Rahul Awale

Toronto, ON | 437-410-0945 | awaler181@gmail.com | Linkedin | Github

Professional Summary

Machine Learning & AI Engineer with hands-on experience building scalable AI systems, including RAG pipelines, semantic search, and production-grade APIs. Experienced in designing, deploying, and optimizing end-to-end ML solutions using FastAPI, Docker, and cloud infrastructure. Strong focus on real-world impact, system design, and performance optimization.

Educational Qualifications

Loyalist College, Belleville, ON, Canada
Post Graduate Certificate in AI and Data Science **2024-2026**

Islington College, Kathmandu, Nepal
BSC(Hons.) in Computing **2019-2022**

Technical Skills

Programming: Python, SQL, Dart

Machine Learning & AI: Supervised & Unsupervised Learning, Deep Learning, Feature Engineering, Model Evaluation, NLP, Retrieval-Augmented Generation (RAG), Predictive Modeling

Data & Analytics: Data Cleaning, Exploratory Data Analysis, Statistical Analysis, Data Visualization, Data Pipelines

Systems & Deployment: API Development, ML Model Deployment, Containerization, Cloud Deployment (AWS), Scalable System Design

Projects

CivicAI - Scalable GenAI Platform (RAG + Async Processing + Cloud Deployment)

- Designed and deployed a production-grade RAG system using FastAPI, Next.js, and pgvector for semantic search
- Built an asynchronous processing pipeline (Redis + Celery) to handle document ingestion and query workloads efficiently
- Deployed on AWS EC2 with Docker, Nginx, and Cloudflare, enabling scalable and reliable system access
- Optimized vector retrieval and query performance for large-scale document datasets. GitHub

Belleville By-Law Assistant - GenAI RAG System

- Reduced manual by-law lookup time by enabling citation-grounded Q&A over scanned PDFs using OCR + RAG.
- Integrated Llama3 (Ollama) and Zephyr-7B to optimize latency, accuracy, and offline inference.
- Built Gradio and Streamlit chat UIs with structured answers and page-level citations. GitHub

Car Price Prediction - Regression | Python, Scikit-learn, FastAPI, Docker

- Engineered brand/quality features and trained ensemble models (Random Forest, Gradient Boosting).
- Reduced RMSE compared to a linear baseline on holdout data.
- Deployed via FastAPI microservice packaged in Docker for portable use. GitHub

Experience

Software Developer | Aalaya Soft-tech Pvt. Ltd., Nepal | Jul 2022 - Jan 2024

- Developed and maintained Flutter applications integrated with REST APIs and analytics modules.
- Collaborated with backend teams to design APIs and implement data-driven features.
- Mentored interns, reducing bug backlog by 20% & increasing delivery by 20%.
- Built production systems with a strong focus on performance, maintainability, and user experience.