Henry Wang

647-877-9988 | h352wang@uwaterloo.ca | linkedin.com/in/henry-w-se | github.com/hwang2409

Education

University of Waterloo Expected April 2026
Bachelor of Software Engineering (BSE)

Experience

Software Engineer Intern Jan 2026 – Present
Fish Audio *San Francisco, CA*

- Built an agentic video creation pipeline in React/TypeScript with Zustand state management, orchestrating script generation, asset selection, and TTS audio synthesis into an auto-populated timeline.
- Designed a script editing interface using Next.js with drag-and-drop line reordering, per-character voice assignment, and selective regeneration, giving users granular control over AI-generated TTS before final render.
- Built the speech-to-text product page end-to-end in Next.js with file upload, live browser via WebRTC, and async transcription polling against a Replicate-hosted ASR model.

Software Engineer Intern Apr. 2025 - Aug. 2025
NationGraph *San Francisco, CA*

- Built an entity resolution ML pipeline in PyTorch to normalize 600M+ vendor names, combining TF-IDF embeddings with custom supervised classification models to deduplicate and standardize records at scale.
- Built dynamic Python web scrapers using Selenium and BeautifulSoup to extract and normalize public procurement records from government agency websites, storing structured data in PostgreSQL.
- Reduced analytics query latency by 38% by rewriting PostgreSQL joins, adding targeted B-tree indexes on high-cardinality columns, and introducing a Redis caching layer for frequently accessed procurement datasets.

Projects

Neuronic | *React, FastAPI, PostgreSQL, Redis, Celery, AWS* Jan. 2026 - Mar. 2026

- Deployed an AI-native learning platform on AWS with an async FastAPI backend behind an ALB, Redis-backed session and embedding caches, and a React 19 SPA on CloudFront, maintaining a sub-200ms p95 API latency.
- Built a hybrid retrieval engine combining fastembed vector search with full-text PostgreSQL indexes, powering the automatic knowledge graph construction across the full note corpus.
- Designed a multi-tenant collaboration system with RBAC, edit-suggestion workflows, real-time group feeds, integrated with Google Calendar with OAuth 2.0 and a webhook event bus.

whitematter | *C++, Python, FastAPI, Next.js, CUDA, AWS* Nov. 2025 – Feb. 2026

- Built a deep learning framework in C++ with automatic differentiation, 20 layer types (Conv2D, MultiHeadAttention, LSTM) and GPU backends for Metal and CUDA.
- Implemented SIMD-optimized tensor ops (AVX/NEON) and OpenMP-parallelized GEMM hitting 99%+ MNIST accuracy in 3 epochs.
- Designed a browser-based training platform where users describe a model in natural language, agents generate the model architecture, and the system compiles and runs optimized C++ training code with real-time loss/accuracy streaming via SSE.
- Shipped as a multi-stage Docker container with a FastAPI job queue, distributed training workers, ONNX export, mixed-precision FP16, and one-click deployment to AWS EC2 GPU instances for inference saving.

Technical Skills

Languages: Python, TypeScript, JavaScript, C/C++, SQL (Postgres)
Frameworks: React, Node.js, Flask, FastAPI
Developer Tools: Git, Docker, Google Cloud Platform, Jira, AWS
Libraries: pandas, NumPy, Matplotlib, PyTorch

Awards

National Champion of the Hypatia Math Contest (1/5627)
Score of 124.5 on the AMC12 (Top 5% out of 140,000 participants)
Bronze Medal on the CLMC (Hosted by the University of Waterloo)