

Aaron YU

Role: Head of AI Team in a Listed Company

Expertise: Large Language Models, Multimodal Models, Agent Infra, Post-training (SFT and RL), and Inference Acceleration by Co-design of Hardware and Software

Location: Shenzhen, China

Bio

AI team leader with a Ph.D. in Computer Science and Engineering from The Chinese University of Hong Kong, earned in 2018. Aaron currently leads research and development in foundation models, agent harnesses, and AI productization at a publicly listed company. My work focuses on large language models (LLMs), vision-language models (VLMs), vision-language-action models (VLAs), inference systems for models with hundreds of billions of parameters, and coding-agent infrastructure. Recipient of the Wu Wenjun AI Science and Technology Progress Award, First Prize.

Experience

Head of AI Department, Apr 2025 - Present

- Leading a team of 50-60 persons in cutting-edge AI techniques spanning post-training of large models (e.g., Qwen3, Deepseek v4, GLM 5.2), inference acceleration (with quantization, pruning, kernel optimization, etc).
- Co-designed hardware-software for Transformers (hybrid attention and large MoEs) for edge devices.
- Built AI infrastructure capabilities including high-performance kernel generation and optimization, and PD-disaggregated inference, **improving TTFT and TPOT by more than 50% (best practice)**.
- Built a coding agent to **improve IPD efficiency by more than 100% with agent infra**, and our kernel agent **achieved the First Place in MLsys Challenge of 2026**.

Head of Foundation Models, Mar 2023 - Mar 2025

- Led a team of 20–30 researchers and engineers in training large language models with over 100 billion parameters. The resulting models **ranked No. 1 on the C-Eval leaderboard in September 2023 and No. 1 on the MMBench leaderboard in April 2024**. The models were deployed for customers in the public sector.
- Developed and commercialized vision-language models with more than 2 billion parameters, powering AI products that **generated over RMB 50 million in annual revenue in 2025**.
- Developed cross-modal models for large-scale image and video retrieval, applied in industrial monitoring.

Director of Big Data Team, Dec 2020 - Dec 2022

- Led a 5-10 person team building a transformer-based public **service classification system processing more than 100,000 calls per day with accuracy above 95% and better than manual routing**.
- Built post-trained public-sector large models for call summarization, named entity extraction, and anomaly warning; **deployed them in cities with the revenue tens of millions of RMB**.

AI Researcher, Jul 2019 - Dec 2020

- Designed transformer-based algorithms for large-scale spatiotemporal image data in behavior analysis.

Education

- Ph.D. in Computer Science and Engineering, The Chinese University of Hong Kong
Hong Kong SAR, China, Aug 2014 - Dec 2018
- Master in Communication and Information Systems, Sun Yat-sen University
Guangzhou, China, Sep 2011 - Jul 2014
- Bachelor in Automation, South China University of Technology
Guangzhou, China, Sep 2007 - Jul 2011

Honors

- National High-level Talents (Qiming Project), 2024
- Top 10 Outstanding Youth of Nanshan District, Shenzhen, 2023
- Wu Wenjun AI Science and Technology Progress Award, First Prize, 2021

Publications, Patents & Research Frontiers

- **150+ granted and pending patents** in AI, machine learning, and related technologies.
- **50+ publications** in leading AI conferences (e.g., NeurIPS, AAAI) and peer-reviewed journals (e.g., *Pattern Recognition*, *Neurocomputing*).
- **Collaborative research projects** on LLMs, VLMs and VLA with leading academic institutions, including Harbin Institute of Technology (Shenzhen) with Prof. Min Zhang and Tsinghua University (Shenzhen) with Prof. Wenming Yang.