JIAJUN (NICK) HUO

[Email] jiajunhuo726@gmail.com · [Phone] 267-777-8518 · [GitHub] github.com/nickhuo · [Web] nickhuo.com

EDUCATION

University of Illinois Urbana Champaign Aug 2024 - May 2026

Master of Science in Information Science (STEM) *Urbana, IL*

- GPA 3.9/4.0 · Applied Machine Learning & Database Systems (both A) · Program #1 in Info Systems (U.S. News)

Shenzhen University Sep 2019 - Jun 2023

Bachelor of Science in Mathematics *Shenzhen*

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

WORK EXPERIENCE

Donut Labs Jun 2025 - Aug 2025

Software Engineer Intern *Agentic Browser Backed by Sequoia Capital*

- Developed a high-performance sentiment analysis pipeline utilizing GPT and ANN, processing 50k tweets daily and reducing inference costs by 19× to $0.003 per tweet.
- Created a prompt-tagged Pinecone vector database, achieving 82% tweet labeling accuracy while decreasing GPT API calls by 80%, maintaining consistent performance.
- Designed and implemented an 18-endpoint RESTful FastAPI microservice for LLM inference and content operations, enhancing query performance by 3× with a P95 latency of 220ms; integrated auto-Swagger for streamlined frontend collaboration, reducing integration time by 66% and cutting error rates by 40% through Pydantic validation.
- Established LangSmith observability across the agent stack, developing a closed-loop LLM evaluation framework with automated BLEU/BERTScore metrics and human review, ensuring 100% PR coverage and identifying regressions within 24 hours.

Sonic SVM Mar 2023 - Jul 2024

Software Engineer - Data Solutions *Bitkraft-backed SaaS for 50+ Game Studios, $100M valuation*

- Spearheaded the development of a data governance system, influencing 3 C-level strategic decisions through actionable insights reported directly to the CPO.
- Enhanced data source capacity by 2x and designed robust ETL pipelines on GCP, achieving 99.9% data accuracy through optimized SQL queries.
- Streamlined API log extraction from PostgreSQL to BigQuery using Airflow, resulting in a 43.2% reduction in processing time.
- Engineered a Pub/Sub to Dataflow streaming pipeline, delivering in-game KPIs to BI dashboards with a latency of under 5 seconds.
- Developed and launched Looker Studio dashboards for a team of 7, accelerating insight turnaround from 2 days to just 30 minutes.

PROJECTS

ReplicaGenie

- Architected 4-service LLM platform; modular microservices allow independent scaling, isolated failures & fast rollouts
- Automated LinkedIn/GitHub scraping via Selenium on AWS Lambda; parsed & stored structured data in MongoDB
- Built MongoDB → RabbitMQ CDC stream pushing changes to feature store in <1s, enabling real-time personalization
- Delivered 4-bit LoRA fine-tuning on SageMaker, trimming GPU training cost 77% and guaranteeing reproducibility
- Integrated Opik tracing and auto-eval, detecting hallucination & relevance and lifting LangChain-SageMaker reliability
- Orchestrated 6-service Docker stack-MongoDB replica, RabbitMQ, Qdrant—delivering production-ready RAG

BiteMatch

- Led a 4-member Agile squad, grooming backlog and delivering on 2-week sprints with 100% sprint goal hit rate
- Top 10% in course. Developed smart recipe app to personalize meal choices using React, Node.js, and MySQL
- Created RESTful API with 12+ endpoints handling user auth, recipe CRUD, video search, and review system
- Reduced YouTube API calls by caching videos in DB and using a quota-aware fetch logic to avoid exceeding limits
- Containerized services with Docker & GitHub Actions; push-to-prod cycle shrank from 20 min to <5 min

SKILLS

- **Programming**: Python, TypeScript, Go, JavaScript, SQL, Zsh & Bash, Node.js, React
- **Frameworks**: React, Next.js, Jenkins, Tailwind CSS, Vite, Node.js, Django, Express.js, MongoDB, Redis
- **DevOps**: AWS, Docker, Terraform, GitHub Actions, Vercel, Airflow, Linux/Unix, Git

Actively seeking Fall 2025 intern opportunities. Willing to relocate. Authorized to work in the US