Yuyi Ao

yuyia@andrew.cmu.edu | website | 412-475-7661 | LinkedIn

Education

Carnegie Mellon University | Master in Computational Data Science — Expected Dec 2026

University of Illinois at Urbana-Champaign | Bachelor of Science in Computer Engineering — May 2025
GPA 3.96/4.0, Highest Honors, Dean's List (2023 & 2024)

Zhejiang University | Bachelor of Engineering in Electrical and Computer Engineering — May 2025
GPA 3.96/4.0, National Scholarship, Provincial Outstanding Graduate

Coursework: Machine Learning, Computer Systems, GPU Programming, Numerical Analysis, Probability, Data Mining, Probabilistic Graphical Models, Artificial Intelligence, Digital Systems

Skills

ML/AI: PyTorch, vLLM, DeepSpeed, SGLang, HuggingFace Transformers, OpenAI API, XGBoost, LLaMA Factory, Scikit-learn
Programming Languages: Python, C, C++, CUDA, MATLAB, x86 Assembly, SystemVerilog, Shell/Bash
Systems & Tools: Linux, Git, Docker, Kubernetes, Slurm, Azure, Prometheus, Grafana, GDB, AWS EC2

Experience

Tencent | Machine Learning Engineer Intern | LLMs, AI-Agent, Post-training, Reasoning Model — June 2025 – Aug 2025

- Post-trained **DeepSeek-R1-32B reasoning model** as the core of PUBG Mobile's AI teammate, a feature deployed in one of the world's most-played mobile games with **30M+ monthly players**, boosting players in-game interactive experiences.
- Fine-tuned a unified LLM for chitchat and instructions, enhanced with **Chain-of-Thought** to enable high-quality dialogue responses and behavior-tree function calls; achieved **92%** chitchat and **81%** instruction recall on live queries.
- Collaborated with Human Evaluation team to validate LLM for production readiness and ensure human-level response quality.
- Developed from scratch a scalable training data pipeline for an AI clock assistant, designing a **function-call schema** with structured fields for alarm representation; the pipeline remains in use and supports future advanced alarm designs.

UIUC SSAIL Lab | Research Assistant | LLM Inference, vLLM, SLO-Aware Scheduling — Mar 2024 – Mar 2025

- Constructed an efficient configuration searcher with **XGBoost + Simulated Annealing** for LLM serving engine (e.g. vLLM) to find near-optimal configurations for multi-objective scenarios.
- Co-designed an **SLO-Aware Scheduler** in vLLM that mitigates starvation by **swapping** long-running request's KV cache to CPU with **overlapped offloading**, reducing TTFT tail latency by **28.3%–89.9%** with $<5\%$ TBT overhead.
- Conducted extensive benchmarking and analysis of latency and throughput metrics in both online and offline serving scenarios, identifying queuing delay as the dominant cause of SLA violations under heavy load.

UIUC iSAIL Lab | Research Assistant | Knowledge Graph Embeddings, Information Retrieval — Dec 2023 – Mar 2024

- Implemented SpherE, the first sphere-based embedding model for knowledge graphs, enabling expressive many-to-many relation modeling and achieving **20%+** F1 score improvements over strong baselines (RotatE, HousE) on link prediction tasks.
- Built a benchmark system with F1 and retrieval rate metrics, formulated a standardized evaluation method for set retrieval.
- Redesigned the NCE loss to encourage sphere intersections, enhancing relation modeling and sphere parameter optimization.

Projects

Convolutional Layer Forward-pass Optimization (CUDA, C++)

- Reduced convolutional-layer inference time from 70 ms to 9 ms, ranking #2 among 70 participants on the course leaderboard.
- Applied kernel fusion and im2col transformation to improve data locality and reduce overall global memory accesses.
- Leveraged tensor core warp-level matrix functions to accelerate computation and increase GPU kernel efficiency.

Operating System Design (C, x86 Assembly)

- Built a Unix-like OS with paging, interrupts, file system, and process management features with 3 teammates.
- Implemented system calls and round-robin scheduling algorithms to support responsive multi-terminal operations.
- Developed a file system with write support, adding tab completion and command history to improve interactivity.

Publication

SpherE: Expressive and Interpretable Knowledge Graph Embedding for Set Retrieval — Zihao Li, **Yuyi Ao**, Jingrui He (SIGIR' 24)