Documentation for StratLytics Chatbot

Llama3-RAG-Pipeline: User Guide

Introduction

The Llama3-RAG-Pipeline is a powerful tool designed to help users perform retrieval-augmented generation using the Llama 3.2:3b model. This pipeline allows users to query documents and receive intelligent responses based on the provided data. Additionally, users can interact directly with the LLM (Large Language Model) to ask any general queries before uploading documents. However, once a document is uploaded, the LLM will only respond with respect to the content of the document. After removing the documents, users can revert to interacting with the LLM for any general queries again.

Features of the Llama3-RAG-Pipeline

1. **Document Upload**:
    - Users can upload documents like PDFs directly into the system.
    - The pipeline supports a variety of document types, including text-based PDFs and PDFs containing images and text.
2. **Multi-format Support**:
    - The pipeline is built to handle both text and text-image-based PDFs, allowing a wide range of use cases.
    - Users can test the model using documents that include:
        - Plain text
        - Images alongside text
        - Scanned documents
3. **Real-time Querying**:
    - After documents are uploaded and processed, users can ask questions in real-time, and the pipeline will provide answers based on the content of the uploaded documents.
    - This is particularly useful for applications like legal document reviews, research paper queries, or technical document understanding.
4. **Database Integration**:
    - The system connects to a local database that stores document metadata and content, allowing for efficient retrieval.
    - The database also enables tracking and management of uploaded documents.

How to Use the Pipeline

1. **Uploading Documents**:
    - To upload a document, navigate to the document upload section in the user interface.
    - Click the “Browse Files” button to select your PDF file from your system.

    - Once uploaded, the system will process the document and prepare it for querying.
2. **Querying the System**:
    - After the document has been successfully uploaded and processed, you can input questions or queries into the input section of the interface.
    - The model will use the document's content to generate a response.
    - Queries can be related to specific sections, keywords, or topics found in the document.
3. **Handling Large Documents**:
    - The system is optimized to handle large documents efficiently. However, there may be some constraints based on your system's memory and processing power.
    - For optimal performance, it is recommended to upload documents under 100 MB.

Constraints and Limitations

1. **File Size Limitations**:
    - While the system supports large documents, excessively large PDFs (over 100 MB) may cause slowdowns or failures during processing. It's advised to keep files within this size range.
2. **System Requirements**:
    - The pipeline requires a system with adequate resources. For optimal performance, a minimum of 16 GB of RAM is recommended. A dedicated GPU will further enhance processing speeds, especially for image-heavy documents.
3. **Image Processing**:
    - Although the pipeline can handle PDFs with images, it is primarily optimized for text retrieval. Image-based content may not always be fully interpreted unless accompanied by recognizable text.

Best Practices

- **Use High-Quality PDFs**: Ensure the documents you upload are clear and well-formatted. Scanned documents with poor resolution may hinder the system's ability to accurately retrieve information.
- **Query Specifics**: When querying the system, try to use precise language. The more specific your query, the more accurate the response from the pipeline.
- **Testing Multiple Document Types**: Users can experiment with different document formats to understand how the system performs with various input types. This includes testing PDFs that contain text, images, or a combination of both.