Yueyuan Huang

Phone (+86) 15317676578 · Mail harvyhyy@gmail.com
Homepage: nekoyellow.github.io

Education

Fudan University, Software Engineering, *B.Eng.* Sep. 2021 – Jun. 2026

- **GPA**: 89/100
- **Selected Courses**: Software Engineering, Operating Systems, Computer Vision, Machine Learning Systems

Uppsala University (Sweden), *Exchange Student* Sep. 2023 – Jan. 2024

- **Selected Courses**: Data Structures, Algorithms, Computer Architecture, Database Design

Experiences

Fudan ACM-XCPC, Member Jul. 2024 – Jun. 2025

- Specialized in solving graph theory, simulation, string, and constructive algorithm problems in contests
- Led the team as captain: coordinated weekly training plans, assigned problem sets, and ensured balanced topic coverage
- Composed post-contest reports and analyses to identify mistakes and refine strategies
- Awarded **Bronze Medal** (89/280) in National CCPC 2024 and **Gold Medal** (14/153) in Provincial CCPC 2025 (China Collegiate Programming Contest, a top-tier ACM-style contest equivalent to ICPC regionals in China)

Ant Group, Data Development Intern Jun. 2023 – Aug. 2023

- Built over 10 offline ETL pipelines on MaxCompute for Alipay's transportation discount service, processing millions of daily orders and ensuring timely and reliable business metrics
- Developed real-time stream analytics using Flink to provide minute-level coupon redemption stats for operational dashboards used by multiple business teams
- Designed A/B test instrumentation with product and strategy teams; independently delivered 5+ SQL reports supporting subsidy strategy optimization for tens of millions of users

Projects

LangArch, LLM Inference Framework Jun. 2025 – Aug. 2025

- Designed and implemented a Radix Tree–based prompt caching system enabling inter-request KV cache reuse; outperformed standard Hugging Face Transformers inference by 5.3× in throughput and 2.6× in latency on structured workloads like Tree of Thought
- Developed a Cache-Aware routing algorithm for multi-node data parallel inference; improved cache hit rate by 3.5× and throughput by 1.7× over Round Robin routing
- Re-architected unified scheduling into Prefill-Decoding separation to eliminate phase interference and support independent optimization; further improved throughput by 1.5×

INT-NN, Pure Integer Neural Network Framework Apr. 2025 – May. 2025

- Led development of INT-NN, a C-based integer neural network library for resource-constrained devices, supporting full training and inference
- Integrated Direct Feedback Alignment (DFA) and quantized activations to enable end-to-end training without floating-point operations, enhancing model stability
- Achieved 97.02% accuracy on MNIST with 10× faster training and comparable precision to float models
- Project: INT-NN

TextHydra, Document Image Tampering Detection Nov. 2024 – Dec. 2024

- Designed a multi-stream detection framework using visual, frequency (DCT), and noise features to enhance document-level tampering verification
- Implemented scSE + Transformer fusion modules enabling pixel-level tampering localization with improved spatial reasoning

- Achieved F1 score of 94.27 in Alibaba Tianchi competition, ranking top 3 at submission; method was documented into a paper
- Paper: TextHydra.pdf

xv6-fdu, OS Kernel Extension and Optimization Oct. 2024 – Dec. 2024

- Extended xv6 kernel with priority scheduling, kernel page table isolation, user privilege management, and environment variable support
- Refactored buffer cache and freelist to reduce lock contention and improve multicore scalability
- Adopted as a lab assignment and awarded 3rd prize in East China regional Computer System Capability Competition
- Project: xv6-fdu

RookieDB, Core Database System Implementation May. 2023 – Jun. 2023

- Independently completed Berkeley CS186 course project with full-stack implementation of B+ tree index, query execution, and optimizer
- Supported multi-transaction concurrency and lock-based scheduling, covering buffer pool, logging, and recovery subsystems
- Designed and executed system test suites to validate correctness and identify performance bottlenecks
- Project: rookiedb-sp23

SKILLS

Languages: Mandarin (native), English (fluent; TOEFL 106), Japanese (conversational)
Programming: **Python** (primary; data science, ML), **C++** (fluent; algorithms, high-performance computing), **Go** (proficient; distributed systems, concurrency), **OCaml**, **SQL**
Technologies: **Machine Learning** (proficient in PyTorch; familiar with system internals), **LLM** (algorithm understanding; hands-on with local/distributed deployment), **DevOps** (Docker, Kubernetes, SSH, Linux)
Interests: Skiing (5 yrs; carving), Snowboarding (2 yrs; powder), Badminton (10 yrs)