Stephen Huang

s395huan@uwaterloo.ca | stephenhuang3.github.io | StephenHuang3 | stephenhuang1

Education

University of Waterloo — *Waterloo, Ontario*

Bachelor of Computer Science with AI Specialization — *Sep 2021 - Apr 2026*

Courses Machine Learning, Computer Vision, NLP, Object Oriented Programming, Data Structures, Algorithms, Statistics, Linear Algebra

Skills

Languages Python, Java, C/C++, SQL, Airflow, JavaScript/TypeScript, Linux, R, Bash, Git, Scala, Docker, Jira, Terraform

ML Tensorflow, PyTorch, Keras, JAX, OpenCV, Pandas, NumPy, Transformers, NLTK, Sklearn, VertexAI, CNN, KNN, ANN

Work Experience

Amazon — *Seattle, Washington*

Database Engineer — *Sep 2025 – Dec 2025*

- Automated the deployment of Time to Live (TTL) feature that deletes entries after a specified time without using customer write units, saving customers **$30 million** in costs
- Reduced **DynamoDB** scan units by implementing a time sorted Local Secondary Index, reducing scan volume by **20%**
- Improved test cases for DynamoDB rate limiter to validate correctness under burst traffic, retries, and partial failures

Meta — *Bellevue, Washington*

Machine Learning Engineer — *May 2025 – Aug 2025*

- Increased anti-scraping effectiveness by fine-tuning LLM detection models, reducing **$3 million** in fines
- Built data pipeline to finetune **Llama 3.3** model on **20,000** Instagram sessions to detect scraping, improving F1 score by **4%**
- Integrated multitenant architecture to **16** data pipelines that process a combined **1,000,000** sessions daily, reducing cross team dependencies, pipeline costs by **30%**, and development time by **50%**
- Designed LLM probability model to detect scraping using **probability tokens**, allowing users to optimize precision or recall

Compass Digital — *Toronto, Ontario*

Data Scientist — *May 2024 – Aug 2024*

- Developed a Snowflake **data workflow**, integrating millions of sales data from AWS and GCP storage for data visualization
- Fine tuned **LLMs** such as PaliGemma and Gemini with **LoRA** and **Adapter** based techniques, increasing BLEU score by 88%
- Achieved an 85% F1 Score using **GPT-4o** multimodal capabilities to determine the empty percentage of a grocery shelf
- Trained **deep learning** drink detection models using manually labeled Google VertexAI datasets, achieving a MIoU of **94%**
- Developed a model training pipeline with **Python**, integrating VertexAI datasets, GCS storage, and model hyperparameters
- Achieved an ARI of **0.92** using Sklearn **clustering** algorithm that groups bounding boxes into their respective row
- Created an **Apache Airflow DAG** to automatically update models datasets with recently uploaded data

Met-Scan — *Toronto, Ontario*

Machine Learning Researcher — *Sep 2023 – Dec 2023*

- Researched tracking crowd density in public areas using a late fusion **multimodal deep learning** approach that combines self-collected image and sensor data from security cameras and Bluezone beacons
- Achieved an IoU of **90%** building a person detection model by creating a Euclidean distance algorithm with **ImageAI**
- Architected a combined **CNN** and **ANN** approach to indoor localization with sensor data using **TensorFlow** and **Keras**, achieving an **84%** mAP
- Built data pipeline to process **10,000+** raw signal data points from sensors with **Numpy** and **Pandas**
- Reworked API calls to **multithread** data collection from multiple beacon sensors simultaneously for model training

Marsh McLennan — *New York, New York*

Software Developer — *Jan 2023 – Apr 2023*

- Worked in an **agile** team to expand a **60,000+** user **MERN** (MongoDB, Express, React, Node.js) stack application by implementing fuzzy search capabilities, adding dashboards, and reducing page load times by **80%** with pagination

Projects

HumorLens

Analyzes Humor in The New Yorker comics — *2024*

- Developed an advanced image captioning model using CNN and **Transformer** technologies in TensorFlow and Keras by integrating and processing multiple datasets including Flickr8k and MS COCO with Numpy and Pandas
- Applied **transfer learning** techniques to adapt the image captioning LLM model to the New Yorker Caption Contest dataset, resulting in a 30% increase in humor relevance in generated caption