

# RAGHU SHARMA

## SENIOR MACHINE LEARNING ENGINEER | APPLIED AI, GENAI AND ML PLATFORMS

Jammu, India | +91-7889806911 | raghusharma0404@gmail.com

linkedin.com/in/raghu-sharma-7091b4115 | github.com/raghurs49 | rs-portfolio-teal.vercel.app | upwork.com/freelancers/raghurs

### PROFILE

Senior Machine Learning Engineer with 7+ years building ML systems from experimentation through production across forecasting, computer vision, NLP, GenAI, retail, and enterprise environments. Current focus: production-grade RAG, agent orchestration, secure tool and data access, evaluation, reliability engineering, and cloud-native delivery across GCP and AWS.

### SELECTED ENGINEERING PROJECTS

#### RAG Quality Workbench

FastAPI, Next.js, pgvector, multi-LLM

- Built document ingestion, hybrid retrieval, reranking, citations, abstention, and a comparative evaluation harness using recall@k, MRR, grounded-answer checks, latency, token, and cost traces.
- Added 77 automated tests and real startup/request-path coverage after identifying that dependency overrides could let isolated tests pass while production wiring failed.

#### Reliable LangGraph Support Agent

LangGraph, FastAPI, persistent state

- Designed router/supervisor workflows with durable state, structured outputs, human-approval gates, escalation, retries, timeouts, and circuit breaking around external tool calls.
- Validated routing and failure behavior with 66 automated tests and an 11-scenario evaluation suite through the compiled graph.

#### Secure BigQuery MCP Gateway

MCP, Cloud Run, BigQuery, PostgreSQL

- Separated caller authentication from cloud workload identity and enforced OIDC, RBAC, read-only query policy, PII masking, rate limits, query budgets, and structured audit logging.
- Extended the reference gateway across BigQuery, PostgreSQL, and governed REST connectors, verified by 34 tests plus 7 connector subtests.

#### AWS Bedrock RAG API

Bedrock, S3, IAM, FastAPI, Docker

- Built a deployable reference API with S3-backed ingestion, Bedrock Knowledge Bases, citations, least-privilege IAM, guardrail support, structured logging, CI, tests, and container delivery.

### CORE SKILLS

**ML & data:** TensorFlow, PyTorch, scikit-learn, XGBoost, NLP, computer vision, forecasting, feature engineering, PySpark, SQL, BigQuery

**GenAI systems:** RAG, hybrid retrieval, reranking, evaluation, grounding, abstention, LangChain, LangGraph, MCP, tool calling, structured outputs, vector databases

**Production engineering:** Python, FastAPI, Docker, Kubernetes, CI/CD, testing, observability, retries, fallbacks, model serving

**Cloud:** GCP (Vertex AI, Cloud Run, GKE, Dataproc, Dataflow, BigQuery, Cloud Functions); AWS (Bedrock, S3, IAM, containerized inference)

## PROFESSIONAL EXPERIENCE

---

### Independent AI/ML Engineer (Upwork) | 2026 - Present

Independent consulting and engineering | Remote

- Design and build production-oriented GenAI, RAG, agent-orchestration, ML, data, and cloud systems; current public work includes four tested reference implementations listed above.
- Own technical discovery, architecture, milestone planning, delivery communication, deployment guidance, and handover documentation for client-facing work.

### Machine Learning Engineer (ML/MLOps) | Nov 2023 - Feb 2026

Open Insights | United States, Remote

- Core contributor to Open ML 1.5, the forecasting engine behind OpenINSIGHTS' AI customer-opportunity platform for enterprise retail brands, supporting lifecycle, acquisition, churn, and win-back prediction.
- Built and optimized PySpark feature pipelines and TensorFlow training/inference workflows across hierarchical product and customer datasets; worked across BigQuery, Dataproc, and GCP infrastructure.
- Collaborated directly with the Chief Scientist on production ML delivery and was recognized for dependable ownership of production work.

### MLOps Engineer | Oct 2022 - Oct 2023

Capgemini Technology Services India | Bengaluru, India

- Deployed ML workloads on GCP and modernized execution paths toward serverless, maintainable architectures.
- Implemented carbon-emission measurement within ML workflows to make computational footprint visible during model execution.

### Data Scientist / ML Engineer | Jul 2019 - Oct 2022

Hughes Systique | Gurugram, India

- Productionized computer-vision, NLP, and tabular ML systems across GCP and AWS, covering object detection, classification, segmentation, predictive risk, APIs, and cloud deployment.
- Built custom TensorFlow libraries for Android/iOS edge deployment and maintained internal CI/CD for GCP-based ML delivery.

### Machine Learning Intern | Jan 2019 - Apr 2019

Integration Wizards | Bengaluru, India

- Trained TensorFlow/PyTorch models, prototyped human pose estimation with AlphaPose, and supported a computer-vision system for detecting chemical-fluid leaks in factory environments.

## CREDENTIALS

---

**Google Cloud Professional credentials earned:** Cloud Architect and Machine Learning Engineer

**Additional:** NVIDIA Jetson AI Certification; HackerRank Python; Coursera specializations in TensorFlow, NLP, and deep learning

**Competitive ML:** Kaggle - ranked in the top 11% (800 of 7,000 competitors)

## EDUCATION

---

**B.Tech, Computer Science and Engineering** | Shoolini University, Solan, India | 2015 - 2019 | OGPA: 8.14