Rajan Chavada

ML Software Engineer · AI Systems · Agentic AI

RajanChavada111@gmail.com | linkedin.com/in/rajan-chavada | github.com/RajanChavada | chavada.vercel.app

Education

University of Western Ontario — London, ON

B.Sc. Computer Science, GPA: 3.7/4.0 — *Expected May 2027*

- **Courses:** Artificial Intelligence, Machine Learning, Data Science, Probability & Statistics, Algorithms, Systems Design, Software Engineering, Networking
- **Leadership:** Executive @ Western AI: led 4-person team building CNN ASL translator for low-income communities · Executive @ Western Design Thinking: pitched GPS car-theft solution to Dean of Engineering

Experience

Borealis AI (RBC AI Research Lab) — Vancouver, BC (Remote)

Machine Learning Software Engineering Intern — *Dec 2025 – Apr 2026*

- Cut inference latency 25% and doubled throughput on client-capacity serving via KServe with serverless NVIDIA Triton
- Migrated 3 production model pipelines onto the AI Farm H100 cluster via GitHub Actions CI/CD, automating Airflow inferencing schedules with Grafana + Prometheus observability
- Deployed Celery, worker, and Flower pods on OpenShift behind a FastAPI gateway with mTLS ingress and certificate-based auth, maintaining 99.95% uptime on mission-critical endpoints

Intact Insurance | AWS — Toronto, ON

Site Reliability Engineering Intern — *Sept 2025 – Dec 2025*

- Cut infrastructure provisioning time by 65% building Terraform IaC libraries across 190+ AWS accounts
- Hit 95% compliance coverage automating cloud cost optimization with Amazon Bedrock ETL pipelines
- Cut security audit latency by 75% across 50+ Kubernetes clusters using DaemonSet vulnerability scanning

Royal Bank of Canada | Amplify – Hedge Funds Research — New York, NY (Remote)

Software Engineering Intern — *May 2025 – Aug 2025*

- **Patent-pending (2025):** agentic orchestration with RAG retrieval and gated validation for regulated financial systems
- In production across 18,000+ front-office traders at RBC Capital Markets, saving 6 hours of manual research per trader workflow
- Architected on FastAPI + LangGraph with Kafka ingesting FactSet, 13F filings, Bloomberg, and internal client data

Royal Bank of Canada | Capital Markets — New York, NY (Remote)

Software Engineering Intern — *May 2024 – Aug 2024*

- Built a RAG agent over 300,000+ financial documents on RBC's internal LLM gateway, surfacing LLM-driven analysis and summarized insights, cutting client support ticket triage by 22%
- Authored reusable React components and Node.js modules powering mission-critical client-facing dashboards (4,500 DAU); profiled with Lighthouse CI across 12+ build configurations to cut page load times

Royal Bank of Canada | Global Equities — Toronto, ON

Software Engineering Intern — *May 2023 – Aug 2023*

- Built a full-stack CVE triage platform (React + FastAPI) auto-ingesting feeds from MySQL and JIRA REST API on a cron schedule, cutting developer triage time by 40% and prod errors by 30%
- Adopted across 3 trading desks (45+ traders) and 100+ Capital Markets engineering teams; designed TTL caching and priority-based alert routing, achieving 99.8% SLA for on-call engineers

Projects

Rosetta – *Open-source agentic-coding CLI, 1,800+ users, adopted at RBC Borealis AI* | npm | GitHub

- One-command CLI (`rosetta init`) provisioning agentic-coding rules, MCP server hooks, and IDE prompts across 9 IDEs including Cursor, Windsurf, and Claude Code; shipped as a multi-module Node.js npm package
- Onboarded 1,800+ technical and non-technical developers to agentic-coding workflows; pitched to and adopted internally by RBC Borealis AI

Neurovn – *Open-source visual workflow editor for agentic AI* | Live App | PyPI | GitHub

- Drag-and-drop canvas to wire up multi-agent workflows; projects cost and P95 latency for any cyclic or DAG agent graph before a single API call, using Tarjan's SCC + topological sort with Rust-bound tiktoken for instant token accounting
- Open-sourced frontend + backend with PyPI distribution; live canvas IDE for developers, designers, and founders prototyping agentic systems

NVIDIA Alert Triage Agent – *Multi-agent LLM incident response for GPU training clusters*

- Multi-agent system ingesting live alerts from internal GPU training clusters; auto-classifies severity and generates root-cause analyses for on-call engineers
- Cut MTTR by 70–80% via LangGraph orchestration over NVIDIA NIM with Kafka alert streaming

Technical Skills

Languages: Python, TypeScript, JavaScript, SQL, C++, Java
Frameworks: FastAPI, LangChain, LangGraph, Next.js, React, Flask, Node.js
Cloud & DevOps: AWS, GCP, Docker, Kubernetes, Terraform, GitHub Actions, CI/CD
Data & ML: PyTorch, RAG, Vector Databases, Kafka, PostgreSQL, Supabase, NVIDIA Triton