Ri Hong

ri.hong16@gmail.com | github.com/Ri-Hong | linkedin.com/in/ririhong

10× Hackathon Winner · Applied Machine Learning Engineer · AI/ML Systems · Cloud Infrastructure

EDUCATION

University of Waterloo 2022 – 2027

Bachelor of Computer Science (Co-op) · GPA: 3.9/4.0

– Relevant coursework: Algorithms, Data Structures, Object-Oriented Programming, Databases, Distributed Systems, Operating Systems, Artificial Intelligence, Machine Learning, Computer Networks

SKILLS

Languages: Python, Go, Java, C++, TypeScript, JavaScript, C
Technologies: Kubernetes, Terraform, Docker, AWS, GCP, gRPC, Temporal, Linux, PostgreSQL, React, Next.js
AI/ML: PyTorch, TensorFlow, scikit-learn, LangChain, OpenAI GPT-4/5, CUDA, MLflow, DVC
Tools: Jenkins, Git, OpenTelemetry, CI/CD, Monitoring & Observability

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 in a **distributed computing environment** and cutting deployment time by 24%; participated in **code reviews** and maintained **technical documentation** following **SDLC best practices**.
– Designed and launched observability platform measuring engineering productivity and **AI** coding tool adoption, providing actionable insights for 400+ engineers; demonstrated **operational excellence** through monitoring and troubleshooting production systems.

Walnote.ai Aug 2025 – Present

Founder & CTO *Toronto, ON*

– Developed production **GenAI** platform using **LLM/VLM** (**GPT-5**) for automated video generation; built **distributed Python + FastAPI** architecture processing 1,000+ animations in a **cloud-native**, **scalable** system.
– Accelerated generation with **distributed** GPU rendering, reducing render latency from 12s to 2s using **fault-tolerant** streaming pipelines; practiced **SDLC** workflows and **CI/CD** principles.

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%; wrote **clean**, **maintainable code** following **design patterns** and **CI/CD best practices**.

PROJECTS

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

– Engineered 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**, **fault-tolerant** system.

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

– Engineered high-performance **AI** system with custom **NVIDIA 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 **GenAI** study assistant using **LLM/VLM** (**GPT-4**) with **LangChain** for **agent** orchestration; developed RAG architecture integrating **VLM** capabilities for document understanding, demonstrating end-to-end **GenAI** development.