Ri Hong

riri.hong@gmail.com | rihong.ca | github.com/Ri-Hong | linkedin.com/in/ririhong

Skills

Languages: Python, Java, C/C++, JavaScript, TypeScript, Go, HTML, CSS
Technologies: React, Next.js, Docker, Kubernetes, Terraform, FastAPI, gRPC, AWS, Linux, PostgreSQL

Experience

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

- Engineered multi-region infrastructure with Terraform, Kubernetes, and Flux on GCP, enabling horizontally scalable systems and cutting deployment time by 24% for production workloads.
- Designed and launched an observability platform measuring engineering productivity and AI tool adoption, providing actionable insights for 400+ engineers across multiple product teams.
- Developed production-ready dashboards and analytics workflows that surfaced correlations between tool usage and code quality, driving data-informed adoption strategies.

Base Power Jan 2025 – Apr 2025
Markets Infrastructure Engineer *Austin, TX*

- Led fullstack and infrastructure initiatives in a mission-critical trading environment, collaborating with algorithm developers to improve system performance in a Series B ($850M) startup.
- Improved trading simulation reliability and scale by transitioning from local to cloud-native execution using Temporal Cloud, enabling 1000s of auto-retriable workflows and eliminating single-node failure risks.
- Discovered and resolved a performance bug in real-time market data transformation logic, reducing complexity from $O(n^2)$ to $O(n \log n)$ by applying a sort-and-search optimization.
- Used OpenTelemetry traces to uncover a performance bottleneck caused by blocking BigQuery writes; implemented non-blocking async publishing using Go routines, achieving a 32% speedup.
- Implemented Protobuf and gRPC to enable seamless communication between Python algorithm services and Go microservices, ensuring type safety and reducing serialization overhead by 20%.
- Rewrote market controller UI using React and Next.js, allowing on-call traders to execute trades within seconds.

Trend Micro Summer 2024
Software Developer Intern *Ottawa, ON*

- Upgraded the legacy Deep Security Manager from JDK 8 to JDK 11, modernizing the codebase and enhancing compatibility with contemporary tools for over 250 million global customers.
- Revamped the Jenkins CI/CD pipeline to support JDK 11, achieving a 35% increase in automation efficiency and accelerating deployment timelines by 15%.
- Refactored monolithic codebases into microservices, cutting deployment errors by 30% and improving scalability for future development.
- Troubleshot and resolved over 40 installation and deployment issues on Linux EC2 instances, enhancing system reliability and uptime.

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

- Launched an AI platform combining GPT-5 with Manim to auto-generate explainer videos; processed 1,000+ animations with a Celery + FastAPI pipeline.
- Accelerated generation by segmenting code generation and running distributed GPU rendering, reducing render latency from 12s to 2s; leveraging a pipelined streaming workflow to deliver near real-time playback.
- Scaled secure rendering and delivery with Docker, FFmpeg, Cloudflare R2, and PostgreSQL; raised $20k pre-seed and led product strategy from prototype to production.

Projects

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

- Engineered a production-grade sentiment analysis service achieving 92.5% accuracy using DistilBERT, deployed on GKE with Terraform.
- Implemented end-to-end MLOps pipeline with MLflow for experiment tracking, DVC for data versioning, 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.

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

- Engineered a high-performance Neural Style Transfer system with custom CUDA kernels for Gram matrix computation, achieving 8x faster style transfer compared to CPU-only implementation.
- Implemented an efficient VGG19-based feature extractor with L-BFGS optimization, delivering production-quality artistic style transfer in under 30 seconds on consumer GPUs.
- Built a scalable REST API with FastAPI for style transfer requests, containerized with Docker for easy deployment.

Trasee | *Pyodide, React, TypeScript, Vite, React Flow* Oct 2025

- Built a real-time Python code visualizer that intelligently recognizes data structures (linked lists, trees, graphs) and renders them interactively in the browser, helping users understand algorithms visually.
- Engineered a two-phase static + runtime analysis pipeline using Python's ast module and sys.settrace within Pyodide WebAssembly, enabling accurate inference without external APIs.
- Won "Best Revolutionizing Learning Hack" at Hack the Valley X among 100+ teams.

HomeLab | *Linux, Proxmox, Ansible, Kubernetes* Apr 2025

- Deployed a Kubernetes cluster on Proxmox with Terraform + Ansible, enabling scalable, self-healing microservices and secure HTTPS traffic with Nginx + Cloudflare routing.

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, AI, ML