

ROSHAN ARYAL

AI / ML Engineer | Full-Stack Developer

Auckland, New Zealand • +64 22 388 9135 • roshanaryaal@gmail.com • linkedin.com/in/roshanaryaal • github.com/roshanaryal1 • roshanaryal.com

PROFESSIONAL SUMMARY

AI Engineer at Apprento.io building HireIQ, a Claude-powered candidate-intelligence platform now live in recruiter workflows for a NZ/AU specialist sales-recruitment firm partnered with 50+ tech companies. End-to-end builder of production AI systems: RAG pipelines, agentic tool calling, in-browser vector embeddings, LLM evaluation harnesses, and the full-stack UI around them. Final-year Bachelor of Information Technology (AI and Data Science) at Otago Polytechnic Auckland, graduating December 2026. Open to graduate and mid-level AI / ML / full-stack roles across New Zealand and remote.

CORE SKILLS

AI and Machine Learning: Anthropic Claude API, Groq Cloud (Llama 3.1), Retrieval-Augmented Generation (RAG), Vector Embeddings (ONNX, all-MiniLM-L6-v2), Agentic Tool Calling, OpenAI Function Calling, LLM Evaluation, Prompt Engineering, Hugging Face, LlamaIndex, scikit-learn, Random Forest, KNN, GridSearchCV, SHAP, pandas, NumPy, TensorFlow

Languages: Python, JavaScript (ES2022), TypeScript, Java, SQL

Frontend: React 18, Next.js, React Router v6, Tailwind CSS, Bootstrap 5, Recharts, Streamlit, HTML5, CSS3

Backend: FastAPI, Node.js, Spring Boot, REST APIs, JWT, OpenAPI 3, Webhooks

Cloud and DevOps: AWS, Firebase (Firestore, Auth, Hosting, Cloud Functions), Vercel, Docker, GitHub Actions, CI/CD, OWASP Top 10

Databases: PostgreSQL, Firestore, IndexedDB

Tools and Practices: Git, Playwright, Jest, ESLint, Stripe API, Agile, Code Review, WCAG 2.1 AA, Content Security Policy

PROFESSIONAL EXPERIENCE

AI Engineer | Apprento.io

April 2026 – Present

Python • FastAPI • Anthropic Claude API • React • Docker • AWS • Auckland, NZ

- Designed and shipped end-to-end HireIQ, an AI report-generation pipeline that produces a client-ready candidate assessment (executive summary, scorecard, fit analysis, targeted probe questions) in under 30 seconds from a CV plus interview transcript, replacing a 30 to 45 minute manual baseline.
- Architected document ingestion (pdfplumber, python-docx) feeding a structured Claude prompting layer that emits a typed report shape, with role-aware scoring rubrics so the same engine handles different candidate seniority levels appropriately.
- Built a stateless processing layer to meet New Zealand Privacy Act 2020 obligations on candidate data, minimising what is stored and reducing compliance surface area.
- Implemented a fallback API path so an upstream LLM outage cannot block recruiters mid-workflow, plus a human-in-the-loop review gate before any candidate outcome is actioned.
- Delivered the React frontend handling upload, role selection, and inline report rendering; containerised with Docker and deployed on AWS with a Vercel frontend.

FEATURED PROJECTS

RentMate | AI-Powered Equipment Rental Marketplace

Live

React 18 • Firebase • Groq Cloud (Llama 3.1) • ONNX Embeddings • Stripe • GitHub Actions • rentmate.roshanaryal.com

- Solo-built, production-deployed three-role marketplace (renter, owner, admin) with 9 distinct production AI features routed through a single server-side proxy; 441 commits to v1.11.0.
- Shipped a production RAG chatbot with a 199-chunk knowledge base, dual retrieval (semantic embeddings plus TF-IDF keyword scoring), and in-browser ONNX vector embeddings via @xenova/transformers (all-MiniLM-L6-v2), eliminating any embedding-server cost.
- Engineered an agentic tool-calling system: 5 read-only Firestore tools wired to the LLM via OpenAI function-calling format, with a streaming agentic loop that resolves tool calls mid-stream before producing the final reply.
- Hardened against prompt injection and hallucination with a layered defence: tagged-envelope user input, client-side short-circuit on injection patterns, server-side tool allowlist per auth state, and an output sanitiser as a final filter.

- Built a 50-query RAG evaluation harness with retrieval-rate, keyword-recall, and p50 / p95 latency metrics, wired as a CI gate that fails the pipeline if pass rate drops below 90%. Currently 96% on the live dataset.
- Implemented server-side LLM cost controls: per-user 50 req/day quota plus platform-wide 5,000 req/day kill-switch backed by Firestore transactions, exponential-backoff retries on 429s, and a 25% prompt-token reduction.
- Built a Stripe-style developer portal with API key management, request logs, p95 latency, error-rate metrics, webhook endpoints, and an interactive API playground; published an OpenAPI 3 spec.
- Set up production infrastructure on Firebase Hosting with a custom domain, Google Trust Services SSL, CSP / HSTS / X-Frame-Options headers, JSON-LD structured data, and a SaaS-grade GitHub Actions pipeline (lint, build, RAG regression, unit tests, Cloud Functions syntax check, per-PR Hosting preview channels).

Road Accident Risk Predictor | Regression on Crash Data

Live

Python · scikit-learn · Random Forest · GridSearchCV · SHAP · Streamlit roadaccident-roshanar-aryal.streamlit.app

- Trained and tuned three regression baselines (Linear Regression, Random Forest, Gradient Boosting) on real road-crash data, with stratified splits to preserve the rare-event distribution.
- Selected Random Forest at 90.5% R² on the held-out test set after GridSearchCV hyperparameter tuning, choosing it over a comparable Gradient Boosting result for retraining stability.
- Added SHAP feature-contribution charts inline in the UI so every prediction is auditable in a safety-sensitive context, with a Streamlit risk-gauge interface a non-technical user can drive directly.

Quiz Tournament Platform | Multi-Role Real-Time System

Live

React · Spring Boot · PostgreSQL · JWT · Vercel quiz-app-froentend.vercel.app

- Designed a three-role platform (admin, host, participant) with JWT role-claim RBAC, keeping the Spring Boot backend stateless and removing the need for a session/Redis dependency.
- Engineered tournament-state synchronisation across concurrent participants using a server-side round-version pattern with a 1-second poll, delivering near-real-time updates without the operational overhead of a websocket layer.
- Modeled tournaments, rounds, questions, and submissions in PostgreSQL where consistency guarantees mattered more than horizontal scale.

Diabetes Prediction Model | Health-Outcomes Classification

2025

Python · scikit-learn · pandas · KNN Imputation · GridSearchCV

- Built a binary classifier on the Pima Indians Diabetes dataset using KNN imputation for missing values and a stratified split to preserve class balance.
- Threshold-tuned for recall over raw accuracy to fit a screening-tool use-case, achieving ROC AUC of 0.81.

EDUCATION

Bachelor of Information Technology | Otago Polytechnic, Auckland International Campus

2024 – Dec 2026

(Expected)

Specialisation: Artificial Intelligence and Data Science. Coursework: Machine Learning, Software Development, Cloud Computing, Automation and Robotics.

CERTIFICATIONS

Anthropic: Claude Code in Action (Oct 2025) · AI Fluency: Frameworks and Foundations (Sep 2025)

Generative AI and LLMs: Hands-On AI: RAG using LlamaIndex (Apr 2026) · Vector Databases in Practice: Deep Dive (Apr 2026)

Cloud: AWS Cloud Foundations, Amazon Web Services (Dec 2025)

Machine Learning: Applied Machine Learning: Feature Engineering (Nov 2025) · Machine Learning with Scikit-Learn (Oct 2025) · Machine Learning with TensorFlow on Google Cloud (Jul 2025)

Python and Data: NumPy, SciPy, Matplotlib and Pandas A to Z, Udemy (Sep 2025) · Python Zero-to-Hero Intermediate, Udemy (Aug 2025)