

Xiang LIU

Homepage: <https://dominic789654.github.io/>

Google Scholar: [VtK5lwUAAAAJ](#)

Email: xliu886@connect.hkust-gz.edu.cn

Mobile: +86-15617865777

EDUCATION

-
- **The Hong Kong University of Science and Technology (Guangzhou)** Guangzhou, CHINA
PhD in DSA Thrust, supervised by Professor Xiaowen Chu Sep 2023 - Aug 2027
 - **New York University** NY, USA
Visiting PhD in CDS, supervised by Professor Eunsol Choi July 2025 - Jan 2026
 - **The University of Hong Kong** Hong Kong SAR, CHINA
Master of Science(MSc) in Computer Science Sep 2022 - Aug 2023
 - **George Mason University** VA, USA
Bachelor of Science in Computer Science; GPA: 3.71/4.0 Aug 2018 - Aug 2022
Honors/Awards: Dean's List (2018-2020)

SELECTED RESEARCH

-
- **ChunkKV: Semantic-preserving kv cache compression for efficient long-context llm inference**
NeurIPS 2025
 - Xiang Liu*, Zhenheng Tang*, Peijie Dong, Zeyu Li, Yue Liu, Bo Li, Xuming Hu, Xiaowen Chu
 - KV Cache, Inference, GPU Memory, Long-Context
 - **LISA: Layerwise Importance Sampling for Memory-Efficient Large Language Model Fine-Tuning**
NeurIPS 2024
 - Rui Pan*, Xiang Liu*, Shizhe Diao, Renjie Pi, Jipeng Zhang, Chi Han, Tong Zhang
 - Parameter Efficient Fine-tuning, Layer-wise Optimization, GPU Memory Optimization
 - **Should We Really Edit Language Models? On the Evaluation of Edited Language Models**
NeurIPS 2024
 - Qi Li*, Xiang Liu*, Zhenheng Tang, Peijie Dong, Zeyu Li, Xinglin Pan, Xiaowen Chu
 - Model Editing, Benchmark, Model Robustness
 - **LongGenBench: Long-context Generation Benchmark**
EMNLP Findings 2024
 - Xiang Liu, Peijie Dong, Xuming Hu, Xiaowen Chu
 - Long-context Generation, Long-context LLMs, Logical Coherence

EXPERIENCE

-
- **NYU Center for Data Science** NY, USA
 - **Visiting Student Supervisor: Prof. Eunsol Choi** July 2025 - Jan 2026
 - Proposed **DiffAdapt: Difficulty-Adaptive Reasoning for Token-Efficient LLM Inference**, a novel algorithm for efficient adaptive reasoning for LLMs.
 - **HKUST Statistics and Machine Learning Research Group** HK,CHINA
Research Intern Supervisor: Prof. Tong Zhang Dec 2022 - Aug 2023
 - Proposed **LISA: Layerwise Importance Sampling for Memory-Efficient Large Language Model Fine-Tuning**, a novel algorithm for efficient fine-tuning of LLMs, accepted at *NeurIPS 2024*.
 - Contributed to the LMFlow, a framework that allows fine-tuning and deploying personalized LLMs with minimal cost and effort. And focused on fine-tuning Large Language Models, including LLaMa, Bloom and Vicuna.
 - Conducted research on the Chain-of-thought (COT) method to enhance Large Language Models' logical ability.
 - **Baidu Research Cognitive Computing Lab** Beijing,CHINA
Research Intern Dec 2021 - June 2022
 - Worked on dependency parsing using the Open Information Annotation (OIA) method to convert sentences into directed acyclic graphs (DAGs).

COMPETITIONS

-
- **Kaggle — Feedback Prize - Predicting Effective Arguments Competition** Jun 2022 - Aug 2022
 - Team Leader — Silver Medal (Top 2%)
 - Developed baseline code, and designed various data preprocessing strategies, and model structures.
 - Implemented token classification instead of sequence classification, boosting rank and saving time on training and inference. Acquired proficiency in using Transformers training API.

PROFESSIONAL SKILLS

-
- **Languages:** English, Mandarin
 - **Computing Skills:** PyTorch, Git, Linux