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 | Zeyi WEN | Tel | |
| Thrust/Hub | DSA Thrust, Info Hub | Office | E2 312 |
| Email | wenzeyi@ust.hk | | |

## Section2: Research Project Proposal

| | |
|---|---|
| Project Title | Efficient Hyperparameter Optimization for LLM Inference |
| Project Description (max 800 words) | *Provide a brief summary of the project, objectives, and expected outcomes.*<br>Large language models (LLMs) have demonstrated remarkable capabilities in text generation, summarization, and dialogue systems. However, deploying these models effectively requires careful tuning of inference hyperparameters—such as temperature and top_p—which are crucial for balancing text quality, creativity, and computational efficiency. Hyperparameter optimization for LLMs presents significant challenges, as traditional methods like grid search, random search, and standard Bayesian optimization are often time-consuming, token-intensive, and struggle with the exploration-exploitation trade-off, particularly when evaluating high-cost configurations (e.g., full-scale inference with long prompts). To address these challenges, this project aims to develop an efficient hyperparameter optimization method for LLM deployment, focusing on minimizing token consumption while maintaining or improving text quality. Key research questions include designing advanced searching strategies and implementing multi-fidelity evaluation techniques. The project seeks to establish a principled framework for hyperparameter optimization in LLMs, emphasizing efficiency, scalability, and adaptability. Students will explore both theoretical foundations (e.g., convergence guarantees) and practical implementations, validated through extensive benchmarks on diverse NLP tasks. |
| Proposed Research Duration | Start Date: _2025__ / ___03___ / ___15____<br>End Date: _2025__ / ___09___ / ___01____ |
| Student/Researcher Duties | *List the primary responsibilities and tasks expected from the student/ researcher during the project.*<br>Conduct benchmarking on NLP tasks in terms of text quality and efficiency.<br>Establish an LLM serving system using SOTA frameworks.<br>Design and implement adaptive search strategies to balance exploration-exploitation trade-offs. |
| Technical Skills Required | ☑ Python ☑ Machine Learning ☐ Big Data<br>☐ R ☑ Deep Learning ☐ SQL |

| | |
|---|---|
| | ☐ C/C++ ☐ Other: _____ |
| Preferred Student/Researcher Background | *List preferred coursework, experience, or skills (e.g., statistics, programming, AI).*<br>Math, programming |
| 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) | |