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	Jia Li	Tel	
Thrust/Hub	Thrust of Data Science and Analytics	Office	W2 L6 605
Email	jialee@hkust-gz.edu.cn		

Section2: Research Project Proposal

Project Title	Speeding Up Your LLMs: Advancing Speculative Decoding
Project Description (max 800 words)	Large Language Models have transformed AI applications with their remarkable capabilities in reasoning and generation. Despite these advances, the autoregressive decoding mechanism of LLMs creates significant computational bottlenecks, increasing costs and energy consumption. Speculative decoding offers a promising solution by pairing a smaller "drafter" model with a larger "verifier" model, potentially reducing computational burden while maintaining output quality. Our research addresses two key optimization questions in speculative decoding. First, what is the ideal size relationship between drafter and verifier models? Smaller drafters operate faster but may have lower prediction accuracy, creating an efficiency-quality tradeoff. Second, how should knowledge be organized within drafter models? Should drafters mirror the comprehensive knowledge of larger models, or should we develop domain-specialized drafters that can be selectively deployed based on query content? To determine optimal model size ratios, we will benchmark a 70B parameter verifier model against drafters ranging from 0.5B to 7B parameters (approximately 0.7% to 10% of the verifier's size). We'll measure token acceptance rate, throughput, resource utilization, latency, and output quality for each configuration. This analysis will produce an optimization curve identifying the efficiency sweet spot where marginal gains from increasing drafter size begin to diminish. For knowledge distribution, we'll compare two strategies. The generalist approach will distill comprehensive knowledge from the verifier into smaller models optimized for broad prediction capability. The specialist approach will create multiple drafters focused on distinct domains like science, programming, and creative writing. We'll implement domain-

	specific knowledge distillation techniques to train multiple drafters. Both strategies will undergo evaluation using multiple benchmarks. Building on these insights, we'll develop an intelligent routing system that analyzes incoming queries to identify their domain characteristics, selects the optimal drafter based on query classification, and dynamically adjusts parameters based on real-time performance metrics. This framework will be tested against single-drafter baselines using diverse inputs. Our research will deliver empirical guidelines for drafter-to-verifier size ratios, comparative analysis of generalist versus domain-specialized knowledge strategies, an open-source adaptive drafter selection system, and efficiency analyses compared to traditional decoding approaches. Through systematic investigation of both model size relationships and knowledge organization strategies, this research will establish foundational principles for next-generation LLM inference systems that balance computational efficiency with generation quality, making advanced AI capabilities more accessible and sustainable.
Proposed Research Duration	Start Date: March / 01 / 2025 End Date: Dec / 31 / 2025
Student/Researcher Duties	Maintaining code repositories, designing and conducting experiments, optimizing algorithms, managing datasets, and developing evaluation tools, while actively contributing to research documentation, team collaboration, and result dissemination.
Technical Skills Required	☑ Python ☑ Machine Learning ☐ Big Data ☐ R ☑ Deep Learning ☐ SQL ☐ C/C++ ☐ Other: _________
Preferred Student/Researcher Background	Machine learning, programming (Python, PyTorch, Triton), and experience with large language models, graph data, or reinforcement learning,
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)	