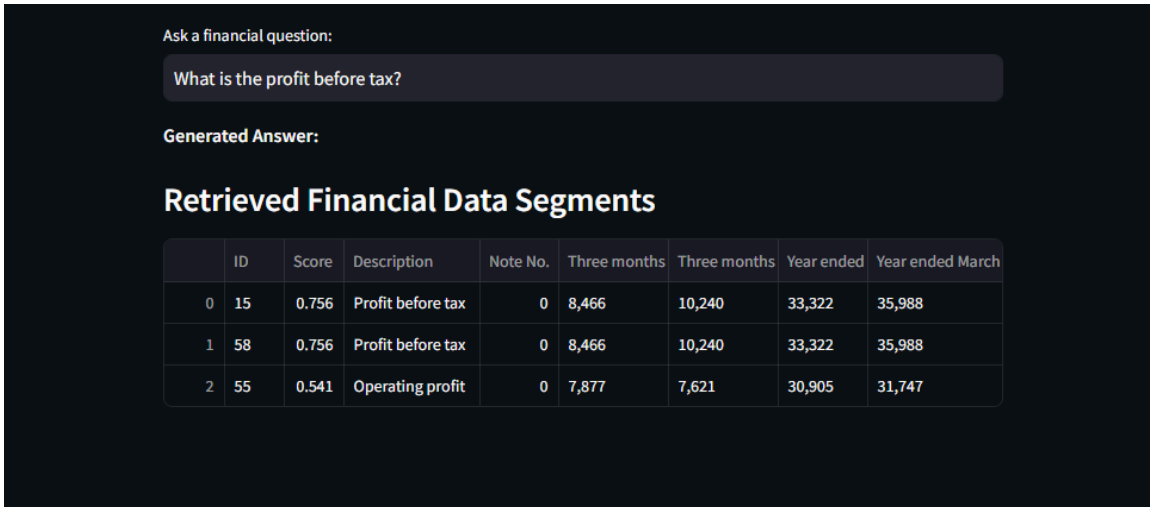
	ID	Score	Description	Note No.	Three months ended March 31(2023)	Three months ended Ma
0	3	0.772	Total income	0	38,112	40,652
1	20	0.753	Other comprehen	0	0.0	0.0
2	63	0.753	Other comprehen	0	0.0	0.0

Query 3: Show the operating margin for the past 6 months.
The first most result scores 0.476 followed by 0.472 and 0.47
Note: Score decreased due the less similarity in words like (margin and profit)



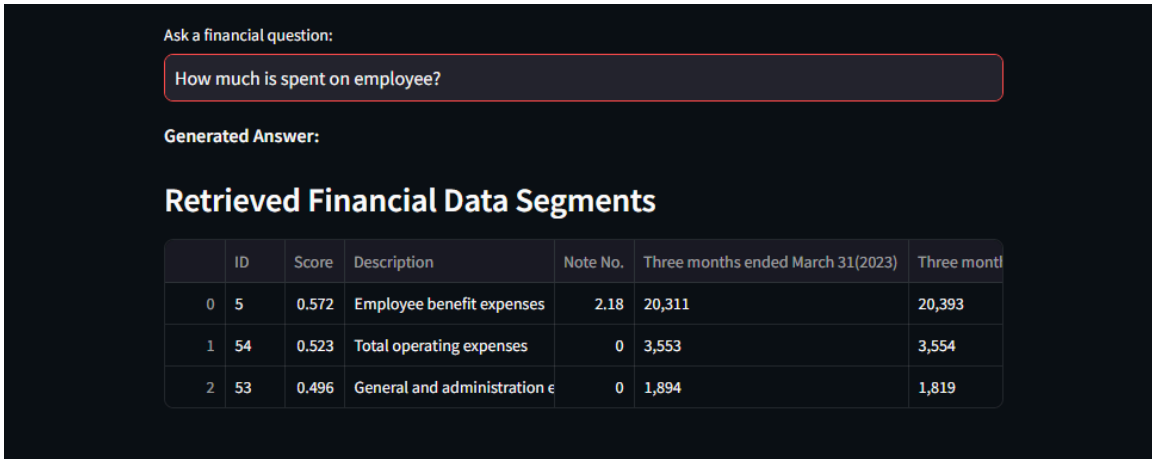
	ID	Score	Description	Note No.	Three months ended March 31(2023)	Three months
0	55	0.476	Operating profit	0	7,877	7,621
1	54	0.472	Total operating expenses	0	3,553	3,554
2	51	0.47	Operating expenses	0	0.0	0.0

Query 4: What is the profit before tax?
The first most result scores 0.756 followed by 0.756 and 0.541
Note: Two recurring results, led due their presence in two different tables



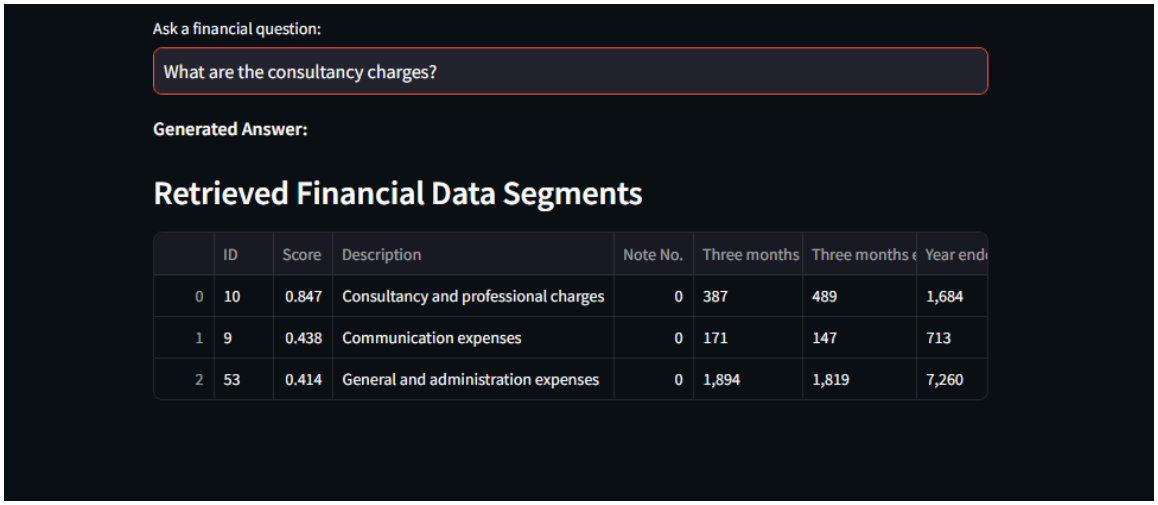
	ID	Score	Description	Note No.	Three months	Three months	Year ended	Year ended March
0	15	0.756	Profit before tax	0	8,466	10,240	33,322	35,988
1	58	0.756	Profit before tax	0	8,466	10,240	33,322	35,988
2	55	0.541	Operating profit	0	7,877	7,621	30,905	31,747

Query 6: How much is spent on employee?
The first most result scores 0.572 followed by 0.573 and 0.496
Note: Score would have increased if the more information was provided like benefit.



	ID	Score	Description	Note No.	Three months ended March 31(2023)	Three montl
0	5	0.572	Employee benefit expenses	2.18	20,311	20,393
1	54	0.523	Total operating expenses	0	3,553	3,554
2	53	0.496	General and administration e	0	1,894	1,819

Query 7: What are the consultancy charges?
The first most result scores 0.847 followed by 0.438 and 0.414



	ID	Score	Description	Note No.	Three months	Three months e	Year end
0	10	0.847	Consultancy and professional charges	0	387	489	1,684
1	9	0.438	Communication expenses	0	171	147	713
2	53	0.414	General and administration expenses	0	1,894	1,819	7,260

Future Improvements

- Enhance PDF parsing for more complex tables.
- Usage of LLMs for more relatable human-friendly responses.