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 Database Systems & Applied Machine Learning · 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 national awards in ML modeling (supply-chain, finance, computer-vision)

WORK EXPERIENCE

Donut Labs Jun 2025 – Aug 2025

***Software Engineer Intern** – 1st AI Agentic Crypto Browser (Sequoia-backed)* *Remote · NYC HQ*

- Orchestrated an autonomous crypto agent integrating 46 actions across 16 providers, boosting task completion rate by 22%
- Mitigated tool-use hallucination risks by implementing constrained decoding, reducing invalid tool-call errors by 85%
- Engineered a context reuse strategy with Redis caching for model intermediate states, cutting GPU costs by $2K/month
- Replaced a costly GPU-based model with RAG pipeline for 50K daily inputs, reducing daily inference cost by 92.5%
- Architected and scaled 18 FastAPI microservices, handling at a P95 latency of 220ms to support agent's reasoning core

Sonic SVM Mar 2023 – Jul 2024

***Software Engineer, Data Solution** – Bitkraft-backed SaaS for game studios* *Shenzhen*

- Led the development of data governance system, drove 3 C-level strategic decisionson on product roadmap, report to CPO
- Scaled data platform 2× by integrating new sources while maintaining <1hr SLA through incremental processing
- Automated PostgreSQL→BigQuery ETL workflows, cutting data processing time 43% and enabling real-time analytics
- Created Prometheus + Grafana dashboards tracking KPIs that influenced product roadmap and secured 10+ partnerships

Tencent Nov 2020 - May 2021

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

- Drove the full SDLC for campus app, mitigating risks and unblocking dependencies to launch from scratch in 12 weeks
- Translated product requirements into technical specifications for a real-time notification system serving 50k+ students
- Leveraged user engagement metrics from A/B tests to guide feature priorities, boosting student engagement by 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

- Owned the backend development, creating a RESTful API with 12+ endpoints for user auth, CRUD, and video search
- 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, TypeScript/JavaScript, Go, SQL, Java, HTML/CSS, Bash/Shell
Technologies	Backend: FastAPI, Django, Node.js | Frontend: React, Next.js | Data: PostgreSQL, MongoDB, Redis
Cloud/DevOps	AWS (EC2, S3, Lambda, SageMaker), GCP, Docker, Kubernetes, Terraform, Kafka, CI/CD, Prometheus

Seeking Fall 2025 internship · 2026 new-grad SDE roles · Open to relocation · CPT/OPT-eligible (U.S.)