Assignment on "Big Data Integration"

Due: March 31, 2025

Drawing inspiration from the topics covered in class, this assignment challenges the students to design, implement, and evaluate an original **schema alignment** solution using **Large Language Models** (LLMs). The implementation will be tested on the BIRD benchmark, a text-to-SQL dataset. BIRD contains over 12,751 unique question-SQL pairs with schema annotations (the ground truth) and 95 datasets that cover various domains, such as blockchain, hockey, healthcare and education.

Task Description

You will:

- develop a system that, for each natural language question in the BIRD benchmark, identifies the Source Tables (STs) that contain data relevant to answering the question;
- evaluate the results against the BIRD ground truth computing the overall recall, precision, and F1-score for detected STs.

Your solution must rely primarily on LLMs (e.g., GPT, Llama, or open-source alternatives) to automate these steps. You are encouraged to experiment with prompting strategies or hybrid approaches combining LLMs with traditional methods (e.g., based on similarity).

Deliverables

1. Technical Report (PDF):
   - Problem Analysis: Challenges in schema alignment and LLM applicability.
   - Methodology: Detailed description of your approach (e.g., LLM prompts, algorithms).
   - Results: Quantitative evaluation (tables/graphs of metrics) and qualitative examples.
   - GitHub Link: Clearly visible URL to your code repository.
2. GitHub Repository:
   - Code: Fully documented, executable implementation.
   - Data: Instructions to reproduce experiments on BIRD (subset or full benchmark).

Guidelines

- Work independently, but you may use any (open source) tool that you like;
- Deadline: Submit the report on Moodle by March 31, 2025;
- Support: you can contact Divesh (divesh@research.att.com) for questions/hints/suggestions about the assignment.

Tips for Success

- Start early — experimenting with LLMs and large benchmarks can be time-intensive.
- Use the BIRD validation set for iterative testing before final evaluation.
- Analyze failure cases to refine your solution (e.g., ambiguous attribute names).

Good luck!