

Qian Yu

Singapore | yuqian05@u.nus.edu | +65 8654 7249 | [linkedin.com/in/yuqiannemo](https://www.linkedin.com/in/yuqiannemo) | github.com/yuqiannemo

Education

National University of Singapore

B.Comp. in Computer Science; Second Major in Mathematics

Aug 2023 – June 2027

Singapore

- GPA: 4.7/5.0; Dean's List (top 5%); focus in Artificial Intelligence.
- Teaching Assistant: Programming Methodology; Data Structures and Algorithms.

University of Washington

Exchange Student

Sep 2025 – Dec 2025

Seattle, WA

- Coursework: Advanced Algorithms, Linear Optimization, and Introduction to Real Analysis.

Experience

TikTok

Backend Engineer Intern, Recommendation Architecture

Jan 2026 – Aug 2026

Singapore

- Designed and deployed a C++ rule engine for configurable trace sampling/filtering, reducing end-to-end tail latency from 30 ms to 10 ms in production and supporting multi-business (ForU/Lemon8) differentiated sampling strategies.
- Built a Go common fragment framework with item-level full-trace enhancements, standardizing trace structures to reduce 70% of redundant node noise in observability views and enabling fine-grained debugging for 20+ business metrics.
- Developed configurable trace statistics capabilities, cutting Lemon8 case investigation time from 3+ days (offline processing) to 0.5 day and expanding access from algorithm engineers to product and operations teams.
- Automated a one-click AI workflow with periodic inspection to diagnose duplicate recommendation cases, reducing manual investigation time from one day to 30 minutes and serving 2,000+ users through MCP tools and agent orchestration.

Beijing Freedo Technology Co., Ltd.

Computer Vision Research Intern

Dec 2024 – Jan 2025

Beijing, China

- Developed multi-view image-analysis and contour-extraction pipelines for parameterized 3D building reconstruction, fine-tuning LCNN-style models for architectural image datasets.

Computing for Voluntary Welfare Organisations

Software Engineer Intern

May 2024 – Aug 2024

Singapore

- Developed full-stack case-management features using React.js and Go for volunteer welfare organizations in Singapore.
- Built volunteer onboarding and workflow-management modules for Lions Befrienders Service Association, improving usability and streamlining internal service workflows.

Technical Skills

Programming: C++, Go, Python, C, Java, SQL

Backend: distributed systems, concurrency, networking, observability, message queues, REST APIs

Frameworks/Tools: Linux, PostgreSQL, React/TypeScript, FastAPI, PyTorch, vLLM, Git

Research Experience

National University of Singapore

Undergraduate Researcher, [LLM DNA](#)

Aug 2026 – Present

Singapore

- Investigating how sampling seeds and decoding configurations affect the stability of functional LLM fingerprints while extending the LLM DNA study from 305 toward 2000+ open- and closed-source models.
- Designed a seed-controlled dose-response study across 12 models, 9 temperature/top-p settings, and 32 generation seeds (3,456 runs; 345,600 responses), revealing that sensitivity to decoding shifts is strongly model- and operating-point-dependent.
- Developed paired held-out SVM and multivariate KDE analyses over 76,800-D functional fingerprints, obtaining 0.0% median wrong-side error for both 410 cross-family and 25 same-family contrasts and motivating checkpoint-level attribution with family metadata.

Research Experience (continued)

National University of Singapore

Undergraduate Researcher, *AgenticServing*

May 2025 – Oct 2025

Singapore

- Co-designed a multi-GPU LLM serving architecture that represents agentic workflows as DAGs and coordinates a centralized router with per-GPU schedulers and vLLM executors to reuse computation across concurrent pipelines.
- Formulated a prefix- and graph-aware placement objective combining token-level longest-common-prefix similarity, reusable initial subgraphs, queue length, and KV-cache pressure to balance computation reuse against worker load.
- Designed cross-workflow DAG merging, critical-path front-edge batching, and segment-level KV-cache eviction based on future reuse, pending time, and cache size; established an evaluation framework covering workflow completion latency, cache-hit ratio, and GPU utilization across chain, fan-out/fan-in, and hybrid workloads.

Selected Projects

npcTrading Modern C++, message bus, GTest, Binance Spot

Oct 2025 - Jan 2026

- Engineered a message-driven trading system separating market data, order management, risk, and execution, with netting and hedging modes for position handling.
- Implemented an L2 order book with VWAP, order-book imbalance, and market-impact calculations to support price discovery and execution decisions.

Hunyuan3D Agent Tencent Rhino-bird Open-source Training Program

Aug 2025 – Sep 2025

- Developed an LLM-based scene-decomposition agent that converted complex text prompts into structured objects and spatial relationships for downstream 3D scene generation.
- Orchestrated an end-to-end text-to-3D pipeline integrating Qwen multi-view image generation, Hunyuan3D 2.1 asset creation, code-generating LLMs, and Blender scene composition and rendering.
- Recognized as an Outstanding Student in the Tencent Hunyuan LLM track for contributions to the open-source project.

Tower Networking, file synchronization, systems design

DubHacks 2025

- Built a LAN file-synchronization prototype on Raspberry Pi for direct file transfer between nearby devices without relying on cloud storage.

ChoreoAI Python, Flask, Three.js, Supabase, EDGE

HackMIT 2025

- Integrated music-conditioned dance generation with a Flask API and interactive Three.js avatar, supporting synchronized playback and 3D motion export in a weekend prototype.
- Built the processing workflow for audio upload, generation status, motion conversion, and downloadable outputs, turning a research model into an end-to-end prototype.

WanderMind Next.js, FastAPI, Claude Haiku

CalHacks 12.0

- Built an AI travel planner that converted open-ended preferences into personalized itineraries through a typed frontend, Python API, and LLM orchestration.

Recognition & Activities

Hudson River Trading Explore APAC

Selected Participant – 1 of 15 across APAC

May 2025

Singapore

- Selected for an immersive program featuring technical sessions on algorithmic trading and software engineering led by HRT engineers.

Asia Business Plan Competition

Finalist

Jun 2025

Singapore

- Proposed an edge-AI shrimp-disease detection system using CNN inference on edge devices for aquaculture deployment in Southeast Asia.

GIC Computing for the Community Fellowship Award

Fellowship Recipient

Aug 2024

Singapore

- Recognized for engineering contributions to nonprofit technology initiatives in Singapore.

Languages: Mandarin (native), English (IELTS 8.0)