# Jiaxiang Tang

jathon.tang@gmail.com | Tangjackson | Google Scholar | LinkedIn

## Summary

Third-year Ph.D. student in Computer Science at the University of Minnesota, focused on **LLM systems**, including **distributed training/inference**, **post-training and alignment systems**, and **high-performance AI communication**. Experienced in building real-world testbeds, profiling computation and communication bottlenecks, and co-designing runtime, networking, and kernel-level optimizations for efficient and robust AI systems.

## Research Interests

LLM systems (inference and post-training) • Collective communication • RDMA/SmartNIC networking • RLHF and alignment • Kernel optimization for AI workloads

## Education

**University of Minnesota** **Sep 2023 – Present**
*Ph.D., Computer Science*

- Advisors: Zhi-Li Zhang and Ali Anwar.
- Research: distributed LLM systems, post-training/alignment systems, and high-performance AI communication.

**The Hong Kong University of Science and Technology (HKUST)** **Sep 2021 – Aug 2023**
*M.Phil, Internet of Things (IoT) | GPA: 3.52/4*

- Thesis: *Secure Embedding Aggregation for Federated Representation Learning* (ID privacy via a new distributed protocol).
- Research: distributed ML systems with an emphasis on scalability and privacy.
- Advisor: Songze Li.

**University of California, Berkeley** **Aug 2019 – Dec 2019**
*Exchange, Mathematics | GPA: 3.85/4*

- Selected coursework: Complex Analysis, Probability, Statistics, and PDE.

**Nankai University** **Sep 2017 – Jul 2021**
*B.S., Mathematics (Boling Class) | GPA: 3.6/4*

- Honors: Boling Scholarship (2017), Innovation Scholarship (2017–2018), and Nankai International Scholarship.
- Selected coursework: Java, Data Structures, Operating Systems, and Advanced Algebra.

## Research Experience

**Research Assistant** **Aug 2023 – Present**
*University of Minnesota (UMN)*

- **GPU sharing system optimization:** Investigated decentralized and communication-aware designs for FlexLLM-style co-serving of RL training and LLM inference; studied bursty inference interference, SLO violations, and runtime scheduling across computation and communication layers.
- **Multi-value alignment for LLMs:** Designed a scalable post-training system for multi-value alignment, enabling stable integration of multiple alignment objectives while improving robustness and controllability across heterogeneous value mixtures.
- **AI networking on RDMA/RoCE:** Built real-world RDMA testbeds to study training traffic patterns, congestion, and tail latency; evaluated system design knobs and proposed improvements grounded in empirical measurements.
- **Collective communication for LLM workloads:** Profiled and optimized distributed execution with **torch.distributed** (NCCL/Gloo); built workload-faithful microbenchmarks, profilers, and NS-3 simulations to attribute end-to-end latency to computation and communication components.
- **Agentic communication analysis:** Developed an LLM-based agent system that automates workflows across XCAP business software, logs, traces, and profiling tools to diagnose network and collective communication bottlenecks and generate debugging and optimization recommendations.

**Research Assistant** **May 2021 – Aug 2023**
*Hong Kong University of Science and Technology (HKUST)*

- Designed privacy-preserving federated learning protocols using cryptographic primitives (MPC/HE) to satisfy privacy constraints.
- Worked on scalable and secure distributed system designs for IoT and edge settings, including protocol design and system evaluation.

**Research Intern** **Jun 2020 – Sep 2020**
*Pennsylvania State University*

- Developed a recommendation system for sports and exercise programs to personalize user experience.
- Improved algorithm accuracy by **15%** through feature engineering and model tuning.

## Industry Experience

**ML Algorithm Intern** **Nov 2020 – Feb 2021**
*Huiyan Tech (Tianjin)*

- Implemented state-of-the-art speech diarization and recognition algorithms (**EEND**) using Chainer and Kaldi.
- Improved speech recognition accuracy in noisy environments by **10%**.
- Performed extensive testing and optimization and integrated modules into production with engineering teams.

## Honors & Awards

- **IEEE INFOCOM NSF Travel Grant, 2026**.
- **IEEE ISIT Travel Grant, 2023**.
- **Nankai International Scholarship, 2019**, Nankai University.

## Publications & Manuscripts

**Manuscripts / Under Review**

- **multi-Queue RDMA in AI Data Centers for Scalable Training Acceleration.**
- **Multi-Path Video Streaming Optimization in 5G Networks.**
- **MASS: Multi-Value Alignment for LLM at Scale.**

**Conference**

- **AHE: Adaptive Homomorphic Encryption for Federated Learning.** IEEE INFOCOM 2026.
- **SecEA: Secure Embedding Aggregation for Federated Representation Learning.** IEEE ISIT 2023; IEEE TIFS 2025.
- **HERL: Tiered Federated Learning with Adaptive Homomorphic Encryption using Reinforcement Learning.** IEEE TPS 2025.

## Teaching

**Teaching Assistant, University of Minnesota** **Sep 2024 – May 2025**

- CSCI 1113: Introduction to C/C++.
- CSCI 5707: Introduction to Databases.

**Teaching Assistant, HKUST** **Sep 2022 – Dec 2022**

- ECE: Convex Optimization.

## Skills

- **Distributed/Comms:** PyTorch Distributed, NCCL, Gloo; collective communication, MPI.
- **Networking:** RDMA/RoCE fundamentals; congestion and tail-latency measurement; packet/trace tooling (e.g., tcpdump, Wireshark).
- **LLM Systems:** vLLM, SGLang, Ray, LangChain.
- **Kernel/Performance:** CUDA, kernel profiling, Nsight Systems, torch.profiler.
- **Systems:** Linux, Docker, NS-3.
- **Languages:** Python, C/C++, CUDA.