Challenges Faced and Solutions

Integrating the LLM with Preprocessed Text and Handling Large Token Counts

A significant challenge was managing the large amount of tokens generated from the PDF content. The Meta Llama3 model supports up to 8192 tokens. Ensuring the token limit was not exceeded was crucial.

Solution:

To tackle this, I used the GPT-2 tokenizer to count tokens in the preprocessed text. By encoding the text and calculating the number of tokens, I ensured the token count stayed within the acceptable limit. For cases where the token count exceeded the limit, I truncated the text, reserving tokens for user queries and system messages. This approach allowed smooth integration with the LLM.

Potential Improvements

Advanced PDF Structure Recognition

Implementing advanced PDF structure recognition to differentiate headers and paragraphs could enhance response accuracy.

Handling Multiple PDFs

Supporting multiple PDF documents in a single session would improve usability and flexibility.

Highlighting Relevant Text

Highlighting the relevant parts of the PDF text from which responses are generated would improve transparency and user insight.

Scalability Considerations

Caching and Resource Utilization

Implementing advanced caching mechanisms would reduce redundant API calls and improve response times, enhancing scalability.

Load Balancing and Distributed Processing

For large-scale deployment, load balancing and distributed processing would ensure the application handles high traffic and multiple requests efficiently.

Enhanced Error Handling

Improving error handling and user feedback would enhance the user experience and reliability. Comprehensive logging and monitoring can help identify and resolve issues promptly.