AI Developer Take-Home Assignment: Document Processing with LLMs

Objective

The goal of this take-home exercise is to build a pipeline that uses Large Language Models (LLMs) to analyze legal contracts, extract specific clauses and summarize key terms. This exercise is designed to evaluate your skills in document processing, prompt design and LLM-based information extraction.

Dataset

- CUAD – Contract Understanding Atticus Dataset
  - Publicly available at: https://www.atticusprojectai.org/cuad
  - Contains over 13,000 legal clauses from 510 contracts, annotated with 41 types of clauses.
  - For this assignment you should use a smaller subset 50 contracts.

Task

1. Data Loading & Preprocessing

- Load a subset of contracts from CUAD.
- Extract the full contract text from PDF files.
- Normalize text.

2. LLM-Powered Information Extraction & Summarization

- Part A – Clause Extraction:
  - Use an LLM (API-based or open-source) to identify and extract key clauses:
    - Termination conditions.
    - Confidentiality clauses.
    - Liability clauses.
- Part B – Contract Summary:
  - Generate a concise 100–150 word summary of the contract highlighting:
    - Purpose of the agreement.
    - Key obligations of each party.
    - Notable risks or penalties.

Expected Deliverables

- Code: Python script or notebook implementing the document processing pipeline.
- Output File: CSV or JSON with columns like: [contract_id, summary, termination_clause, confidentiality_clause, liability_clause].
- Readme: Instructions on how to run the code, approach explanation including flow diagram of the solution.
- Bonus (Optional):
    - Implement semantic search over clauses using embeddings.
    - Experiment with few-shot examples to improve clause extraction.

Upload the above to a GitHub (or GitLab/Bitbucket) repository and share the link with us by replying to the email containing this assignment. Please ensure the repository is public so we can access it and its contents do not have any mention of Uptitude.

Evaluation Criteria

- Accuracy: Quality of clause extraction and summaries.
- Code Quality: Readability, modularity and documentation.
- LLM Utilization: Efficient use of LLMs, prompt engineering and handling large text.
- Reproducibility: Easy setup and execution.
- Creativity: Additional features, optimizations or comparisons between models.