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 | Wei WANG | Tel | 13003187004 |
| Thrust/Hub | DSA/INF | Office | W4 515 |
| Email | weiwcs@ust.hk | | |

### Section2: Research Project Proposal

| | |
|---|---|
| Project Title | Enhancing Natural Language Querying for Structured and Unstructured Data via Deep Reasoning Models |
| Project Description<br>(max 800 words) | 1. Introduction<br>The rise of natural language (NL) interfaces for querying structured (e.g., SQL databases) and unstructured (e.g., text documents) data promises democratized access to information. However, existing deep reasoning models like Deepseek’s R1 face critical limitations in handling ambiguous queries, cross-data source reasoning, and complex contextual understanding. This project aims to systematically evaluate the constraints of such models and develop novel methods to improve their performance for NL-based data querying.<br><br>2. Objectives<br><br>Evaluate Limitations: Analyze the effectiveness of Deepseek’s R1 and similar models in querying structured/unstructured data, identifying gaps in reasoning accuracy, ambiguity resolution, and cross-data integration.<br><br>Improve Ambiguity Handling: Design techniques to enable models to interpret and disambiguate user intent in complex queries (e.g., temporal or spatial dependencies).<br><br>Cross-Data Query Support: Develop methods to unify reasoning across structured (tables, graphs) and unstructured (text) sources, ensuring coherent answers.<br><br>3. Challenges<br><br>(a) Underexplored Model Effectiveness: No rigorous study exists on how state-of-the-art (SOTA) reasoning models like R1 perform on hybrid data querying tasks.<br>(b) Ambiguity Resolution: Current models struggle with implicit user intent (e.g., “Show sales trends last year” without specifying region). |

(c) Cross-Data Integration: Models lack frameworks to jointly query structured databases and unstructured documents (e.g., linking product tables to customer reviews).

4. Methodology
Phase 1: Baseline Evaluation

Benchmark Deepseek R1's performance on hybrid datasets (e.g., WikiSQL + MS MARCO) using metrics like precision, recall, and response coherence.
Identify failure modes (e.g., schema misalignment, context neglect).

Phase 2: Ambiguity-Aware Query Parsing

Contextual Disambiguation: Train R1 to generate clarification prompts for ambiguous queries (e.g., "Which region's sales?") using reinforcement learning with user feedback simulation.

Dynamic Reasoning Chains: Augment R1 with iterative reasoning modules to decompose multi-step queries (e.g., "Compare Q2 revenue to last year's best month").

Phase 3: Cross-Data Hybridization

Unified Representation Learning: Develop a shared embedding space for structured (table columns) and unstructured (text snippets) data using contrastive learning.

Schema-Aware Retrieval: Integrate graph-based neural networks to map NL queries to hybrid data schemas, prioritizing relevance across sources.
Phase 4: Validation

Test enhanced models on real-world datasets (e.g., corporate reports + SQL databases) and synthetic benchmarks with controlled ambiguity.
Compare against SOTA baselines (e.g., ChatGPT-4, DEBERTa) on task-specific accuracy and latency.

5. Expected Contributions

Theoretical: A taxonomy of limitations in deep reasoning models for NL-based querying, with actionable insights for model design.

| | |
|---|---|
| | Technical: Novel methods for ambiguity resolution and cross-data integration, validated through reproducible experiments.<br>Practical: Open-source tools for adapting hybrid querying pipelines to vertical domains (e.g., healthcare, finance).<br>6. Broader Impact<br>While focused on NL-to-data querying, the methodologies (e.g., context-aware parsing, schema mapping) could generalize to domains like automated customer support or legal document analysis. This work bridges the gap between unstructured NL flexibility and structured data precision, advancing toward more intuitive human-data interaction. |
| Proposed Research Duration | Start Date: __2025___ / MAR__ / _02____<br>End Date: __ 2026___ / MAR__ / _01____ |
| Student/Researcher Duties | - |
| Technical Skills Required | ☑ Python<br>☐ R<br>☐ C/C++<br>☐ Machine Learning<br>☐ Deep Learning<br>☑ Other: _Interest in cognitive aspect of learning is a plus_<br>☐ Big Data<br>☑ SQL |
| Preferred Student/Researcher Background | *List preferred coursework, experience, or skills (e.g., statistics, programming, AI).* |
| 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 | Mon / Wed / Fri morning |
| Preferred Time | n/a |
| Meeting Mode | ☑ In-Person ☑ Online |
| Venue (if in-person) | |
| Meeting Link (if online) | On request |