

Jinxi (Jace) Yu

+1 (626) 265-0641 | yujinxi1104@hotmail.com | linkedin.com/in/jinxiy1104 | jinxiy1104.github.io

EDUCATION

University of California, Los Angeles (UCLA)

B.S. in Computer Science

GPA: 3.97/4.00 | Awards: Dean's Honors List

Relevant Coursework: Object-Oriented Programming (C++), ML, NLP, DL, Algorithms, Operating System, Data Mining

Los Angeles, CA

Sep 2024 – Jun 2028

RESEARCH & WORK EXPERIENCE

UCLA Moonshot Math Harness (Prof. Kai-Wei Chang, Nanyun Peng, Terence Tao)

Research Assistant

Los Angeles, CA

Jul 2026 – Present

- Built a **long-horizon multi-agent reasoning harness** for autonomous mathematical research that has **solved multiple open mathematical problems**, orchestrating specialized explorer, verifier, and consultant agents over a rich library of **skills and MCP tools**, with per-role capabilities statically enforced by the harness rather than governed by prompts
- Designed a **persistent graph-based memory system** maintaining an evolving exploration graph across long-running executions — preserving intermediate results and competing proof directions — with **self-evolving** memory curation and iterative planning that prioritize high-value branches and adapt problem-solving strategy

NLP Lab UCLA (Natural Language Processing)

Research Assistant

Los Angeles, CA

Mar 2026 – Present

- Built a **privacy-preserving safety guard** for LLM-based multi-agent systems, training an edge-featured graph attention network over inter-agent communication graphs via **federated learning** to detect and block risky agents without sharing raw interaction traces
- Outperformed the **centralized-training ceiling** on 3 agent-safety benchmarks and cut live attack-success rate by **43%** at above-unguarded utility as a 0.5M-param, ~50 ms CPU sidecar; first-author paper in submission to **AAAI 2027** (arXiv:2609.02967)
- Amortized per-query MAS topology design into a **16-entry VQ-VAE codebook** with a reward-weighted code predictor and an execution-grounded MLP reranker, achieving **best accuracy on all 6 benchmarks (84.6 vs. 83.0 best prior)**, **> 120×** faster than graph-diffusion designers (**2.4 ms**) with **21.9–33.2%** fewer LLM tokens; first-author paper in submission to **AAAI 2027** (arXiv:2609.02264)

SCAI Lab UCLA (AI and Data Mining)

Research Assistant

Los Angeles, CA

Feb 2026 – Present

- Developed **Gaussian speculative decoding** in PyTorch, pairing lightweight draft models with single-pass parallel verification to accelerate inference of pretrained time series foundation models by up to **3.1×** across 5 model families without retraining
- Proved a **capped-geometric throughput law** making speedups predictable before deployment, and a **bounded-deviation guarantee** on forecast fidelity to the target model; first-author paper in submission to **AAAI 2027** (arXiv:2511.18191)

Yiyue Intelligent Technology Co., Ltd.

Full-Stack Developer Intern

Nanjing, China

Jul 2025 – Sep 2025

- Developed key modules (grade & course management, role-based notifications) for a university web using **Vue.js** and **Java/Spring Boot**
- Used **Spring Data JPA** to perform CRUD operations on **MySQL** without raw SQL, ensuring data integrity across **50+** test cases
- Collaborated via **Git** feature branches and code reviews; documented **REST APIs** to streamline frontend-backend integration

PROJECT HIGHLIGHTS

EgoX - AI-Powered Personal Career Wiki | Next.js, Supabase (Postgres + pgvector), Vercel AI SDK, Drizzle ORM

Sole Developer

Apr 2026 – Present

- Built a **multi-tenant AI career-wiki SaaS** that converts raw career material into a structured knowledge base via **LLM structured extraction** (Zod-constrained outputs), powering **grounded RAG chat** with inline citations (HNSW vector search + knowledge-graph expansion), source-traceable resume generation, and voice mock interviews (TTS/STT) with **LLM-as-judge** rubric scoring
- Engineered a **sealed LLM gateway** metering every call end-to-end: **atomic budget reservations** via single conditional UPDATE (eliminating check-then-act races), full cost audit trail, tiered model routing (~ **5×** cost cut), and batched embeddings cutting ingest latency **15s → 600ms**; shipped **Stripe** billing with idempotent webhooks, **Postgres**-backed rate limiting, and server-side **multi-tenant isolation**
- Migrated long-running LLM jobs (graph build, repo ingest) from request handlers to an **asynchronous worker pipeline** on **AWS SQS + Lambda/Fargate** with a job-state table, retries and dead-letter queue, and idempotent workers, eliminating **serverless timeout** failures; provisioned via **AWS CDK**, with GitHub Actions **CI** (typecheck/lint) and a scheduled billing reconciliation job alerting on cost drift

Mock-Trader | Trading Simulation Platform & Agent Evaluation Sandbox

Sole Developer

Dec 2025 – May 2026

- Built a trading platform with **Node.js/TypeScript**, **PostgreSQL** trade ledgers, position reconstruction, rule engines, live market data
- Open-sourced an **npm MCP server** exposing **16 tools** for AI trading-agent evaluation: order execution with fees/slippage, prop-firm-style risk rules, and strict replay backtests on Alpaca/Polygon/CSV data, hiding future bars server-side to prevent **look-ahead bias**

SKILLS

Languages: Python, C++, Java, Go, JavaScript, TypeScript, HTML/CSS

Frameworks/Tools: PyTorch, CUDA profiling, Node.js, React, Vue, Redis, Pandas, Spring Boot, Bash, Git, MySQL, Docker, PostgreSQL