AI Engineering Assignment: Similar Questions Evaluation Framework

Background

At Lytmus AI we help students solve Physics, Math, and Chemistry problems. To avoid hallucinations and provide contextual solutions, we feed in previously solved similar questions inside the prompt and let them learn better to build great solutions. However, we need to understand:

1. **Quality Assessment**: Are the retrieved similar questions actually good representatives of the input question?
2. **Impact Measurement**: How much do these similar questions improve solution quality compared to solving without them?

Your Task

Build a comprehensive **LLM-as-a-Judge framework** that evaluates both the relevance of similar questions and measures their impact on solution quality.

What You'll Receive

- **Dataset**: 500 questions with their associated "similar questions" and “similarity score” (production data)
- **Subject Distribution**: Mix of Physics, Math, and Chemistry problems
- **Solution Approach:** These similar questions will also have the summarized approach with which they were solved

Requirements

Core Deliverables

1. Similar Question Relevance Evaluator

Design an evaluation system that assesses how well "similar questions" represent the input question across multiple dimensions:

- **Conceptual Similarity**: Do they test the same underlying concepts/principles?
- **Structural Similarity**: Are the problem structures analogous?
- **Difficulty Alignment**: Is the difficulty level appropriate?
- **Solution Approach Transferability**: Can the solution method be meaningfully applied?

2. Solution Builder

Create a framework to build the solutions of the given question for the below cases:

- **Without Similar Questions**:
- **With Similar Questions And their summarized approaches**

We would like to understand how you design prompts and make use of similar question approaches

3. Comparative Analysis System

Build a system that can:

- Compare solutions generated with vs. without similar questions
- Quantify improvement metrics
- Identify when similar questions help vs. hurt performance
- Generate actionable insights for prompt engineering

Output Format

- Please share the work either in any executable files or python notebooks
- BEST : Initialize a repo and build projects there. I would love to see the commit history
- Highlight the quantifiable metrics on 1 and 3
- You SHOULD write code as if you are going to deploy it on production

Bonus Challenges (Optional)

- **Agent-Based Approach**: Use multi-agent systems for more robust evaluation
- **Meta Prompting :** Improve your initial solution generation prompts with the feedback you obtain
- **A/B Testing Framework**: Design experiments for production deployment

Questions?

Feel free to reach out for clarifications, but part of the assignment is making reasonable assumptions and documenting your decision-making process.
