

Conversational AI Assignment 2

Group : 70

Members:

- ATUL SHARMA
- MONISH. K.B
- SHIVAM RAVI
- VENKATRAMAN SOUMYA P G VENKATRAMAN
- VOBANI VENKATESWARLU

Problem Statement:

- Implement Basic RAG
- Implement Advanced RAG using **Memory-Augmented Retrieval**

Setup the application:

1. Run below command from the root of the directory to create a virtual environment

```
python -m venv ./venv
```

2. Activate the virtual environment.

For linux/mac:

```
source ./venv/bin/activate
```

For Windows:

```
.\venv\Scripts\activate
```

3. Install Dependencies:

```
pip install -r requirements.txt
```

4. Run the application:

```
streamlit run Group_70_rag_app.py
```

This will start the **Web-UI** at <http://localhost:8501/>

Usage:

1. On the UI select either of the two options Basic RAG or Advanced RAG on the side panel.
2. Upload a PDF file containing the company financial data. Optionally you can also select the **chunk size** on the side panel to chunk the data into specific chunk sizes, before uploading the documents.

Note: Sample file **Group_70_sample_data_Q3_24.pdf** has been provided for reference.

3. Enter any query in the query box and hit enter to generate response.

Note: Optionally you can also select the number of chunks to refer dynamically by selecting the **top_k** parameter from the sidebar.

4. **Answer**, along with the **confidence scores** will be shown on the UI. Also, **top retrieved documents** will be available for exploration under an expander tab.
5. We have also provided a Test Case execution option, where you can test the application against pre-configured high confidence and low confidence test case scenario.

Reference Screenshots

- Basic RAG with sample test questions

The screenshot displays the 'Financial Document Q&A' web application interface. On the left is a 'Configuration' sidebar with two sections: 'RAG Type' where 'Basic RAG' is selected with a red radio button, and 'Chunk Size' with a slider set to 700. Below this, 'Top K Chunks' is set to 2. The main content area is titled 'Financial Document Q&A' and 'Upload Financial Documents'. It shows a file upload area with a 'Browse files' button and a list of uploaded documents, including 'PDF Solutions Management Report Q3 24.pdf'. A green status bar indicates 'Processed 33 text chunks from 1 documents'. Below this is the 'Ask Questions' section with a text input field containing 'What is the revenue of Q3, 2024?'. The 'Answer' section shows the response: 'The revenue for Q3, 2024 is \$46.4M, up 11% over Q2 2024 and up 10% over Q3 2023.' The 'Confidence Score: 0.61' is displayed with an orange progress bar. A 'View Relevant Document Chunks' link is visible. The bottom section is 'Run Test Cases', which contains two columns of test queries. The first column has a query 'What is the Integrated yield ramp revenue for Q3, 2024?' with a 'Test High-Confidence Query' button, resulting in an answer and a confidence score of 0.61. The second column has a query 'Who is the latest Prime Minister of India?' with a 'Test Off-Topic Query' button, resulting in an answer stating the information is not available and a confidence score of 0.18.

- Advanced RAG with sample questions

Configuration

RAG Type

☐ Basic RAG

☒ Advanced RAG

Chunk Size

500 700 1000

Top K Chunks

1 2 5

Financial Document Q&A

Upload Financial Documents

Upload PDF files

Drag and drop files here

Limit: 200MB per file • PDF

Browse files

PDF Solutions Management Report Q3 24.pdf 478.6KB

×

Processed 33 text chunks from 1 documents

Ask Questions

Enter your question about the financial documents

What is the revenue of Q3, 2024?

Answer

The revenue of Q3, 2024 is \$46.4M, up 11% over Q2 2024 and up 10% over Q3 2023. The revenue of Analytics revenue is \$44.8M, up 17% over Q3 2024, and up 13% over Q3 2023. The revenue of Integrated yield ramp revenue is \$1.7M, down 53% over Q3 2024, and down 42% over Q3 2023. Non-GAAP gross margin is 77%, and non-GAAP operating cash flow is \$9.3M. Cash used for capital expenditures is \$4.6M. The revenue by geographic area is as follows: North America: \$35.3M, Asia Pacific: \$9.5M, Europe: \$3.9M, and Rest of World: \$1.0M. The key financial and operating metrics are as follows: Revenue by functional area: Analytics: \$44.8M, Integrated yield ramp: \$1.7M, and Other: \$1.3M. Revenue by geographic area: North America: \$35.3M, Asia Pacific: \$9.5M, Europe: \$3.9M, and Rest of World: \$1.0M. Non-GAAP gross margin: 77%, and non-GAAP operating cash flow: \$9.3M. Cash used for capital expenditures: \$4.6M. Revenue by geographic area: North America: \$35.3M, Asia Pacific: \$9.5M, Europe: \$3.9M, and Rest of World: \$1.0M. Non-GAAP gross margin: 77%, and non-GAAP operating cash flow: \$9.3M. Cash used for capital expenditures: \$4.6M.

Confidence Score: 0.89

View Relevant Document Chunks

▼

Run Test Cases

Execute Test Cases

Test Query

What is the Integrated yield ramp revenue for Q3, 2024?

Test High-Confidence Query

Answer: The Integrated yield ramp revenue for Q3, 2024 is \$1.7M, down 42% over Q3, 2023.

Confidence Score: 0.82

Test Query

Who is the latest Prime Minister of India?

Test Off-Topic Query

Answer: The latest Prime Minister of India is Narendra Modi.

Confidence Score: 0.57

Snippets of Basic and Advanced RAG implementations:

- Basic RAG implementation is available under BasicRAG class in the code.
- Advanced RAG implementation is available under AdvancedRAG class in the code.

Link to deployed App

[Huggingface spaces deployment](#)