

WTL AI Trial Project

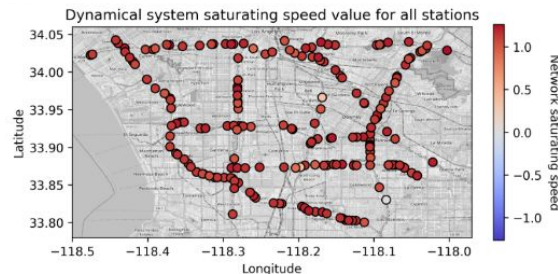
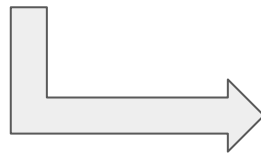
by Michael Lui

About Me

- Recent college graduate of UCLA, go bruins, majored in mathematics of computation (math and computer science)
- Interested in full time roles for data science / Machine Learning
- Have experience (as showcased in this presentation) in DS/ ML
- Extremely thankful for this opportunity to interview with the WTL team



Check out another of my DS/ML projects on my resume/github in examining and predicting LA highway traffic



$$\mathcal{L}(b, W, c) = \sum_{t=0}^{\cdot} \|\hat{\mathbf{v}}_{b, W, c}^{(t+1)} - \mathbf{v}^{(t+1)}\|^2$$

Project program Link:

https://github.com/max09lui/WTL-AI_Trial_Project/tree/main

read README file, to recreate environment to replicate results showcased in this presentation

note: **OpenAI API key** (not free) and **mySQL database** (can be free) is needed to run the program

Project Objective

Objective: “Develop a system where an LLM (Large Language Model) reads an SQL database tracking employee activities and generates natural language summaries based on natural language queries”

Expanding on objective

- Any mySQL database summary has a corresponding mySQL query command to it
- **Updated Objective:** “Develop a system where an LLM (Large Language Model) can generate mySQL query commands to generate metadata summaries on a mySQL database based on user prompts”

Project Outline:

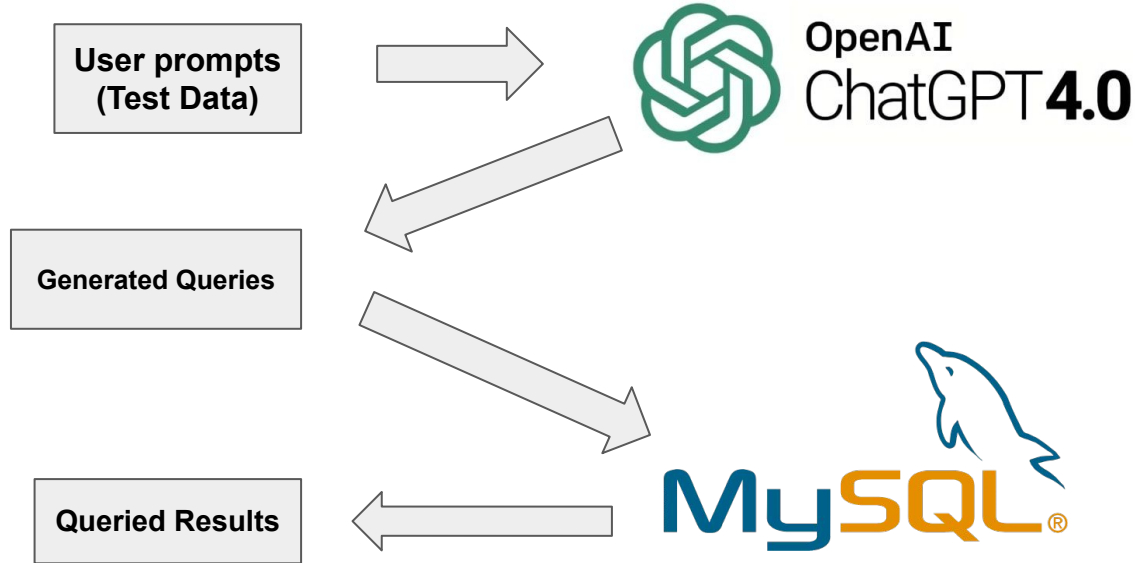
Database:



LLM:



Server:



mySQL Database Creation and Population:

Employee ID: Unique identifier for each employee.

• Employee Name (Added in correspondence of testing data)

- Week Number: (1-10)
- Number of Meetings: Integer representing the number of meetings attended.
- Total Sales (RMB): Decimal representing the total sales value in RMB.
- Hours Worked: Decimal representing the total hours worked.
- Activities: Text field describing activities such as preparation for meetings, sales strategies, challenges faced, and solutions implemented.
- Department: Text field indicating the employee's department (e.g., Sales, Marketing, Product Development, Finance, IT).
- Hire Date: Date indicating when the employee was hired.
- Email Address: Employee's email address.
- Job Title: Employee's job title (e.g., Sales Manager, Data Analyst, Marketing Specialist).

```
-- Employee table (basic info)
CREATE TABLE employees (
  Employee_ID INT PRIMARY KEY,
  Name VARCHAR(20),
  Department VARCHAR(20),
  Hire_Date TIMESTAMP,
  Email_Address VARCHAR(50),
  Job_Title VARCHAR(50)
);

-- EmployeeWeeks table (employee-specific weekly data)
CREATE TABLE employee_weeks (
  Employee_ID INT,
  Week_Num INT CHECK (Week_Num BETWEEN 1 AND 10),
  Num_Meets INT,
  Total_Sales DECIMAL(10,2),
  Hours_Worked INT,
  Activities VARCHAR(50),

  PRIMARY KEY (Employee_ID, Week_Num), -- Composite key
  FOREIGN KEY (Employee_ID) REFERENCES employees(Employee_ID)
);
```

LLM System Prompt

```
export const LLM_PROMPT=`With the SQL table of:
```

```
-- Employee table (basic info)
```

```
CREATE TABLE employees (
```

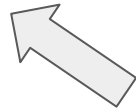
```
  Employee_ID INT PRIMARY KEY,
```

```
  Name VARCHAR(20),
```



mySQL table metadata
(partially omitted to reduce redundancy)

```
, generate SQL commands only that correspond by question given by the user, return as a the SQL command only with no line breaks and no SQL label`
```



**specific instructions on how to reply to user
prompt to get an SQL query command**

Test Data (LLM User Prompts)

Filtered Data into clean/unclean questions, clean meaning there is a correct corresponding answer in the database

$$\{\text{Training Data}\} = \{1, \dots, 20\}$$

$$\{\text{Filtered Data}\} = \{\text{Training Data}\} \setminus \{3, 11, 14, 15, 18, 19\}$$

motivation: removed questions have no distinct relevancy to the mySQL tables' metadata, which could lead to ambiguity in correctness of the response as well as likely hallucination from the LLM when creating a mySQL query command

(Unclean data will be added back to training data later on)

Examples of Issues in Unclean Training Data

3: What was the sales revenue of 'Wei Zhang' for the week starting on '2024-08-28'?

This question has no correspondence based on the metadata from the mySQL tables

19: Who achieved the highest sales revenue in a single week, and when?

Is this question asking for highest revenue in the sales department or highest sales (metadata) in total revenue (not given)

Benchmark 1:

In first created program, correctly queried 100% of filtered training data

Minor Issues: small variations of what was returned,

IE: due to ambiguity of the question “find employee with *feature x y z*” it would return either the entire employee metadata, employee name or employee ID.

System Improvement / Benchmark 2:

Idea: how to correctly incorporate and test unclean data in training data?

4.0 model is extremely sophisticated if a prompt is valid it would most likely result in a valid generated mySQL query command, if invalid then would give an invalid response thus:

