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@hkust-gz.edu.cn		

Section2: Research Project Proposal

Project Title	DeepSeek-V3 Inference Efficiency Optimization
Project Description (max 800 words)	The Mixture of Experts (MoE) model, represented by DeepSeek-V3, is a highly sparse model architecture. A current research focus involves leveraging offloading technology to migrate the sparse parameters of MoE models to the CPU, enabling local deployment of MoE models on a single GPU (e.g., RTX 4090) with high-memory configurations. While KTransformer provides a viable implementation pathway, it currently only supports a batch size of 1. In light of this, we aim to optimize KTransformer for high-throughput performance.
Proposed Research Duration	Start Date: 31 / March / 2025 End Date: 31 / August / 2025
Student/Researcher Duties	Conduct benchmarking on NLP tasks in terms of efficiency. Establish an LLM serving method using SOTA frameworks. Design and implement offloading and parallelism strategies.
Technical Skills Required	☑ Python ☑ Machine Learning ☐ Big Data ☐ R ☑ Deep Learning ☐ SQL ☑ C/C++ ☐ Other: _____
Preferred Student/Researcher Background	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)	