Tiannuo Yang

Updated September 5, 2025

Contact Information

Ginsburg Hall (GCS), 1031 Downey Way
Los Angeles, CA 90089
tnyang2000@gmail.com
https://tiannuo-yang.github.io/

Research Interests

I am broadly interested in AI systems, especially generative AI such as LLMs, RAG, agents, and world models. I advocate for AI for all. Overall, my research is about two questions: how to understand efficiency insights underneath groundbreaking AI techs; and how to supercharge AI deployment in practice with full-stack designs?

To this end, I like to collaborate with AI people. Check out our latest work, SearchAgent-X, a collaboration with the Search-R1 team, on how to accelerate cutting-edge LLM search agents!

Education

University of Southern California, Los Angeles, CA, United States
Ph.D. Student in Computer Science — from August 2025
Advisor: Willie Neiswanger

Nankai University, Tianjin, China
Master in Computer Science (Exempted from Entrance Exam) — August 2022 – Present
Advisor: Professor Yusen Li GPA: *3.63/4.0*
Thesis: Automated Performance Tuning Techniques for Parallel Applications

University of Science and Technology Beijing, Beijing, China
Bachelor in Information Management and Information System — August 2018 – June 2022
Major GPA: *3.97/4.0* Cumulative GPA: *3.89/4.0*

Southern Taiwan University of Science and Technology, Taiwan, China
Major in Information Management (Exchange Program) — September 2019 – January 2020
Cumulative GPA: *4.3/4.3*

Research Experience

University of Illinois at Urbana-Champaign, Urbana, IL, United States
Retrieval-Augmented Generation — May 2024 – Present
Working with Professor Minjia Zhang on GPU-enhanced retrieval augmented generation, shedding lights on key concerns like batching strategies and latency-quality tradeoffs.

Nankai University, Tianjin, China
Datacenter, System, Machine Learning for System — August 2022 – Present
Working with Professor Yusen Li on automatic performance tuning and hardware resource isolation for job collocations within multi-core systems.

Ant Group, Beijing, China
Vector Retrieval, Vector Database Optimization — June 2023 – January 2024
Worked as a research intern under Dr. Jianguo Li and Wen Hu on optimizing AI infrastructure - vector database, enhancing CodeFuse services (a coding large language model platform).

University of Chinese Academy of Sciences, Beijing, China
Mixed Integer Programming, Heuristic Algorithm — September 2020 – September 2021
Worked as an undergraduate research assistant under Professor Guanghui Zhou to develop data-driven combinatorial optimization problem (vehicle routing optimization).

Publications

Supercharging AI Systems

**Denotes equal contribution.* †*Denotes corresponding authorship.*

T. Yang, Z. Yao, B. Jin, L. Cui, Y. Li, G. Wang, X. Liu. “Demystifying and Enhancing the Efficiency of Large Language Model Based Search Agents.” *arXiv*, 2025

K. Cheng, Z. Wang, W. Hu, T. Yang, J. Li, S. Zhang. “SCOOT: Towards SLO-Optimized LLM Serving via Automatic Inference Engine Tuning.” *The Web Conference (WWW)*, 2025. Oral Presentation.

T. Yang, W. Hu, W. Peng, Y. Li, J. Li, X. Liu, G. Wang. “VDTuner: Automated Performance Tuning for Vector Data Management Systems.” *International Conference on Data Engineering (ICDE)*, 2024. Deployment on Ant Group’s CodeFuse platform.

T. Yang, R. Chen, Y. Li, X. Liu, G. Wang. “CoTuner: A Hierarchical Learning Framework for Coordinately Optimizing Resource Partitioning and Parameter Tuning.” *International Conference on Parallel Processing (ICPP)*, 2023.

Research Survey

Y. Zhou*, X. Lin*, X. Zhang*, M. Wang*, G. Jiang*, H. Lu*, Y. Wu*, K. Zhang*, Z. Yang*, K. Wang*, Y. Sui*, F. Jia* Z. Tang*, Y. Zhao*, H. Zhang*, T. Yang*, W. Chen*, Y. Mao*, Y. Li*, D. Bao*, Y. Li*, H. Liao*, T. Liu*, J. Liu*, J. Guo*, X. Zhao, Y. WEI, H. Qian, Q. Liu, X. Wang, W.K. Chan, C. Li, Y. Li, S. Yang, J. Yan, C. Mou, S. Han, W. Jin, G. Zhang, X. Zeng. “On the Opportunities of Green Computing: A Survey.” *arXiv*, 2023
(Writing Section: 6.4 Resource Optimization).

Operations Research (Undergraduate Thesis)

T. Yang†, Z. Chu, B. Wang. “Feasibility on the Integration of Passenger and Freight Transportation in Rural Areas: A Service Mode and an Optimization Model.” *Socio-Economic Planning Sciences* (SCI JCR Q1), 2023.

Research Projects

Qiyuan Laboratory Innovation Fund — November 2023 – November 2024
Worked as a core member on resource isolation mechanism for a multi-tenant cache system (i.e., Cachelib by Facebook).

CCF-Ant Research Fund on Green Computing — January 2023 – January 2024
Worked as a leader to write project proposal and conclusion, conduct research on AI infrastructure (vector database optimization) and practical platform deployment.

National Natural Science Foundation (NSF) of China — January 2023 – Present
Working as a core member on improving resource utilization in cloud with QoS guarantee.

Major Project of National NSF of China — December 2022 – Present
Working on real-time scheduling of cluster robots for major equipment manufacturing.

Honors And Awards

Ph.D. Student Fellowship, USC	2025
Excellent Graduate, Nankai University (8 out of 157)	2025
1st-Class Gongneng Scholarship, Nankai University	2023, 2024
National 3rd Prize, Massive Storage Competition	2022
National 2nd Prize (1493/41826 = top 3.6%), Contemporary Undergraduate Mathematical Contest in Modeling	2020
Top Ten Singers on Campus, USTB	2020

Talks And Services

Invited talk on ”Towards Efficient LLM Search Agents”
by Di Wu, at ByteDance, June 6th 2025
at MLSys Reading Group, Nankai University, June 5th 2025
by Haosen Shi, at CUHK, June 4th 2025

Conference Talk at
the 40th ICDE at Utrecht, the Netherlands, May 2024
the 52nd ICPP, Online, August 2023

Reading Group Founder and Leader for

Machine learning system research at Nankai-Baidu Joint Lab (from October 2024)

Reviewer for

the 2025 ACM Web Conference

Open Source

SearchAgent-X: Highly efficient system for reasoning-search interleaved LLM agents.
https://github.com/tiannuo-yang/SearchAgent-X

VDTuner: Automated performance tuning framework for vector data management systems.
https://github.com/tiannuo-yang/VDTuner
(Chinese Blog: https://mp.weixin.qq.com/s/1JgXM5WSWBTv7fA0TLGfqw)

G-VRP-IPD-TW: Mathematical model and real-world dataset for a complex combinatorial optimization problem in transportation.
https://github.com/tiannuo-yang/G-VRP-IPD-TW