Case Study : NLP Engineer

Objective

Leverage NLP and machine learning techniques to extract, process, and analyze financial data from a P&L statement in PDF format. Use a Large Language Model to generate forecasting scenarios based on historical financial data.

Part 1: Data Preparation and Understanding (± 2h)

1. **PDF Data Extraction**:
    - Extract P&L knowledge from a provided PDF document. Assess and choose the appropriate tools for PDF parsing
    - Discuss the challenges and limitations encountered during PDF data extraction.
2. **Tabular Data Manipulation**:
    - Convert extracted data into a structured tabular format suitable for analysis. Include steps for data cleaning and validation.
    - Prepare the data for analysis by organizing it into quarterly financial metrics over multiple years.

Part 2: Forecasting with LLM (± 4h)

1. **RAG Model Creation**:
    - Briefly describe the concept of a Retrieval-Augmented Generation model and its relevance to forecasting tasks.
    - Outline the steps to integrate RAG with a LLM, focusing on how it can enhance forecasting by leveraging both retrieved information and generative capabilities.
2. **LLM Forecasting Scenarios**:
    - Utilize a LLM to generate forecasting scenarios for the P&L data. Explain how you format the input to the LLM and interpret its output.
    - Explore different prompts or configurations to generate a range of forecasting scenarios (e.g., optimistic, pessimistic, most likely).

Part 3: Evaluation and Presentation (± 2h)

1. **Forecast Evaluation**:
    - Discuss methods to evaluate the plausibility and accuracy of the LLM-generated forecasts. Consider comparing against baseline models or historical trends.
    - Reflect on the limitations and potential biases of using LLMs for financial forecasting.
2. **Technical Documentation**:
    - Provide well-documented code for all parts of the exercise, including data extraction, manipulation, and forecasting.
    - Summarize the approach, challenges, and key findings in a concise report.

Submission Guidelines

- **Code**: Submit your code in a Python notebook format (Jupyter Notebook, Google Colab). Ensure your code is clean, well-commented, and organized.
- **Report**: Include a PDF or Markdown report summarizing your methodology, challenges faced, solutions implemented, and a discussion on the forecasting scenarios generated by the LLM.

Evaluation Criteria

- **Technical Skills**: Ability to extract and manipulate data, create and integrate RAG models, and generate meaningful forecasts using LLM.
- **Problem-Solving**: Ingenuity in overcoming extraction and data manipulation challenges, as well as creativity in generating and evaluating forecasting scenarios.
- **Code Quality**: Clarity, organization, and documentation of code.
- **Analytical Thinking**: Insightfulness in interpreting LLM outputs and evaluating forecasting scenarios.