Kaiwen Chen

Ph.D. in Computer Science and Engineering ✉ kwchen@link.cuhk.edu.hk 🔗 https://kaiwenkevinn.github.io/

Education

The Chinese University of Hong Kong *HONG KONG SAR, China*
Ph.D. in Computer Science and Engineering *2024 Aug – Present*

- Advisor: Prof. Hong Xu
- **Research Area:** Machine Learning System (AI Infra)

Nanjing University *Nanjing, China*
B.E. in Software Engineering *2020 Sep – 2024 Jun*

- GPA 4.46/5.0, 3.82 / 4.0 (Wes Algorithm)

Research Interest

I am broadly interested in System Design for Machine Learning (AI Infra), including the following topics:

- Efficient Serving: KV cache and memory management, scheduling, parallelism, and hardware-aware optimizations.
- Distributed Training Systems: Scalable frameworks for large-scale model training.

Publications

1. MemShare: Memory Efficient Inference for Large Reasoning Models through KV Cache Reuse
Kaiwen Chen, Xin Tan, Minchen Yu, Hong Xu. In arXiv Preprint, Jul 2025.

2. Echo: Simulating Distributed Training At Scale
Yicheng Feng, Yuetao Chen, ***Kaiwen Chen***, Jingzong Li, Tianyuan Wu, Peng Cheng, Chuan Wu, Wei Wang, Tsung-Yi Ho, Hong Xu. In arXiv Preprint, Dec 2024.

Internships

Microsoft Research Asia *Beijing, China*
Research Intern *Nov 2023 – May 2024*

- Investigated BGP routing scalability. Contributions were acknowledged in the final paper.
- Built a forecasting system for routing table growth using multi-layer LSTM.
- Designed and ran simulations on data center network devices to evaluate behavior and performance when routing entry limits are exceeded.
- Proposed and assessed improved route aggregation and selective Add-Path strategies to reduce table pressure, enhance stability, and improve resource efficiency.

Meituan *Beijing, China*
Software Engineer Intern *Jun 2023 – Aug 2023*

- Assisted in developing backend security engine to detect and handle risk events in the coupon system.
- Developed user-friendly front-end pages to help operators set up and manage promotional rules.

Selected Awards

Full Postgraduate Scholarship, The Chinese University of Hong Kong 2024-2028
Outstanding Graduate Student, Nanjing University 2024
People's Scholarship, Nanjing University 2021-2023
Outstanding Student (5% annually), Nanjing University 2021-2023

Skills

Languages: Chinese, English
Programming: Python, VLLM, Pytorch, CUDA, C++, Java

Academic Services

Reviewers: AAAI

Teaching

Teaching Assistant:

- CSCI 3150, Introduction to Operating Systems, CUHK. 2024 Fall
- CSCI 3320, Fundamentals of Machine Learning, CUHK. 2025 Spring