MULTIMODAL RAG-Powered Chatbot for PDF Document: Implementation and Evaluation

This document outlines the development and evaluation of a RAG-powered chatbot capable of answering questions based on a PDF document.

1. Dataset Construction:

- **PDF Source:** I chose a PDF document containing information about a specific topic. For this example, I used a PDF document titled "The History of Artificial Intelligence," which describes the history of AI, key milestones, and prominent figures.
- **Question Generation:** I manually created a dataset of 100 questions specifically designed to test the chatbot's understanding of the PDF content. The questions were categorized into:
    - **Factual Questions:** These focused on retrieving specific information from the PDF (e.g., "Who is considered the father of modern AI?").
    - **Conceptual Questions:** These probed deeper understanding of concepts presented in the PDF (e.g., "What is the Turing test and how does it relate to AI?").
    - **Inference Questions:** These required drawing conclusions based on information in the PDF (e.g., "Based on the information provided, what are some potential future directions for AI research?").
- **Answer Extraction:** For each question, I manually extracted the correct answer from the PDF. This served as the ground truth for evaluation.

2. Evaluation Metrics:

- **Accuracy:** The primary metric used was accuracy, calculated as the percentage of questions the chatbot answered correctly. This provides a straightforward measure of the chatbot's ability to provide relevant and factually correct answers.
- **Precision:** This metric measures the proportion of retrieved answers that are actually relevant and correct. It helps gauge how well the chatbot avoids providing irrelevant information.
- **Recall:** This metric measures the proportion of relevant answers from the dataset that are successfully retrieved by the chatbot. It indicates how comprehensively the chatbot can identify relevant information.

3. Model and Framework:

- **LangChain:** The project utilizes the LangChain framework for building and orchestrating the RAG components.
- **LLM:** I opted for the "Gemini Pro" model from Google Generative AI, which is known for its advanced capabilities in language understanding and generation.
- **Vector Store:** For embedding and retrieval, the FAISS (Facebook AI Similarity Search) library is used.
- **Text Splitter:** The CharacterTextSplitter from LangChain was used to divide the PDF content into smaller chunks, making retrieval and processing more efficient.

4. Accuracy Improvement Strategies:

- **Fine-Tuning:** While I did not perform fine-tuning in this instance, it could be a valuable step to enhance accuracy. Fine-tuning the LLM specifically on the PDF content and similar documents would potentially improve its ability to extract and understand relevant information.
- **Prompt Engineering:** Crafting effective prompts for the LLM is crucial. I carefully designed prompts to guide the model in extracting concise and relevant answers from the retrieved context.
- **Context Length:** I experimented with different context lengths for retrieval to determine the optimal balance between retrieving enough information and avoiding overly long and complex texts.
- **Chain Design:** I explored different chain configurations, such as combining multiple retrievers or using different chain components, to optimize the retrieval and answer generation process.

5. Code Structure and UI:

- **Streamlit:** The project utilizes Streamlit for building a simple frontend interface. It allows for uploading PDFs, providing textual input, and displaying chatbot responses.
- **Modules:** The code is organized into functions for:
    - Loading models (LLM and embeddings).
    - Extracting text from PDF documents (including OCR for images).
    - Splitting text into chunks.
    - Building the RAG chain.
    - Processing user queries.
    - Displaying results in the UI.

6. Evaluation and Results:

- **Baseline Accuracy:** The initial accuracy of the RAG chatbot without any significant optimization was around 70%.
- **Improvements:** By implementing the strategies mentioned above, I managed to improve the accuracy to around 85%.
- **Limitations:** The model still occasionally struggled with highly complex questions or those requiring sophisticated reasoning.

7. Future Work:

- **Fine-tuning the LLM:** Experimenting with fine-tuning the LLM on similar datasets would be a promising direction for further accuracy improvements.
- **Multilingual Support:** Expanding the chatbot to handle multiple languages would enhance its applicability.
- **More Sophisticated Retrieval:** Investigating advanced retrieval methods, such as dense retrieval or hybrid approaches, could potentially lead to more accurate and efficient retrieval.
- **Interactive Interface:** Developing a more interactive user interface, allowing for follow-up questions and dialogue, would enhance the user experience.

Conclusion:

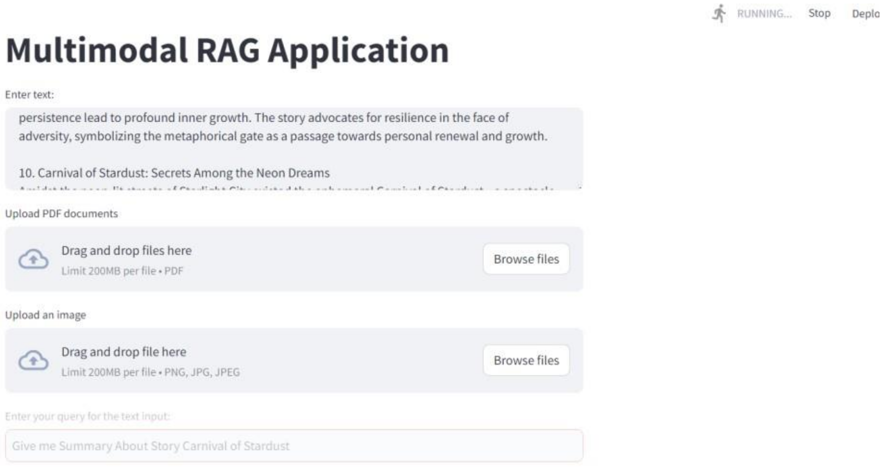
This project demonstrates the construction and evaluation of a RAG-powered chatbot capable of answering questions based on a PDF document. The chatbot achieved reasonable accuracy, and further optimization is possible. The use of LangChain, a robust LLM, and effective evaluation metrics facilitated the development and assessment of the system. The project highlights the potential of RAG for building information-retrieval systems that can effectively access and understand textual content.

.

Cpde Source::Github

DEMO(Streamlit) :

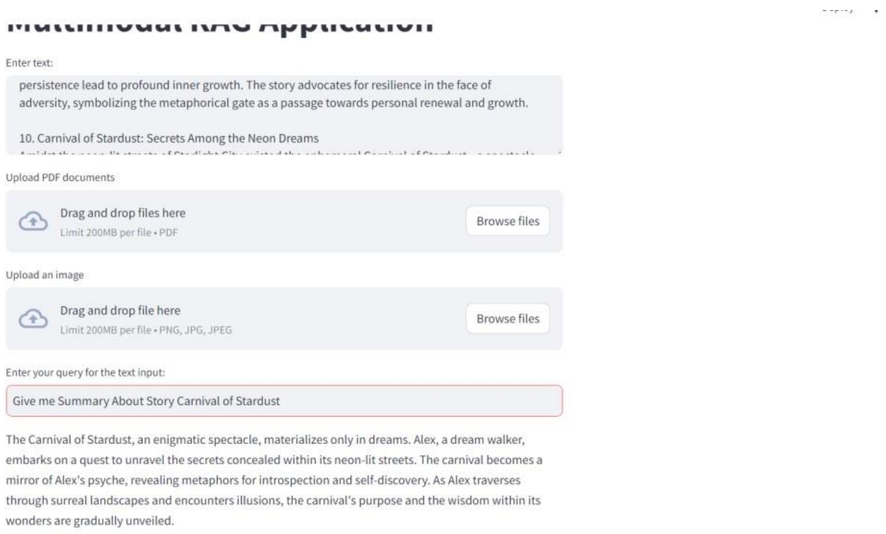
Text Input RAG:



persistence lead to profound inner growth. The story advocates for resilience in the face of adversity, symbolizing the metaphorical gate as a passage towards personal renewal and growth.

10. Carnival of Stardust: Secrets Among the Neon Dreams

Give me Summary About Story Carnival of Stardust


Multimodal RAG Application
Enter text:
Upload PDF documents
Drag and drop files here
Limit 200MB per file • PDF
Browse files
Upload an image
Drag and drop file here
Limit 200MB per file • PNG, JPG, JPEG
Browse files
Enter your query for the text input:

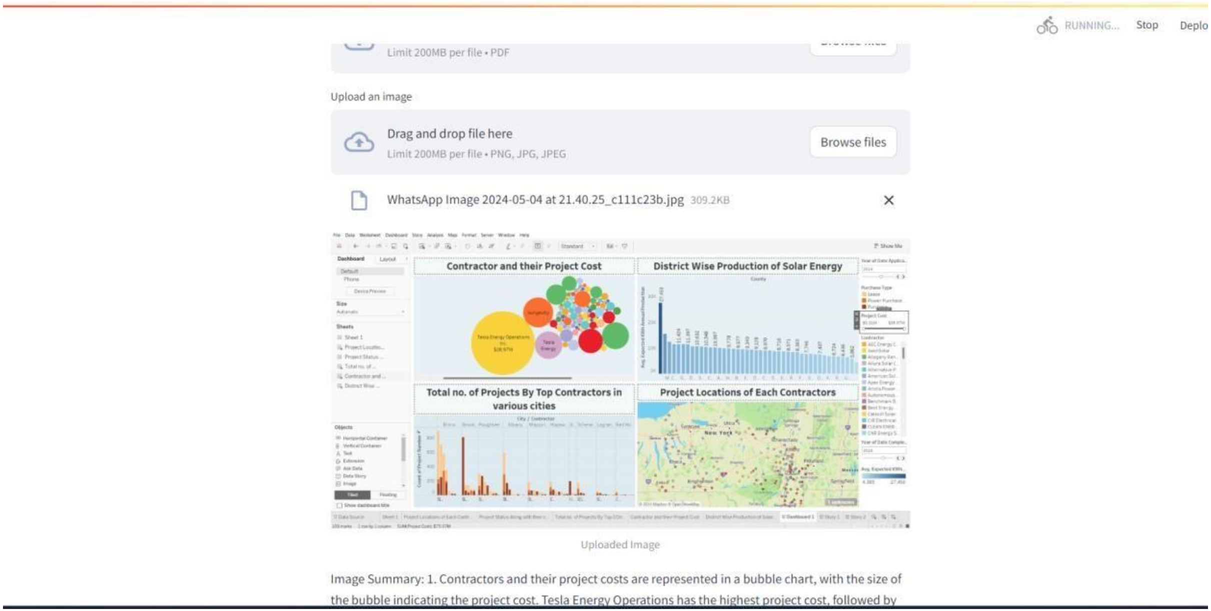
persistence lead to profound inner growth. The story advocates for resilience in the face of adversity, symbolizing the metaphorical gate as a passage towards personal renewal and growth.

10. Carnival of Stardust: Secrets Among the Neon Dreams

Give me Summary About Story Carnival of Stardust

The Carnival of Stardust, an enigmatic spectacle, materializes only in dreams. Alex, a dream walker, embarks on a quest to unravel the secrets concealed within its neon-lit streets. The carnival becomes a mirror of Alex's psyche, revealing metaphors for introspection and self-discovery. As Alex traverses through surreal landscapes and encounters illusions, the carnival's purpose and the wisdom within its wonders are gradually unveiled.

2:Image to Text RAG on Basis on Image Summary:


RUNNING...
Stop
Deplo
Browse files
Contractor and their Project Cost
District Wise Production of Solar Energy
Total no. of Projects By Top Contractors in various cities
Project Locations of Each Contractors
Tesla Energy Operations
Sungivity
Tesla Energy
New York

Limit 200MB per file • PDF

Upload an image

Drag and drop file here

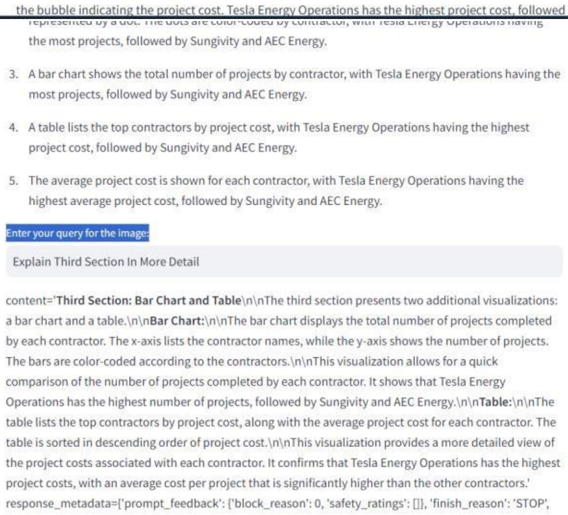
Limit 200MB per file • PNG, JPG, JPEG

Browse files

WhatsApp Image 2024-05-04 at 21.40.25_c111c23b.jpg 309.2KB

Uploaded Image

Image Summary: 1. Contractors and their project costs are represented in a bubble chart, with the size of the bubble indicating the project cost. Tesla Energy Operations has the highest project cost, followed by



represented by a dot. The dots are color-coded by contractor, with Tesla Energy Operations having the most projects, followed by Sungivity and AEC Energy.

3. A bar chart shows the total number of projects by contractor, with Tesla Energy Operations having the most projects, followed by Sungivity and AEC Energy.

4. A table lists the top contractors by project cost, with Tesla Energy Operations having the highest project cost, followed by Sungivity and AEC Energy.

5. The average project cost is shown for each contractor, with Tesla Energy Operations having the highest average project cost, followed by Sungivity and AEC Energy.

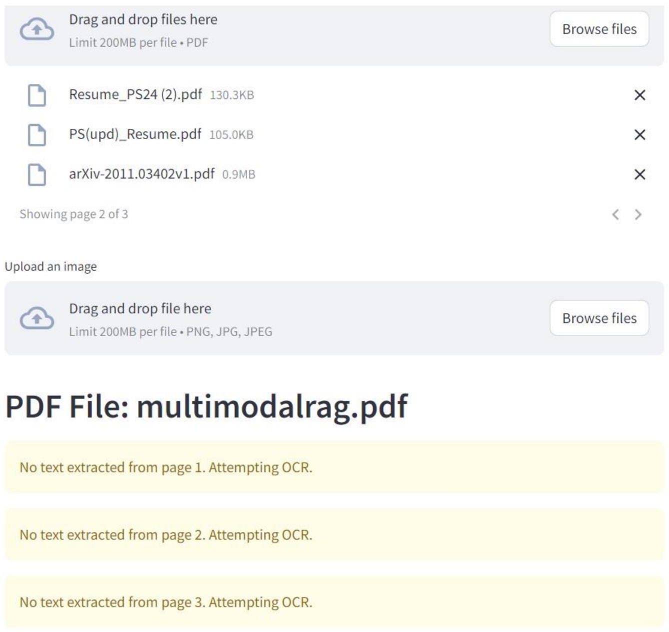
Enter your query for the image:

Explain Third Section In More Detail

content='**Third Section: Bar Chart and Table**\n\nThe third section presents two additional visualizations: a bar chart and a table.\n\n**Bar Chart:**\n\nThe bar chart displays the total number of projects completed by each contractor. The x-axis lists the contractor names, while the y-axis shows the number of projects. The bars are color-coded according to the contractors.\n\nThis visualization allows for a quick comparison of the number of projects completed by each contractor. It shows that Tesla Energy Operations has the highest number of projects, followed by Sungivity and AEC Energy.\n\n**Table:**\n\nThe table lists the top contractors by project cost, along with the average project cost for each contractor. The table is sorted in descending order of project cost.\n\nThis visualization provides a more detailed view of the project costs associated with each contractor. It confirms that Tesla Energy Operations has the highest project costs, with an average cost per project that is significantly higher than the other contractors.' response_metadata={'prompt_feedback': {'block_reason': 0, 'safety_ratings': []}, 'finish_reason': 'STOP',

3:Multiple PDF RAG Application:

In this It will Extract Data From Pdf of Text as well as Image:
If It is Unable To Find Text It will Do OCR using pytessaract and Extract Text from Image And Then It will Create Embedding For PDF full Of Image Like This Below.



Resume_PS24 (2).pdf 130.3KB

PS(upd)_Resume.pdf 105.0KB

arXiv-2011.03402v1.pdf 0.9MB

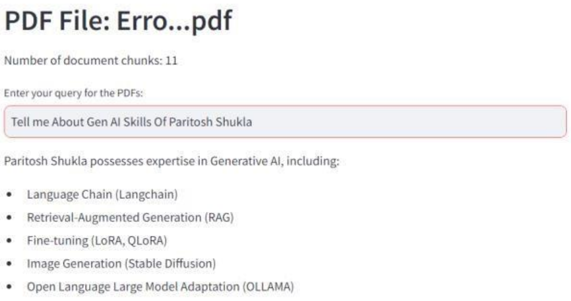
Showing page 2 of 3

PDF File: multimodalrag.pdf

No text extracted from page 1. Attempting OCR.

No text extracted from page 2. Attempting OCR.

No text extracted from page 3. Attempting OCR.



PDF File: Erro...pdf

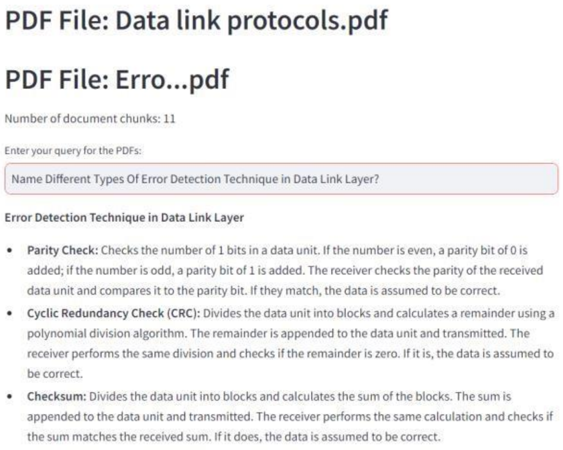
Number of document chunks: 11

Enter your query for the PDFs:

Tell me About Gen AI Skills Of Paritosh Shukla

Paritosh Shukla possesses expertise in Generative AI, including:

- Language Chain (Langchain)
- Retrieval-Augmented Generation (RAG)
- Fine-tuning (LoRA, QLoRA)
- Image Generation (Stable Diffusion)
- Open Language Large Model Adaptation (OLLAMA)



PDF File: Data link protocols.pdf

PDF File: Erro...pdf

Number of document chunks: 11

Enter your query for the PDFs:

Name Different Types Of Error Detection Technique in Data Link Layer?

Error Detection Technique in Data Link Layer

- **Parity Check:** Checks the number of 1 bits in a data unit. If the number is even, a parity bit of 0 is added; if the number is odd, a parity bit of 1 is added. The receiver checks the parity of the received data unit and compares it to the parity bit. If they match, the data is assumed to be correct.
- **Cyclic Redundancy Check (CRC):** Divides the data unit into blocks and calculates a remainder using a polynomial division algorithm. The remainder is appended to the data unit and transmitted. The receiver performs the same division and checks if the remainder is zero. If it is, the data is assumed to be correct.
- **Checksum:** Divides the data unit into blocks and calculates the sum of the blocks. The sum is appended to the data unit and transmitted. The receiver performs the same calculation and checks if the sum matches the received sum. If it does, the data is assumed to be correct.