Sample Project Report

Despite the efforts made to solve this challenge, several key insights and improvements can further enhance the accuracy of the solution.

Here is a step-by-step account of the approach and challenges faced:

Project Structure Preparation: I first prepared the complete project structure, ensuring all necessary components were in place, including modules for PDF extraction, question matching, LLM integration, and Slack communication.

Implementation and Conclusion: I started the implementation by carefully integrating Slack, a new concept for me. Initially, Slack integration was smooth, with some outputs appearing correctly on both the terminal and Slack. However, I noticed that some outputs were incorrect.

Challenges and Solutions:

1. **Model Limitations:** During implementation, I noticed that the model had token limits, impacting the output's accuracy. I attempted to rectify this by breaking the text into smaller chunks but faced challenges with context continuity.
2. **Dummy Data Testing:** To verify the system, I tested with dummy data, and the output was correct. This highlighted the importance of token limits and the PDF's context.
3. **PDF Content Structure:** The initial content of the PDF was not in a proper format, impacting the output. After restructuring the content, I received better results.
4. **API Limits:** I encountered API limit issues during testing, restricting further experimentation.

Alternative Approach: To solve the problem, I switched to a Retrieval-Augmented Generation (RAG) implementation using open-source models. This approach yielded correct outputs. Due to time constraints, I had to work hard to refactor the existing approach to the new one, requiring additional time for complete implementation.

Enhancing Accuracy: Summary

Despite the efforts to solve the challenge, the following steps can significantly improve accuracy:

1. Robust text extraction and preprocessing.

2. Efficient handling of large documents through chunking and context management.
3. Precise question matching using both exact and semantic techniques.
4. Optimized use of LLMs with fallback mechanisms.
5. Hybrid and RAG approaches for enhanced retrieval and generation.
6. Continuous learning and performance optimization for long-term improvements.