YANCHEN LIU

University of Southern California, Los Angeles, CA, USA
+86-189-173-18020 | jamesnulliu@gmail.com | https://jamesnulliu.github.io

RESEARCH INTERESTS

Natural Language Processing

- Reinforcement Learning from Human Feedback
- Reasoning
- Test-time Computing
- Inference Acceleration

Machine Learning

- In-vehicle System Security
- Super-Resolution for Meteorological Data

Software Engineering

- CUDA Programming
- Fast Kernel Development
- AI Compiler

EDUCATION HISTORY

M.Sc. in Computer Science **Sep. 2025 – Sep. 2027**
University of California, LA, CA, USA
B.Eng. in Computer Science **Sep. 2021 – Jun. 2025**
Shanghai University, Shanghai, China

HONORS AND AWARDS

1. [2024] First Prize and Group Competition Award in 2024 ASC Student Supercomputer Challenge Global Final.
2. [2022] First-Class Academic Scholarship for outstanding academic performance, Shanghai University.

PROFESSIONAL EXPERIENCE

Graduate Research Intern **Jun. 2025 - Present**
University of Southern California | INK Lab
RLHF and Reasoning for LLMs

Machine Learning Engineer Intern **Jul. 2024 - Jun. 2025**
Shanghai AI Laboratory
LLM Inference Engine, AI Compiler and Model Fine-tuning

- Extended vllm for LLM inference on in-house TPUs, including researching and adapting Speculative Decoding, Paged-Attention and Continuous-Batching.
- Developed high-performance kernels with MLIR to ensure seamless compatibility and optimal performance of LLMs on in-house TPUs.
- Conducted research on LLM knowledge injection and fine-tuning for kernel fusion and translation across different hardware platforms.

Undergraduate Research Assistant **Mar. 2023 - Apr. 2025**

Shanghai University | Shanghai University Cyber Security Lab

Vehicle Modeling, Simulation, and Intrusion Detection

- Combined deep learning and traditional math modeling to assess cyber security and functional safety in in-vehicle communication.
- Developed a novel efficient gradient descent solver for Multi-Dimensional Hawkes Process, and implemented algebraic simplification, vectorization and JIT compilation for optimization.
- Introduced a novel MDHP-LSTM structure for improved feature extraction in vehicular ECU communication system and related applications.

Team Leader of Shanghai University Super-Computing Team **Sep. 2023 - Jul. 2024**

Shanghai University | SHUSCT

Super-Computing

- Participated in 2024 ASC Student Supercomputer Challenge.
- Assembled, benchmarked and optimized a high-performance computing cluster for running large language models and super-computing programs.
- Developed a custom LLM inference engine with various parallelism and scheduling policies; Integrated FlashAttention, quantization and pruning techniques to speed up inference.
- Profiled and improved performance of various super-computing programs by applying vector instructions, OpenMP, loop unrolling, and MPI.

Machine Learning Research Intern **Jun.2023 – Jul. 2024**

Shanghai University | East China Air Traffic Control Bureau

Super Resolution for Meteorological Data

- Deployed cutting-edge deep learning models including PanGu and FourCastNet to improve meteorological monitoring, forecasting, and super-resolution for the East China Air Traffic Control Bureau, replacing traditional numerical methods.
- Utilized key-point and semantic constraints, combined with feature map fusion and redesigned loss functions, to improve the up-sampling process in super-resolution models, addressing the issue of semantic detail loss in certain areas.

PUBLICATIONS

1. [2025 | Preprint] J. Lv, X. He, **Y. Liu**, A. Shen, X. Dai, Y. Li, J. Hao, J. Ding, Y. Hu, S. Yin. "HPCTransEval: A Benchmark of High-Performance GPU-to-CPU Transpilation with Pre-trained Large Language Models".
2. [2025 | Preprint | Code] Q. Liu†, **Y. Liu†**, R. Li, C. Cao, Y. Li*, X. Li*, P. Wang, R. Feng, "MDHP-Net: Detecting an Emerging Time-Exciting Threat in IVN".
3. *[2025 | Preprint]* Z. Xu, A. Shen, D. Kong, X. Dai, J. Liu, **Y. Liu**, L. Wang, S. Wei, Y. Hu and S. Yin*, "LLMEngine: Disaggregated Mapping and Memory Management Co-scheduling for Wafer-scale Chips".
4. [2024 | IEEE Internet of Things Journal | Code] Q. Liu, X. Li, K. Sun, Y. Li* and **Y. Liu***, "SISSA: Real-Time Monitoring of Hardware Functional Safety and Cybersecurity With In-Vehicle SOME/IP Ethernet Traffic".
5. [2024 | MDPI Future Internet | Code] Li, X., R. Li, and **Y. Liu**. "HP-LSTM: Hawkes Process–LSTM-Based Detection of DDoS Attack for In-Vehicle Network".