Data Quality Task for QA Analyst/Engineer

Objective:
Clean and standardize entity names (sponsoring government agencies and awarded companies) from SAM.gov (USA) and the E-Procurement Government of India databases.

- https://sam.gov/data-services
- https://eprocure.gov.in/eprocure/app
    - Please do in depth research about the above sources to find the right url to find active tenders and historic tender records dataset
    - Within this find the "entities" i.e. the sponsoring government. Here candidates need to only focus on two key attributes
      A. the attribute that mentions the sponsoring government agency like the Department of Defense for SAM. gov or the Ministry of Road and Highway for E Procurement India i.e. the government agency that is giving the contract.
      B. The supplier or the company that won the contract or was awarded the contract, so it could be Tata Steel or Bouygues construction. Now note the bigger focus is on the company and supplier name. Say Tata Groups can have multiple entities for example Tata Steel and Tata Steel might be mentioned with the Tata Steel Europe or Tata Steel USA or other such names. The objective is to have consistent and reliable final names that you are supposed to clean up.
    - Candidate can show
      A. Manual ways they fixed a small sub sample of 100 records.
      B. if they can provide a basis of automation of this using LLMs or python operations and execute on his, they will get extra points.

Part 1: Data Cleaning and Standardization
Task: Manually clean and standardize a subset (100 records) of entity names from the provided datasets.
Focus: Pay special attention to the company/supplier names, ensuring consistency and reliability in the final names. For example, different iterations of "Tata Steel" should be standardized.

Part 2: Automation Proposal and Script Development
Task: Develop a basic automation script or method using Python and language models (OpenAI API, Llama2, etc.) to standardize entity names in the datasets.

Expectation:
Provide a working script or a detailed proposal for automating the standardization process. Demonstrate the script's effectiveness on a subset of the data.

Part 3: Scalability and Production Readiness
Task: Document how the proposed method can be scaled and implemented in a production environment.

Details:
Include considerations for continuous data updating and processing large volumes of data. Explain how the method adheres to data quality and standards.

Evaluation Criteria
Standards & Quality: Accuracy and consistency in the final cleaned and standardized data.
Scalability: The potential of the method to handle large datasets efficiently in a production environment.
Documentation: Clarity and comprehensiveness of the documentation, including reasoning for scaling the solution.

Deliverables
Candidates should submit a Google Drive folder containing:

1. Python Scripts: Code for data cleaning and standardization.
2. Sample Data: Original and final cleaned datasets (100 records minimum).

Documentation:
- Detailed explanation of the methods used.
- Plan for scaling the solution to a production environment.

Additional Task Details
- Data Sources: Use SAM.gov and the E-Procurement Government of India for sourcing data.
- Tools and Languages: Utilize Python, OpenAI API, Llama2, or other open-source language models.