Challenge Problem: Rapid Prototyping of LLM-Enabled Materials Data Ingestion as part of datascribe.cloud platform

Challenge Overview:

As part of the selection process for this project, students are tasked with a one-week challenge problem to rapidly prototype a basic version of the proposed web application (The explanation of the main project is provided in page 3). This challenge is designed to test the student's ability to quickly understand the problem, apply their technical skills, and deliver a functional proof of concept within a constrained time frame.

Problem Description:

Students are required to develop a simplified version of the **Materials Insight** application, focusing on the core functionality of ingesting materials science data and utilizing an LLM (Large Language Model) to perform basic data extraction and summarization. The prototype should be capable of the following:

1. **Data Ingestion**: Implement functionality to allow users to enter prompt questions such as:
    a. (i) "Can you list all material properties of AZ31 alloy?",
    b. (ii) "Can you provide a stress-strain curve for the AZ31 alloy tensile test at room temperature?",
    c. (iii) "Can you plot the stress-strain curve for SS316 alloy and compare it with AZ31?"
    d. Store these queries and their responses in a SQL database, specifically linking them to the relevant alloy (e.g., AZ31, SS316). Ensure the system minimizes hallucination by cross-referencing reliable data sources before generating responses.
2. **LLM Integration**: Use an LLM to extract key material properties and summarize the findings on the screen in a visually nice manner.
3. **User Interface**: Develop a simple web interface where users can enter these prompts and view the result in some nice front-end; allow users to download the tables, or save the graphs.
4. **Output**: Display the extracted data in a structured format (e.g., table, list, graphs) on the web interface.

Technologies and Tools:

- Frontend: React.js, CSS, JavaScript

- Backend: Node.js
- Database: PostgreSQL for storing material information
- LLM Integration: OpenAI GPT-4 API or similar for language model capabilities, Langchain, Langraph and/or other LLM-related tools.
- Cloud Platform: Google Cloud for deployment
- Version Control: Git/GitHub for code management

Submission Requirements:

- **Video Presentation**: Teams/Students must submit a video presentation (10-15 minutes) explaining their development process, design choices, and how they integrated the LLM. The presentation should include a live demonstration of the prototype in action.
- **Code Submission**: The codebase for the prototype should be submitted via a GitHub repository, along with basic documentation explaining how to set up and run the prototype.
- **Prototype**: The working prototype should be deployed on a public cloud platform (e.g., Google Cloud) or accessible via a local server with clear instructions on accessing it.

Evaluation Criteria:

- **Functionality**: The prototype's ability to correctly ingest data and extract relevant information using the LLM.
- **Creativity and Innovation**: The team's approach to solving the problem, including any unique features or optimizations implemented.
- **Usability**: The simplicity and intuitiveness of the user interface and overall user experience.
- **Presentation Quality**: Clarity, professionalism, and thoroughness of the video presentation, including how well the team explains their process and decisions.
- **Code Quality**: The organization, readability, and documentation of the submitted code.

Challenge Timeline:

- **Challenge Start Date**: Sep 10, 2024
- **Submission Deadline**: Sep 17, 2024, 12 midnight
- **Evaluation and Selection**: The evaluation of submissions will be completed within 3 days after the submission deadline. Selected individuals will be notified and invited to participate in the full project.

This challenge not only assesses the students' technical abilities and creativity but also their time management and ability to deliver under pressure. The results of this challenge will play a significant role in the selection process for the capstone project.

Note: Integration to the DataScribe.cloud platform is not required as part of the challenge problem.

For questions/inquiries, please reach out to Dr. Vahid Attari (attari.v@tamu.edu) or Mrinalini Mulukutla(mrinalini.mulukutla@tamu.edu).

Main Project: Development and Integration of a Web Application Utilizing Large Language Models for Materials Science Data Ingestion and Analysis

Project Duration:

Two Semesters (Fall 2024 – Spring 2025)

Project Team:

- **Students:** 1-3 Undergraduate Students in Computer Science/Software Engineering and Materials Science/Engineering.
- **Advisors:**
    - Primary Advisor: Dr. Vahid Attari, Department of Materials Science and Engineering
    - Co-Advisor: Dr. Raymundo Arroyave, Department of Materials Science and Engineering

Problem Statement:

The goal of this project is to develop a web-based application, **Materials Insight**, that leverages Large Language Models (LLMs) such as GPT-4 to assist researchers and engineers in the field of materials science. The application will provide tools for data ingestion, organization, analysis, and documentation of materials science data, enhancing the efficiency and accuracy of materials research. The project involves designing and implementing an LLM system for the datascribe.cloud online platform for vector-based searches, enhancing the search and retrieval capabilities for stored metadata. The system should also associate the search results with relevant metadata stored in SQL database, allowing for comprehensive exploration and retrieval of related information.

Objectives:

1. **Develop a Web Interface**: Integrate an intuitive web-based user interface in datascribe.cloud website where researchers can query materials science data using the LLM interface in various formats (e.g., text, tables, experimental results).
2. **Integrate LLM Capabilities**: Implement LLM functionalities to:
    - Extract and organize key material properties and experimental results from web and unstructured text sources.
    - Assist in summarizing, tabulating, and graphing data.
    - Provide insights and interpretations of the data.
3. **Automate Data Processing**: Automate routine data processing tasks such as unit conversions, calculations of derived properties, and standardization of data formats.
4. **Literature Review and Reference Management**: Develop features to assist users in conducting automated literature reviews and managing materials data, and references within the application.
5. **Generate Reports and Documentation**: Enable the generation of reports, formatted LaTeX documents, and presentations summarizing the materials data and findings.
6. **Deploy and Test**: Deploy the application on a cloud platform and conduct user testing with materials science students and faculty to refine the application.

Project Deliverables:

1. **Fully Functional Web Application:** A deployed web application accessible to users within the university, featuring:
    - Data input and ingestion capabilities.
    - LLM-driven data extraction, interpretation, and summarization tools.
    - Automated data processing and report generation features.
2. **Documentation**: Comprehensive user documentation, including a user guide and technical documentation detailing the architecture and functionality of the application.
3. **Final Report and Presentation**: A detailed final report summarizing the project, the methodologies employed, the outcomes, and future work recommendations. This will be accompanied by a final presentation to the department.

Technologies and Tools:

- Frontend: React.js, CSS, JavaScript
- Backend: Node.js
- Database: PostgreSQL for storing user data and material information

- LLM Integration: OpenAI GPT-4 API or similar for language model capabilities, LangChain, Langraph and/or other LLM-related tools.
- Cloud Platform: Google Cloud for deployment
- Version Control: Git/GitHub for code management

Expected Learning Outcomes:

- Technical Skills: Gain experience in full-stack web development, API integration, and cloud deployment.
- Project Management: Develop project planning, time management, and teamwork skills.
- Materials Science Knowledge: Understand the specific data needs and challenges in materials science research.
- AI/ML Application: Learn how to apply AI/ML models to solve domain-specific problems in materials science.

This project provides a unique opportunity to develop a practical, real-world application that bridges computer science and materials science, utilizing cutting-edge AI technology. The resulting application will not only aid researchers but also provide the student team with valuable interdisciplinary experience.