Ri Hong

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

Skills

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

Experience

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

- Engineered multi-region infrastructure with **Terraform**, **Kubernetes**, Kustomize, and Flux on GCP and Groq hardware, enabling horizontally scalable AI/ML workloads and cutting deployment time by 24%.
- Designed and launched an observability platform to measure engineering productivity and AI coding tool adoption (Claude Code, Cursor, AmpCode), providing actionable insights for 400+ engineers.
- Developed dashboards and analytics workflows that surfaced correlations between AI tool usage and code quality, driving data-informed adoption strategies.

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.

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

- Led fullstack and infrastructure initiatives in a mission-critical environment, collaborating with algorithm developers to improve energy trading 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 OpenTelmetry traces to uncover a performance bottleneck caused by blocking BigQuery writes during simulations; implemented non-blocking async publishing using **Go** routines, achieving a 32% speedup.
- Implemented **Protobuf** and **gRPC** to allow for 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
May 2024 – Aug 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%.
- Overhauled critical development tools, including install4j, powermock, and JAX-WS, enabling 25% faster integration testing cycles.
- 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.

GeeseHacks
May 2024 – Feb 2025
Lead Software Developer *Waterloo, ON*

- Led a team of 8 developers to develop the software infrastructure for a first-time hackathon with 600+ attendees, achieving 98% uptime during the event and securing $25k in funding.

- Engineered and implemented a robust DevOps strategy to manage multiple databases, ensuring 100% schema synchronization across three environments and reducing deployment errors by 30%.
- Pioneered the use of **GitHub Actions** pipeline to reduce manual work, increasing developer productivity by 43%.
- Cultivated a quality-first culture by incorporating **Jest** for unit tests and **Cypress** for E2E testing, achieving 98% test coverage and reducing production bugs by 35%.
- Optimized application performance and search visibility using **Next.js**, achieving a 25% faster page load time and improving SEO rankings by 20%.

Microgreen Solar Corp. Sept 2023 – Dec 2023
Software Developer Intern *Toronto, ON*

- Spearheaded and meticulously documented a full-stack **AWS** Cloud project, from design to deployment, for processing EnergyPak lithium battery data.
- Utilized **AWS** services including API Gateway, **Lambda**, **DynamoDB**, IoT Core, and Amplify, enhancing project functionality and reliability.
- Implemented a Progressive Web App (PWA) frontend with **Next.js**, optimizing navigation with page routing and Server-Side Rendering (SSR).
- Dramatically reduced data retrieval latency and operational costs by 318% through effective use of browser indexedDB caching.

Projects

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.
- Designed a performant time-travel debugger, storing execution diffs for thousands of steps while maintaining smooth playback with debounced updates and memoized **React** renders.
- Integrated Monaco Editor for a VS Code-like UX and **React Flow** for custom graph visualizations with auto-layout algorithms and drag interactions.
- Won "Best Revolutionizing Learning Hack" at Hack the Valley X among 100+ teams.

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.

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.

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.

Skip the Walk | *Terraform, Mappedin, Commander.js, Slack API* Sept 2024

- Created a CLI tool to solve the last-mile-delivery problem, enabling volunteers to deliver pizza directly to hackers without leaving their workspace.
- Used Mappedin for room labeling and pathfinding, and **Terraform** to automate pizza orders.
- Awarded "Best Use of Terraform" at Hack the North 2024, Canada's largest hackathon.

PharmFill | *Google Cloud Vision, Gemini, Python Pillow* Jan 2024

- Led the development of a mobile app designed to streamline the process of converting and communicating prescriptions between physicians, pharmacists and patients.
- Integrated **Google Cloud Vision** OCR and **Gemini** Pro for accurate text extraction, classification, and form annotation from prescription receipts.
- Won "MedX AI Challenge" and placed Top 10 at DeltaHacks X among 95 teams.

TrailGuide | *Google PaLM, Infobip, FastAPI* Nov 2023

- Created an SMS-based hiking assistant delivering real-time weather updates, safety check-ins, and trip info to enhance user experience and safety.
- Implemented an advanced query classification system using Google's **PaLM** LLM, enabling accurate and responsive user interactions.
- Awarded "Best Use of Infobip" at Hack Western 10, distinguishing our project among 15 competitive teams.

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

- Engineered an AI study assistant using **GPT-4** for personalized learning tools like summaries and flashcards from user-uploaded notes.
- Led the RAG process architecture, integrating **GPT-4** with **MongoDB** Vector Search and **Flask** for the backend.
- Achieved 4th place at Hack the Valley 8 among 80 teams.

Blockmind AI | *GPT-4, Polygon, LangChain, Faiss* June 2023

- Developed an AI assistant to simplify blockchain tasks like minting NFTs, token transfers, smart contract audits, and transaction validation.
- Implemented the T3 stack for a full-stack web demo, using IPFS/Filecoin for decentralized NFT storage and The Graph for market queries.
- Won Worldcoin Best AI Use Case and Pool Prizes from UMA and The Graph at ETHGlobal Waterloo.

Bill in Bytes | *React Native, Google Cloud Vision* May 2023

- A practical application designed to streamline the management of paper receipts by converting them into digital records, enabling easy tracking of expenses and simplified tax filing.
- Integrated **Google Cloud Vision** API to incorporate optical character recognition (OCR) technology, transforming physical receipts into digital data.
- Submitted to GryphHacks 2023, won Best UI/UX and People's Choice award (1 out of 20 teams).

Ingredify | *React Native, Cohere, MongoDB* May 2023

- A mobile app that allows users to take a picture of an ingredients list, then provides extensive information about ingredients, including health effects and allergenic properties.
- **Cohere** API used to generate descriptions, classify components, retrieve the best explanation from our **MongoDB** database, and provide the summary of an ingredient.
- Submitted to MetHacks 2023, won Best Use of Cohere (1 out of 35 teams).

Personal Website | *ReactJS, CSS* Jan 2023

- Developed a frontend **React** website to showcase my work experience and projects.
- Used Midjourney AI image generation, Photoshop and pure CSS to achieve the parallax effect on landing page.
- Integrated Google Analytics to track number of site visits (100+ site visits).

Foodtastic | *HTML, CSS, Javascript* May 2022

- A website that uses the Flipp API and Spoonacular API to compile a list of recipes and supermarkets to find the cheapest ingredients so that university students can make healthy and affordable food.
- HTML and CSS used for the frontend and Javascript used to connect to the APIs.
- Submitted to Jamhacks 6, won Best Hack for University Students (1 out of 24 teams).

Smart Door Lock | *Arduino (C++), Android Studio (Java), IoT* Apr 2021

- An ESP 32 module with an LCD display, pin pad, and servo to lock/unlock a door.
- Comes with a companion Android app to remotely toggle, monitor the state of the lock and Android SQL Database used to store the log of users. The system allows for multiple users with unique pins to be recognized.
- Communicates over IoT MQTT.

Fluencity | *ReactJS* Aug 2021

- A web app used to train reading and speech fluency using the Symbl.ai text-to-speech API.

- Won Best Veteran Hack and Best use of Symbl.ai in XHacks 2021.

Computer Science Club Website | *ReactJS, Firebase* May 2021

- Developed the official website of Bethune C.I.'s computer science club to promote computer science and attract new members.
- Lead the process of planning, prototyping, frontend design, and backend integration in a team of four.
- Implemented **Google Firebase** to authenticate executives and used the Firestore database to store announcements, executives, and projects.

Publications

Pruned Graph Generating State Space Models (in preparation) Oct 2025

Research Paper

- Investigates the effect of pruning edges from the Minimum Spanning Tree (MST) in GG-SSMs to reduce FLOPs while preserving accuracy.
- Proposes a pruning strategy to sparsify graph structure, improving computational efficiency and scalability of GG-SSMs for high-dimensional data.

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, Artificial Intelligence, Machine Learning, Operating Systems, Distributed Systems, Computer Vision