JIAJUN (NICK) HUO

jiajunhuo726@gmail.com · 267-777-8518 · github.com/nickhuo · nickhuo.com

EDUCATION

University of Illinois Urbana Champaign — Aug 2024 – May 2026
Master of Science in Information Science (STEM) — *Urbana, IL*

GPA 3.88/4.0 · Distributed Systems & Database Systems & Applied Machine Learning · U.S. News #1 Information Systems

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*

- Governed MCP tool-use across 46 actions with enforced OpenAPI contracts and rollout guards, raising task success 22%
- Engineered a context reuse strategy with Redis caching for model intermediate states, cutting GPU costs by $2K/month
- Built a self-improving RAG pipeline with data flywheel on LanceDB, decreasing real-time per-tweet analysis costs by 86%
- Designed LangSmith tracing + closed-loop eval (BLEU/BERTScore); 100% PR coverage, regressions flagged <24 h

Business Intelligence Group — Aug 2025 – Present
Software Engineer & Consultant (Capstone) — *Urbana, IL*

- Deployed an AI chatbot for airline pilots to accelerate access to flight data, directly addressing operational workflow gaps
- Implemented CI/CD pipelines with Jenkins and Docker on AWS, cutting deployment time by 70% and improving reliability

Sonic SVM — Mar 2023 – Jul 2024
Software Engineer, Data Solution – Bitkraft-backed SaaS for game studios — *Shenzhen*

- Launched a self-service analytics platform on GCP with Looker, enabling product teams to answer 80% of data questions
- Scaled data ingestion 10x using Kafka & Spark to support 200% user growth and international product expansion
- Automated ETLs with Airflow & dbt, cutting daily processing time from 4 hours to 15 mins and saving $7K+ monthly
- Built monitoring with Prometheus & Grafana, reducing critical data incident response time from hours to under 5 minutes

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

PROJECTS

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

- Delivered 4-bit LoRA fine-tuning on SageMaker, trimming GPU training cost 77% and guaranteeing reproducibility
- 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

Go KV storage

- Implemented a SkipList-based in-memory index, enabling support for complex data types (lists, zsets) and efficient lookups
- Designed a concurrent compaction process using goroutines to merge log files and reclaim disk space from obsolete records
- Optimized disk I/O with memory-mapped files and built a client-server architecture with a custom TCP-based protocol

SKILLS

Languages Python, TypeScript/JavaScript, Go, SQL, Java, Scala, Rust, Swift, HTML, CSS, Bash/Shell
Technologies Backend: SpringBoot, Django, Node.js | Frontend: React, Next.js | Data: PostgreSQL, MongoDB, Redis
Cloud/DevOps AWS (EC2, S3, Lambda, Dynamo), Git, Linux, Terraform, Docker, Kubernetes, Kafka, Prometheus