香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

数据科学与分析学域
DATA SCIENCE AND ANALYTICS THRUST
信息枢纽
INFORMATION HUB

DSA Research Experiences for Undergraduates

Research Project

Section1: Faculty Information

Full Name	Yuyu LUO	Tel	13717902271
Thrust/Hub	DSA Thrust, Info Hub	Office	E2-6F-615
Email	yuyuluo@hkust-gz.edu.cn		

Section2: Research Project Proposal

Project Title	LinguaSQL: Unlocking NL2SQL Potential Through Multilingual and Multi-Dialect Prompting
Project Description (max 800 words)	LinguaSQL investigates a novel approach to enhance Natural Language to SQL (NL2SQL) systems by leveraging the multilingual and multi-dialect capabilities of Large Language Models (LLMs) through advanced prompting techniques. The core idea is to generate a more diverse and robust set of candidate SQL queries by prompting the LLM with the same natural language question translated into multiple languages (e.g., English, Chinese) and by requesting SQL output in multiple dialects (e.g., MySQL, PostgreSQL). This approach aims to mitigate the limitations of relying on a single language and dialect, potentially capturing nuances and variations that might be missed otherwise. Objectives: 1. Develop and implement a multilingual and multi-dialect prompting framework for NL2SQL. This involves designing effective prompts for LLMs that incorporate techniques like Chain-of-Thought and few-shot learning, tailored to different languages and SQL dialects. 2. Evaluate the impact of multilingual and multi-dialect prompting on the quality of generated candidate SQL queries. This will involve rigorous experimentation and analysis using standard NL2SQL evaluation metrics. 3. Develop and implement a robust candidate fusion and ranking mechanism to effectively combine and select the best SQL query from the diverse set of candidates generated. Expected Outcomes: 1. Demonstrate a significant improvement in NL2SQL performance compared to existing NL2SQL methods. 2. Provide insights into the effectiveness of different prompting strategies for multilingual and multi-dialect NL2SQL.

3. Potentially lead to more robust, accurate, and generalizable NL2SQL systems that are accessible to a wider range of users and databases.

Detailed Methodology:

LinguaSQL introduces a new framework for NL2SQL that systematically exploits the capabilities of LLMs across diverse linguistic and SQL dialect landscapes through *multilingual and multi-dialect prompting*. The key components of this framework include:

1. **Multilingual Prompting:** The input natural language question will be translated into multiple languages (e.g., English and Chinese). Specifically crafted prompts, incorporating techniques such as Chain-of-Thought (CoT) and few-shot examples, will be designed for each language. These prompts will guide the LLM to generate candidate SQL queries based on each translated question, leveraging the potential for different linguistic representations to capture different nuances of the query intent. We will investigate methods to ensure semantic equivalence across the translated questions and to resolve any inconsistencies.
2. **Multi-Dialect Prompting:** We will design prompts that instruct the LLM to generate candidate SQL queries in multiple SQL dialects (e.g., MySQL, PostgreSQL). This leverages the LLM's potential knowledge of the syntactic and semantic variations between dialects, enabling the creation of a more diverse and potentially more accurate set of candidate queries. Prompts will be designed to encourage the generation of both common SQL structures and dialect-specific features.
3. **Candidate Fusion and Ranking:** The candidate SQL queries generated from different languages and dialects (via different prompting strategies) will be combined and ranked using a scoring mechanism. This mechanism will consider factors such as execution results (when available), structural similarity, and LLM-provided confidence scores. The objective is to select the most accurate and robust SQL query.
4. **Equivalence Verification:** We will explore methods for verifying the semantic equivalence of both the translated natural language questions and the generated SQL queries across languages and dialects. Only equivalent representations will be used for generating the final candidate pool. This will help mitigate the risk of introducing errors through translation or dialect-specific variations.

Proposed Research Duration	Start Date: __2025_____/___03_____/___20_____ End Date: ___2025_____/___08_____/____30____
Student/Researcher Duties	1. **Literature Review:** Conduct a thorough review of existing literature on NL2SQL, LLMs, SQL dialects, and *especially* advanced prompting techniques (e.g., Chain-of-Thought, few-shot learning, prompt tuning). 2. **Dataset Preparation:** Prepare and preprocess datasets for training and evaluation. 3. **Prompt Engineering:** This is a *central* responsibility. The student will design, implement, and iteratively refine prompting strategies for both multilingual question reformulation and multi-dialect SQL generation. 4. **Model Development:** Implement and experiment with different LLMs and prompting techniques, integrating them into the *LinguaSQL* framework. This will involve working with LLM APIs and potentially fine-tuning models. 5. **Experimentation and Evaluation:** Design and conduct rigorous experiments to evaluate the performance of the *LinguaSQL* approach, using standard NL2SQL metrics (e.g., execution accuracy, exact set match accuracy). This will involve setting up a robust evaluation pipeline. 6. **Analysis and Interpretation:** Analyze experimental results, identify trends, draw conclusions about the effectiveness of different prompting strategies, and diagnose any limitations or challenges. 7. **Documentation and Reporting:** Thoroughly document the research process, including prompt designs, experimental setups, results, and code. Prepare reports and presentations summarizing the project's progress and findings. 8. **Collaboration:** Actively participate in discussions with the faculty mentor and potentially other collaborators, providing regular updates, seeking feedback, and contributing to the overall research effort.
Technical Skills Required	☑ Python ☑ Machine Learning ☐ Big Data ☐ R ☑ Deep Learning ☑ SQL ☐ C/C++ ☑ Other: ___LLM____
Preferred Student/Researcher Background	1. **Programming Proficiency:** Strong programming skills in Python are essential. Familiarity with LLM fine-tuning and deployment/inference, including libraries like vLLM, DeepSpeed, or similar tools, is highly desirable. 2. **Database Knowledge:** Basic familiarity with SQL and database concepts is required. Experience with multiple SQL dialects (MySQL, PostgreSQL) is a significant plus. 3. **Machine Learning Fundamentals:** Understanding of basic machine learning concepts, including model training, evaluation, and hyperparameter tuning. 4. **Excellent Communication Skills:** Ability to communicate research findings effectively, including documenting code and experimental procedures.

Maximum Number of Students/Researchers	☐ 1 ☑ 2

Section3: Pre-Application Research Exposure Meeting

Faculty members are encouraged to schedule a Research Exposure Meeting to introduce students to their projects.

Preferred Date	
Preferred Time	
Meeting Mode	☐ In-Person ☐ Online
Venue (if in-person)	
Meeting Link (if online)	