CMPUT 692 - Topics in Data Management with LLMs

Assignment 1

Due: Oct 1, 2025 at 23:55

Individual Assignment

Objective

The goal of this assignment is to investigate challenges in applying large language models (LLMs) to data management tasks, with a focus on *text-to-SQL generation*. The task involves translating natural language queries into SQL statements given a database schema. A central component of this assignment is the rigorous *evaluation of generated queries* using standard benchmark tools.

Tasks

1. **Query Selection** Select at least 40 queries from the development set of the BIRD dataset. Your selection must:
   - Cover a variety of query lengths and difficulty levels.
   - Span at least three distinct databases.
2. **Query Generation** Use programmatic prompting to generate SQL statements for your selected queries. You may use either closed-weight models (e.g., Gemini, GPT-based) or open-weight models (e.g., Qwen-2.5-Coder-Instruct).
3. **Evaluation** Assess the generated queries using the official evaluation scripts provided with the BIRD benchmark.
4. **Analysis** Analyze the evaluation results. Discuss:
   - Overall accuracy and error rates.
   - Common sources of failure.
   - Observed trends (e.g., by query length, schema complexity, or model choice).
5. **Submission** Submit a single tarball containing:

(a) Your code, or a link to a GitHub repository.

(b) The list of selected queries.

(c) The SQL statements generated for each query.

(d) A written report that includes:

- Evaluation results (outputs from the official scripts).
- Your analysis and discussion of findings.

Important: Any external sources you consult (e.g., papers, repositories, software libraries, or LLMs used) must be explicitly cited in your report.

Constraints

1. Each student must work with a unique set of queries. After making your selection, post your dataset and query IDs on the course forum to avoid duplication.

2. Query selection must reflect diversity across databases and include non-trivial queries.

Rubric (10 points total)

1. **Query Selection (1 point)** Meets requirements for query count, diversity (length, difficulty, database), and uniqueness.

2. **Implementation & Query Generation (2.5 points)** Uses programmatic prompting (e.g., scripts or notebooks). Configures LLM(s) with reproducible results. Clearly documents prompting approach and parameter settings.

3. **Evaluation (2.5 points)** Correctly applies official evaluation scripts. Accurately reports metrics (e.g., execution accuracy, exact match). Appropriately handles errors and failed cases.

4. **Analysis (2 points)** Provides insightful discussion of performance and error cases. Identifies trends across query types, schema complexity, or difficulty levels. Presents analysis in a clear, coherent, and well-organized manner.

5. **Comparison (2 points)** Compares results with at least one published model on the leaderboard. Highlights strengths and weaknesses of your approach relative to prior work.