Gur Raunaq Singh

Senior Machine Learning Engineer / AI Architect

Bangalore, India | Born 09/01/1996 | +91 90151 54091
raunaq.soni@gmail.com | LinkedIn | Github

PROFESSIONAL SUMMARY

Senior AI & ML Systems Architect with 8+ years of experience designing secure, scalable AI platforms in enterprise environments processing high-volume transactional data. Specialized in distributed ML systems, Kubernetes-native orchestration, multi-cloud architectures, and Generative AI (LLMs, RAG, agentic workflows). Strong focus on MLOps, IAM-based access control, auditability, and operational resilience.

TECHNICAL SKILLS

AI/ML: LLMs, RAG, LangChain, LangGraph, Model Monitoring
Cloud & Infra: GCP (Vertex AI, BigQuery), AWS, Azure, Kubernetes, Docker, Terraform
Data & Backend: Python, SQL, Kafka, Redis, PostgreSQL, FastAPI, Django
MLOps: CI/CD, MLflow, IAM, Audit Logging, Encryption, SLA Monitoring

WORK EXPERIENCE

01/2023 – today

Machine Learning Engineer
Aptos India, Bangalore, India
Project: Aptos One - Enterprise Unified Commerce Platform – serving 100+ global retailers, processing billions of transactions annually.

- Architected a Kubernetes-native ML orchestration control plane enabling workload-aware scheduling, checkpoint-based recovery, and distributed job lifecycle management in SLA-bound production environments.
- Designed secure, fault-tolerant data pipelines on GCP (BigQuery + event-driven workflows), implementing IAM-based access controls, encryption practices, and audit logging while improving processing SLAs by 35%.
- Optimized cloud infrastructure and query execution strategies, reducing compute costs by 60% (~USD 240K annual savings) while modernizing legacy systems into scalable, observable cloud-native components.

Technologies used: Python | GCP | Vertex AI | BigQuery | Kubernetes | Docker | Terraform | Redis | CI/CD | IAM

Project: Generative AI & LLM Support Assistant

- Designed an enterprise-grade RAG architecture integrating structured analytics metadata and internal knowledge bases to deliver context-aware responses with improved semantic accuracy.
- Implemented prompt versioning, evaluation workflows, and monitoring mechanisms to reduce hallucinations by 35% while improving traceability and observability of model outputs.
- Deployed containerized inference workflows using MLOps practices, enabling controlled experimentation, scalable inference, and secure knowledge isolation via vector database access controls.

Technologies used: Python | LangChain | LangGraph | Vertex AI | VectorDB | FastAPI | Docker | MLOps

04/2022 – 12/2022

Senior Software Engineer

Turing.com , India (remote)
Project: End-to-End ML Deployment Platform – US-based remote talent platform enabling AI-driven analytics and automation at scale.

- Designed and implemented production-grade ML pipelines using Airflow for ingestion, training, validation, and automated deployment.
- Standardized CI/CD and model versioning practices, reducing deployment time by 60% and improving controlled production rollouts.
- Integrated monitoring and validation checks to enhance production reliability and maintainability of deployed ML systems.

Technologies used: Python | Django | Airflow | GCP | Docker | MLOps

04/2021 – 04/2022 **Senior Software Engineer**
Clickpost Pvt. Ltd., Bangalore, India
Project: Cash Reconciliation Automation Platform – serving 1,000+ enterprise customers handling high-volume CoD transactions.

- Designed and built a high-accuracy financial reconciliation system processing large-scale transactional datasets with strict data integrity and traceability requirements.
- Reduced reconciliation cycle time by 95% (from ~2 days to minutes) through parallelized processing and automated validation workflows, improving operational efficiency and financial visibility.
- Architected secure ingestion pipelines and scalable backend APIs under high-availability constraints, ensuring auditability and controlled access to financial data.

Technologies used: Python | Django | PostgreSQL | AWS | Redshift | Kafka | Docker | Kubernetes

07/2018 – 04/2021 **Software Engineer**
Veda Labs Pvt. Ltd., Delhi, India
Project: Video Analytics Platform – Computer vision platform deployed on-premise for retail analytics using CCTV footage.

- Developed containerized video analytics pipelines for edge devices
- Built a Kafka-based streaming architecture for scalable, real-time processing

Technologies used: Python | OpenCV | Computer Vision | Kafka | Docker

EDUCATION

08/2014 – 06/2018 Bachelor of Technology – Computer Science Engineering
Indraprastha University, New Delhi, India

ACHIEVEMENTS / CERTIFICATIONS

Co-author: **Amazon Sumerian by Tutorials** (ISBN: 1942878877) | Guest Author: **RealPython.com**, **Kodeco.com** | Published multiple technical tutorials on Python, Alexa Skills, and Unity 3D

LANGUAGE SKILLS

English (C2) | Hindi (Mother tongue)

HOBBIES

Travel | Photography | Strength Training | Badminton | Technical Writing