Xiang LIU

Homepage: https://xiangl-ml.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

- **DiffAdapt: Difficulty-Adaptive Reasoning for Token-Efficient LLM Inference** — **ICLR 2026**
  First Author · *Adaptive Inference, Reasoning LLM, Token Efficiency, Entropy Analysis*
- **Reasoning Language Model Inference Serving Unveiled: An Empirical Study** — **ICLR 2026**
  Co-first Author · *LLM Serving, Reasoning System, Empirical Study, Memory Optimization*
- **ChunkKV: Semantic-preserving KV cache compression for efficient long-context LLM** — **NeurIPS 2025**
  First Author · *KV Cache, Inference, GPU Memory, Long-Context*
- **LISA: Layerwise Importance Sampling for Memory-Efficient LLM Fine-Tuning** — **NeurIPS 2024**
  Co-first Author · *Parameter Efficient Fine-tuning, Layer-wise Optimization*
- **Should We Really Edit Language Models? On the Evaluation of Edited LLMs** — **NeurIPS 2024**
  Co-first Author · *Model Editing, Benchmark, Model Robustness*
- **LongGenBench: Long-context Generation Benchmark** — **EMNLP Findings 2024**
  First Author · *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**, a novel algorithm for efficient adaptive reasoning for LLMs (Accepted to ICLR 2026).
  - Conducted comprehensive entropy analysis on reasoning traces to identify "overthinking" phenomena.
- **HKUST Statistics and Machine Learning Research Group** — **HK,CHINA**
  Research Intern ***Supervisor: Prof. Tong Zhang*** — *Dec 2022 - Aug 2023*
  - Proposed **LISA**, a novel algorithm for efficient fine-tuning of LLMs, accepted at *NeurIPS 2024*.
  - Contributed to the LMFlow framework, enabling personalized LLM fine-tuning and deployment (LLaMA, Bloom, Vicuna).
  - Conducted research on Chain-of-Thought (CoT) methods to enhance logical reasoning in LLMs.
- **Baidu Research Cognitive Computing Lab** — **Beijing,CHINA**
  Research Intern — *Dec 2021 - June 2022*
  - Worked on dependency parsing using Open Information Annotation (OIA) to convert sentences into DAGs.

Competitions

- **Kaggle — Feedback Prize - Predicting Effective Arguments Competition** — **Jun 2022 - Aug 2022**
  Team Leader — Silver Medal (Top 2%)
  - Developed baseline code, designed data preprocessing strategies, and optimized model structures.
  - Implemented token classification replacing sequence classification, boosting rank and efficiency.

Professional Skills

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