Xirui Huang

514-690-8818 x36huang@uwaterloo.ca linkedin.com/in/xrhuang10 github.com/xrhuang10

Education

University of Waterloo **Sep 2023 - Apr 2027**

BASc. in Systems Design Engineering, Option in AI, Intelligent Systems Specialization *3.93 cGPA, 91%*

First in Class Scholarship, President's Scholarship, Dean's Honor List 4x, EngSoc Student Council Cohort Leader

Technical Skills

Coding: Python, JavaScript/Typescript, C++, SQL, HTML/CSS, React, FastAPI, Flask, Next.js, Tailwind
AI/ML: PyTorch, TensorFlow, LangChain, Hugging Face, scikit-learn, pgvector, Pinecone, Runpod
Systems: Docker, Kubernetes, AWS (Lambda, EC2), GCP Suite, Auth0, Trigger.dev, Doppler, Terraform
Data: PostgreSQL, Snowflake, AWS (s3, Dynamodb), Databricks, GCP (Bigquery), MongoDB, Supabase, Neon

Experience

Sequence Holdings **Jan 2026 – Apr 2026**

AI Software Engineering Intern - Forward-Deployed *New York, New York*

- Architected a chat mining pipeline that **clusters 40K+ client conversations** via HDBSCAN on vector embeddings, orchestrated through Trigger.dev and Databricks, surfacing recurring and failing workflows to drive template creation
- Rebuilt templates platform from flat list to marketplace with folders, sharing, save/upvote, and AI suggestions, migrating CRUD flows to side panel UIs, writing backfill scripts, and shipping full-stack with zero-downtime feature-flag rollout
- Developed a conversational template generator agent supporting multimodal input with multi-turn refinement and role-aware MCP integration suggestions using Langchain tool, **adopted by over 80% of users** at portfolio company
- Built a chat adoption analytics platform with BigQuery RBAC (user department and role permissions synced to Auth0), real-time Slack alerts on negative feedback, and a client-facing leaderboard on GCP, MongoDB, and Terraform
- Rebuilt portfolio company's user directory by consolidating 3 data sources with cross-platform datastream replication, and cross-env data sync pipelines, **migrating 4 services with zero downtime** and fixing on-call production issues

Wisedocs **May 2025 – Sep 2025**

Machine Learning Intern *Toronto, Ontario*

- Developed an LLM-powered entity extraction system for OCR-scanned PDFs, **obtaining 96% accuracy** by integrating Claude API using structured JSON output and building dataset sourced from Snowflake SQL query via Boto3
- Designed a FastAPI + PostgreSQL prompt store with versioning and multi-field search, deploying to Kubernetes via CI/CD pipelines and built testing suite with Terraform-based Grafana dashboards and E2E pytests
- Fine-tuned Google Gemini on a multimodal task, outperforming prompt engineering baselines and achieving **91% inference accuracy** while producing a cost-benefit analysis against other LLMs based on cost, accuracy, and latency

Tesla **Sep 2024 – Dec 2024**

TPM Cell Engineering Intern *Austin, Texas*

- Coordinated a cross-functional team of 20+ engineers to install high-speed vision upgrades to an existing production line for a cell defect tracking system, **decreasing wasted cathode foil by 5.1%** for the Cybertruck battery cells
- Led on-time install of 3000+ vision parts by consolidating BOMs through efficient data organization and **acquiring over $2.6M** worth of materials and services by driving the RFQ and bidding processes for next gen machine upgrades

Research

SSDP - NeurIPS AI Research Paper | *GPUs, LLMs, Hugging Face, Runpod, FastAPI*

- Co-authored SSDP, a Tree-of-Thought pruning framework achieving 2.3x speedup with 85-90% node reduction for LLM reasoning tasks, accepted to NeurIPS Efficient Reasoning Workshop and **cited by Google Deepmind researchers**
- Benchmarked ToT search algorithms on GSM8K, MATH500, and AIME500 (HF datasets) with LLaMA and Qwen using verifier, embedding, and policy models, deploying FastAPI eval harness on RunPod containerized GPU clusters

SickKids Research Institute - ML Researcher | *U-Net, V-Net, Pytorch, Neural Networks*

- Training 2D U-Net and 3D V-Net on 1,000+ MRI/CT slices to segment brain AVMs in pediatric patients, achieving 85% DICE score, automating a step in neurosurgical radiation planning, ensuring compliance via SSH logins on HPC clusters