Raghav Kochhar

raghavkochhar007@gmail.com — linkedin.com/in/raghavkochhar007 — github.com/Raghav-Kochhar
+1 (778) 652-9739 — Langley, BC, Canada — Canadian Permanent Resident

Summary

ML Engineer building event-driven systems, ML pipelines, and LLM-powered applications on GCP. Skilled in Pub/Sub, Dataflow, Cloud Run, and real-time inference, with experience in RAG architectures and Transformers.

Education

Vivekananda Institute of Professional Studies – Technical Campus — 2021 – 2025

Bachelor of Technology in Artificial Intelligence and Machine Learning (CGPA: 9.03/10).

Experience

Software Engineering and MLOps Intern, Jaipur Robotics — August 2025 – Present

- Engineered stateful Apache Beam/Dataflow pipelines to aggregate sensor streams, implementing complex deduplication logic across adjacent gates to resolve multi-camera overlap.
- Optimized streaming job costs by right-sizing Dataflow workers, cutting compute spend by 50%.
- Built a Cloud Run service for ONNX door-state inference (model trained from scratch, 98% accuracy), handling video processing with FFmpeg and session storage in AlloyDB.
- Architected a video-generation platform (FastAPI + Streamlit) that renders timelapse videos from GCS using Cloud Tasks for async processing.

Data Science Intern (Stealth Startup) — Summer 2024

- Led three production ML systems for traffic analysis, sign language, and emotion classification with 90%+ accuracy.
- Deployed containerized solutions on AWS, reducing inference latency by 25%.

Projects

American Sign Language Interpreter

- Built a real-time sign language recognition system with EfficientNet and LSTM, achieving 95% accuracy.
- Deployed the model via a FastAPI backend and Streamlit interface with live webcam inference.

Speech Emotion Classifier

- Developed a seven-class emotion recognition LSTM model trained on 1,200 audio samples.
- Built a FastAPI inference API and Streamlit UI for real-time and batch audio analysis.

Traffic & Pedestrian Analyzer

- Created a YOLOv8-powered traffic monitoring system achieving 90% mAP with pedestrian and signal detection.
- Integrated traffic-light color recognition for smart-city planning insights.

Automaton: No-Code ML Workflow Platform

- Implemented a No-Code ML workflow platform for data preprocessing, training, and tuning.
- Built interactive dashboards for model benchmarking and performance visualization.

Research

Publication: Efficient Adaptation of Lightweight LLMs (ICAAI 2025, Springer)

- Benchmarked Baseline, T-Free, and STELLA on consumer-grade hardware, validating edge feasibility.
- Achieved 75% embedding reduction with T-Free while maintaining accuracy.

Publication: Health Helix: Connecting You to Better Health (COM-IT-CON 2024)

- Presented at the International Conference on Progressive Computational Intelligence (Taylor & Francis Group); architected a unified healthcare platform.
- Designed engagement workflows that reduced clinic no-show rates via automated AI alerts and scheduling integration.

Technical Skills

- Languages: Python (uv, ruff, mypy, pytest), SQL, Bash.
- ML and Deep Learning: PyTorch, TensorFlow, Hugging Face Transformers, ONNX, Scikit-learn.
- LLMs and GenAI: LangChain, RAG pipelines, Prompt Engineering.
- Computer Vision and NLP: OpenCV, YOLOv8, FFmpeg.
- Web and APIs: FastAPI, Streamlit, SQLAlchemy, Pydantic.
- Data Engineering: Apache Beam, Pandas, NumPy.
- Cloud and DevOps: GCP, AWS, Azure, Docker, Kubernetes, GitHub Actions.