Ri Hong

ri.hong4@gmail.com | github.com/Ri-Hong | linkedin.com/in/ririhong
Software Engineer · GenAI & LLM Development · AI/ML Systems

Education

University of Waterloo 2022 – 2027
Bachelor of Computer Science (Co-op) · GPA: 3.9/4.0

- Relevant coursework: Object-Oriented Programming, Artificial Intelligence, Machine Learning, Computer Vision

Skills

Languages: Python, Java, C++, Go, C, TypeScript, JavaScript, HTML, CSS
Technologies: Kubernetes, Terraform, Docker, React, Next.js, AWS, GCP, gRPC, Temporal, Linux, PostgreSQL, FastAPI
AI/ML: PyTorch, TensorFlow, scikit-learn, LangChain, OpenAI GPT-4/5, CUDA, MLflow, DVC

Experience

Groq Sept 2025 – Dec 2025
Software Engineering Intern *Toronto, ON*

- Engineered multi-region infrastructure with **Terraform** and **Kubernetes** on GCP, enabling horizontally **scalable AI/ML** workloads and cutting deployment time by 24%; followed **SDLC best practices** including **code reviews** and **CI/CD** pipelines.
- Designed observability platform measuring **AI** coding tool adoption for 400+ engineers; delivered technical demos to senior engineers, communicating **technical concepts** using **clear language** and **visual** demonstrations.

Walnote.ai Aug 2025 – Present
Founder & CTO *Toronto, ON*

- Developed production **GenAI** platform using **LLM/VLM** (**GPT-5**) for automated video generation; built **agent services** coordinating **VLM** processing, processing 1,000+ animations through **distributed Python + FastAPI** architecture.
- Architected **SDLC** workflows; accelerated generation with **distributed** GPU rendering, reducing render latency from 12s to 2s using **fault-tolerant** streaming pipelines.

Base Power Jan 2025 – Apr 2025
Software Engineering Intern *Austin, TX*

- Transitioned trading simulations to **cloud-native** execution using **Temporal** Cloud, enabling 1000s of workflows in a **distributed computing environment**; participated in **code reviews** and **technical documentation** following **SDLC** principles.
- Implemented **Protobuf** and **gRPC** for **Python** and **Go microservices** communication, reducing serialization overhead by 20%; maintained **maintainable code** following **design patterns** and **CI/CD best practices**.

Projects

DistilBERT Sentiment Analysis | *PyTorch, GCP, Kubernetes, Terraform* Sept 2025

- Engineered a production-grade **AI/ML** sentiment analysis service achieving 92.5% accuracy using DistilBERT, deployed on GKE with **Terraform** in a **cloud-native** architecture; implemented end-to-end MLOps pipeline with MLflow, DVC, and BentoML for model serving.
- Optimized training with mixed precision and **distributed** GPU training with **Kubernetes**, reducing training time by 40% while maintaining model accuracy in a **scalable**, **cloud-native** architecture.

Neural Style Transfer Engine | *PyTorch, CUDA, FastAPI, Docker* Sept 2025

- Engineered a high-performance **AI** system with custom **CUDA** kernels for Gram matrix computation, achieving 8x faster style transfer; built **scalable** REST API with **FastAPI** and containerized with **Docker** for **production** deployment.

FlaimBrain | *LangChain, GPT-4, MongoDB, Flask* Oct 2023

- Engineered a **GenAI** study assistant using **LLM/VLM** (**GPT-4**) with **LangChain** for **agent** orchestration; developed **agent services** with RAG architecture integrating **VLM** capabilities for document understanding, demonstrating end-to-end **GenAI** and **agent service development** with **LLM/VLM** technologies.