

JIAJUN (NICK) HUO

jiajunhuo726@gmail.com · 267-777-8518 · github.com/nickhuo · nickhuo.com · IL (UTC-6)

EDUCATION

University of Illinois Urbana Champaign

Aug 2024 – May 2026

Master of Science in Information Science (STEM)

Urbana, IL

GPA 3.9/4.0 · A in Applied Machine Learning & Database Systems · U.S. News #1 Information Systems program

Shenzhen University

Sep 2019 – Jun 2023

Bachelor of Science in Mathematics

Shenzhen

GPA 3.79/4.50 (Top 10 %) · 2× merit scholarships · 4 ML-modeling wins in supply-chain, finance and computer-vision tracks

WORK EXPERIENCE

Donut Labs

Jun 2025 – Aug 2025

Software Engineer Intern – 1st Crypto Agentic Browser (Sequoia-backed)

Remote · NYC HQ

- Cut inference cost 99.5% to 0.003¢ by engineering GPT-seeded ANN pipeline that processes 50 k tweets/day for sentiment
- Scaled 18-route FastAPI microservice to 3× QPS with P95 220 ms; reduced 4xx/5xx 40%, frontend adoption to 66%
- Deployed prompt-tagged Pinecone vector store that auto-labeled 82% tweets, cut GPT calls 80% with same accuracy
- Designed NestJS microservice mesh with 3 MCP endpoints, isolating LLM ops and boosting crypto-agent throughput
- Containerized stack with multi-stage Docker & GitHub Actions on AWS EKS, shrinking deploy cycle 20 min → <5 min

Sonic SVM

Mar 2023 – Jul 2024

Software Engineer – Bitkraft-backed SaaS for game studios

Shenzhen

- Led the development of data governance system, drove 3 C-level strategic decisions with quant insights, report to CPO
- Scaled data sources by 2x and architected ETL pipelines on GCP, ensuring 99.9% accuracy with optimized SQL queries
- Automated API log extraction workflow from PostgreSQL to BigQuery using Airflow, reducing 43.2% processing time
- Instrumented Prometheus + Grafana; live KPIs paged execs, steering roadmap and lifting 10+ deals

Tencent

Nov 2020 - May 2021

Technical Product Manager Intern – The largest WeChat ecosystem platform for universities

Shenzhen

- Collaborated 5 interns enginners through agile sprints, launching a campus mini-app from concept to prod in 12 weeks
- Designed event-driven notification stack (Pub/Sub + WebSocket) that serves 50 k students with <2 s latency, 0 downtime
- Set clear success metrics and weekly experiments; user insights shaped feature priorities, lifting student engagement 22%

PROJECTS

ReplicaGenie – Real-time, cost-efficient RAG platform for personalized LLMs

- Architected 4-service LLM platform; modular microservices allow independent scaling, isolated failures & fast rollouts
- Delivered 4-bit LoRA fine-tuning on SageMaker, trimming GPU training cost 77% and guaranteeing reproducibility
- Built MongoDB → RabbitMQ CDC stream pushing changes to feature store in <1 s, enabling real-time personalization
- Automated scraping for 4 data sources via Selenium on AWS Lambda, parsing and persisting structured data in MongoDB

BiteMatch – Smart recipe recommender, ranked top-10% in class

GitHub · Demo

- Created RESTful API with 12+ endpoints handling user auth, recipe CRUD, video search, and review system
- Reduced YouTube API calls by caching videos in Redis and using a quota-aware fetch logic to avoid exceeding limits
- Worked with 4-member Agile squad, grooming backlog and delivering on 2-week sprints with 100% sprint goal hit rate

Claude Code Extension for Raycast

GitHub

- Built TypeScript & React Raycast extension (5 modules) transforming Claude CLI into a GUI, accelerating dev workflows
- Implemented JSONL session search with real-time filters, plus memory limits and timeout guards for reliable performance

SKILLS

Languages Python, Go, TypeScript, JavaScript, Rust, Java, SQL, Bash

Frameworks FastAPI, LangChain, LangGragh Django, Flask, React, Next.js, Node.js, Express.js, GraphQL

DevOps AWS, Docker, Kubernetes, Terraform, Helm, GitHub Actions, CI/CD, Prometheus, Grafana, Airflow