Richard Bai

437-226-7128 | r25bai@uwaterloo.ca | LinkedIn | Github | Personal Website

Education

University of Waterloo — Waterloo, ON
Bachelor of Computer Science — *Sep 2023 – May 2027*

Experience

Software Engineering Intern — May 2025 – August 2025
Shopify — *Toronto, Ontario*

- Incoming **Summer 2025** Software Engineering Intern

Software Engineering Intern — January 2025 – April 2025
Trend Micro — *Ottawa, Ontario*

- Eliminated **$10,000** in annual AWS and Splunk costs by redesigning the **CloudWatch** alarm and **SNS** forwarding microservices, filtering misleading **AWS RDS** CPU, memory and I/O metrics to prevent unnecessary scaling events.
- Reduced **PostgreSQL** and **AWS RDS** storage usage by **20%+** by developing a **Lambda** microservice to fold and cache high-frequency events with **Redis**, minimizing redundant inserts and optimizing historical data queries
- Improved inference accuracy and reduced OpenAI expenses by **$750/month** by optimizing bottlenecks in the **LangChain**-based ASRM companion, utilizing prompt structure caching to lower response latency from **2+ min** to **25 sec**

Full-Stack Engineering Intern — May 2024 – August 2024
Savi Finance — *Toronto, Ontario*

- Expanded user-base by **9,000+ MAU** and improved upload speeds by **17x** by remastering CSV/PDF upload system with **GPT-4o**, **LlamaIndex**, **AWS S3** and **Redis** to accurately parse/remap transactions in a concurrent, mutex-based worker
- Automated **CI/CD** staging/production deployments for **3** micro-services using **Docker, Kubernetes, AWS EKS, AWS ECR** and **cron**, reducing deployment times by **60%+** and ensuring zero-downtime deployments
- Integrated **Plaid** financial accounts into a concurrent web worker for expensive computations, increasing front-end responsiveness by **2500ms+** while performing rigorous E2E-testing with **Jest** and **Cypress** in an **Agile** environment

Software Development Intern — May 2022 – January 2023
Freedo Technologies — *Remote*

- Innovated a patented process using **PyTorch** and **OpenCV** to identify and analyze different roof shapes and their respective dimensions and rebuild them in 3D space using **C++**, speeding up roof processing by **86%**
- Repurposed an unsupervised ML algorithm to condense satellite images of buildings to its most dominant colors, speeding up 3D modelling processing speeds by **40%+** and reducing project processing loads by an average of over **65%**

Projects

elitecode | *Plasmo, React, D3.js, Express, Docker, AWS ECS, MongoDB* — September 2024 - Present

- Developed a web platform and browser extension to make learning LeetCode easier with **Plasmo**, **D3.js** and **React**, allowing users scrape, visualize and practice **3000+** Leetcode problems with an interactive graph
- Innovated a recommendations engine with **PineCone** and **OpenAI ADA-002**, in-line code feedback and question hints with **OpenAI-4o-mini** and **Tavily API**, and continuously deploying it with **GitHub Actions**, **AWS ECS** and **Docker**

marketloo | *Next.js, D3.js, Supabase, PostgresQL, Redis, RabbitMQ* — January 2025 - Present

- Designed a real-time predictions market with **Next.js**, **D3.js**, **Redis Pub/Sub** and **RabbitMQ** to manage subscription channels, **Supabase** for authentication and storage and scaling it to **1500+** users attending **Nosu AI Hack** with **Vercel**
- Provisioned **5** AI Agents to act as market makers, scheduling **cron** jobs to continuously rescrape Devpost pages with **Selenium**, using **OpenAI-4o-mini** and **Redis Pub/Sub** to queue tasks and decouple the agents

Tennis Tracker AI | *PyTorch, OpenCV, Python, Ultralytics, Roboflow* — June 2024 - Present

- Fine-tuned an **Ultralytics YOLOv5l6u** object detection model using **Roboflow** datasets to analyze tennis matches, tracking player and ball movements to compute and provide key match analytics
- Customized a **convolutional neural network (CNN)** built with **PyTorch** and **OpenCV** to extract and map key-points to interpolate player and ball positions, converting them to relative velocities in real-time

Technical Skills

Languages: JavaScript, TypeScript, HTML/CSS, Python, Java, C, C++, Bash, Groovy, YAML, SQL
Libraries/Frameworks: React, React Native, Next.js, Node.js, PyTorch, scikit-learn, Tailwind
Developer Tools: AWS, Terraform, Serverless, Git, Docker, Jenkins, Jira, Confluence, Slack, Figma
Testing Tools: Jest, Cypress, Storybook, Postman, Junit