LLM Challenge: QA Over Documents with OCR Integration

Objective:
The objective of this challenge is to train a large language model capable of performing Question-Answering (QA) tasks over documents. The challenge will focus on developing a Python-based solution that incorporates Optical Character Recognition (OCR) for text extraction from documents. You will have 6 days to complete the challenge and deliver your solution, along with a deck and a remote/onsite presentation.

Challenge Details:

Dataset:

You will be provided with a real estate lease PDF as the dataset for this challenge.

Task:

Your task is to build a Python-based system that performs QA over the given documents using a large language model, such as GPT-3.5 or an equivalent. The solution should integrate OCR functionality to extract text from the documents for analysis and QA.

Technical Requirements:

1. OCR Integration:
    a. Utilize a suitable OCR library or API, such as Tesseract, Pytesseract, PyPDF2 or an equivalent, for text extraction from the documents/pdfs.
    b. Implement the OCR functionality in your Python solution to extract the text accurately.
2. Language Model Training:
    a. Train or fine-tune a large language model using appropriate techniques, such as transfer learning or reinforcement learning.
    b. The trained language model should be capable of answering questions based on the content of the documents.
3. Documentation and Deployment:
    a. Provide clear documentation and comments within your Python codebase to facilitate understanding and future development.
    b. Include instructions and dependencies necessary for easy setup and deployment of your solution.

Deliverables:

1. Python Codebase:
    a. Submit a completed Python codebase that implements the solution, including all necessary functionalities.

    b. Ensure that the code is well-structured, readable, and follows best practices.
2. Preprocessing Scripts (if required):
    a. Include any preprocessing scripts that are necessary for data preparation or OCR text extraction.
3. Deck:
    a. Create a deck summarizing your technical approach, the challenges you faced during the development process, and key findings from your solution.
    b. The deck should provide an overview of your system's architecture and highlight its performance and capabilities.
4. Onsite Presentation:
    a. Prepare an onsite presentation to showcase your solution.
    b. During the presentation, provide a live demonstration of your system's functionality, including question-answering using the trained language model.
    c. Discuss the architecture, performance, and potential future improvements of your solution.
    d. Bonus point: What are the potential commercial applications for OCR + LLM solutions? How would you envision commercializing these solutions, including identifying the customer persona and the specific workflows they address?

Best of luck!