

# Ajmal Rasi

Senior AI Engineer · Computer Vision & Edge AI · GenAI & LLMs (RAG) · Distributed Systems

Bengaluru, India · +91 96563 68936 · kpajmalrasi@gmail.com · github.com/ajmalrasi · ajmalrasi.com

## PROFESSIONAL SUMMARY

---

Senior Applied AI Engineer with 7+ years of expertise designing and scaling production GenAI/LLM ecosystems. Specialized in building end-to-end RAG pipelines, transforming unstructured petabyte-scale seismic data into high-value insights. Proven track record in AI platform engineering, including retrieval optimization, structured extraction, and robust evaluation frameworks (RAGAS). Complemented by a deep foundation in C++ and computer vision, including segmentation, object detection, and ViT/MAE-based foundation models, alongside expertise in edge AI deployment (TensorRT, model quantization), distributed systems, and cloud platforms across GCP and AWS. Committed to data-driven performance engineering, benchmarking, isolating variables, and validating deployments at scale. Open-source contributor to Label Studio (HumanSignal).

## CORE STRENGTHS

---

- **GenAI & LLM Systems** Retrieval-Augmented Generation (RAG) architecture, multimodal RAG, vector search (IVF) & embedding optimization, tool orchestration, structured extraction, prompt engineering, transformer architectures (BERT, GPT, LLaMA), and image generation (Stable Diffusion SDXL, ControlNet).
- **AI Observability & Evaluation** LLM evaluation frameworks (RAGAS: faithfulness, retrieval precision), hallucination detection, operational monitoring (OTLP/Grafana), drift analysis, and production AI performance benchmarking.
- **MLOps** Docker, MLflow, Kubeflow, TFX, Apache Airflow, Kubernetes; comprehensive MLOps for GenAI/RAG, including operational monitoring (OTLP/Grafana) and quality-focused observability (RAGAS for hallucination, drift, and retrieval metrics).
- **Computer Vision & Video Analytics** Vision-language models (VLMs), object detection, multi-object tracking, segmentation, face recognition, pose estimation, number-plate recognition (ANPR); production video intelligence (footfall/people counting, dwell time, demographic profiling, vehicle analytics) and pedestrian tracking; camera calibration, fisheye/surround-view, and photometric alignment for automotive and CCTV camera systems.
- **Edge AI & Inference Optimization** NVIDIA Jetson (Nano to Orin), DeepStream SDK, GStreamer, TensorRT, ONNX, CUDA, Transfer Learning Toolkit (TAO); INT8/FP16 quantization, QAT (Quantization Aware Training) and PTQ (Post-Training Quantization), pruning, layer fusion; NXP/Samsung automotive SOC deployment; latency, throughput, FPS, and memory optimization via systematic benchmarking and profiling for real-time on-device inference.
- **Data Platforms & Distributed Systems** Petabyte-scale ingestion and processing, Dask distributed computing, NumPy/GeoPandas pipelines, workflow orchestration, storage architecture, data quality/integrity at scale.
- **Performance Engineering** Benchmarking, profiling, and root-cause analysis across model inference (latency, FPS, memory footprint), storage (NVMe vs network-attached), distributed throughput, and cloud workloads.
- **Cloud & Infrastructure** GCP (GKE, Cloud Run, Pub/Sub), AWS, Kubernetes, Apache Airflow, CI/CD (Azure DevOps), observability, automation, reliability and scalability.
- **Programming** Python (production-grade), C++, JavaScript, Shell, CMake/Make; SOLID, design patterns, algorithm design, Linux/OS internals.

## EXPERIENCE

---

### Senior AI Engineer, Data Platforms

TGS (Steerwise) | Remote, India | Jul 2022 – Present

*Via Affine Analytics (Senior Software Engineer), Jul 2022 – Jun 2024; direct with Steerwise (Senior AI Engineer), Jul 2024 – Present*

- Architected a Multi-modal Seismic AI Assistant integrating computer vision (ViT/MAE) and LLMs, enabling semantic search, fault segmentation, and intelligent metadata extraction on petabyte-scale datasets.
- Designed and deployed production-grade RAG systems for querying petabyte-scale seismic metadata. Engineered retrieval pipelines using embeddings and vector search (IVF), optimizing context retrieval for high accuracy. Orchestrated complex multi-step retrieval workflows, enabling robust natural-language query resolution over unstructured EBCDIC headers.
- Developed a comprehensive AI Observability & Evaluation framework. Integrated RAGAS for systematic LLM evaluation (faithfulness, retrieval precision, and hallucination monitoring) and established operational monitoring (OTLP/Grafana) to track latency, throughput, and cost-efficiency in production.

- Implemented LLM-based structured extraction pipelines to convert high-entropy, free-text vendor metadata into clean, schema-conformant data. Optimized parsing workflows to ensure high-fidelity context engineering, improving downstream searchability and data integrity.
- Built and scaled a petabyte-scale SEG-Y ingestion and distributed-processing platform, owning file-format parsing, metadata, coordinate systems, quality control, storage architecture, and Dask parallelization with NumPy and GeoPandas.
- Designed and operated cloud-native data platforms across GCP and AWS using Kubernetes, serverless execution, Apache Airflow, event-driven messaging, and Dask; improved reliability through systematic performance profiling and CI/CD in Azure DevOps.

## Software Engineer, Machine Learning

**Visteon Corporation** | Bengaluru, India | May 2021 – Jun 2022 | ADAS

- Led development of a Driver Monitoring System (DMS) for automotive ADAS; fine-tuned RetinaFace to state-of-the-art face-detection accuracy using Caffe, TensorFlow, and PyTorch, enabling real-time driver-state inference in-cabin.
- Optimized and ported CV/DL models to NXP and Samsung automotive SOCs via layer fusion, pruning, and INT8/FP16 quantization, cutting inference latency and enabling real-time, low-latency embedded deployment for automotive AI.
- Engineered in-cabin infrared image-quality and image-processing algorithms that improved downstream face-detection model accuracy under challenging lighting conditions.
- Built a 360° multi-camera surround-view system (fisheye distortion correction, perspective projection, camera calibration) in C++ with OpenCV, Eigen, and OpenGL for real-time automotive visualization.
- Developed a photometric alignment solution for multi-camera surround view, correcting brightness and color inconsistencies across camera feeds; grounded in current academic research and exceeded client expectations.

## Machine Learning Engineer, Computer Vision

**Beyond Drops** | Bengaluru, India | Dec 2020 – Mar 2021

- Built an end-to-end fingerprint biometric system (POC) using deep learning, object detection, and image processing; deployed a Dockerized inference server on GCP behind a REST API for Android integration.
- Implemented real-time, browser-based background subtraction for eKYC video calls (segmentation, OpenPose, MediaPipe, TensorFlow.js); cut in-browser load time via model sharding for execution in Chrome V8.

## Machine Learning Engineer, Computer Vision

**Integration Wizards** | Bengaluru, India | May 2019 – Dec 2019

- Delivered production CCTV video intelligence for enterprise clients: footfall (people counting), dwell-time measurement, demographic profiling, and vehicle analytics on NVIDIA Jetson Nano using DeepStream SDK, Transfer Learning Toolkit (now TAO), and GStreamer, the same edge stack behind modern CCTV and crowd-analytics platforms.
- Optimized detection and classification models via ONNX and TensorRT with FP16/INT8 quantization, QAT (Quantization Aware Training) and PTQ (Post-Training Quantization) on Tensor Core/CUDA devices, improving inference speed (FPS) and cutting memory footprint for real-time multi-stream video processing.
- Designed and built a custom multi-object pedestrian tracker in Python and C++ that outperformed OpenCV trackers and SORT on resource-constrained ARM CPUs, enabling real-time multi-object tracking at the edge through targeted performance optimization.
- Architected edge AI serving with Gunicorn, Celery, and Kafka for object detection, multi-label classification, pose estimation, and automatic number-plate recognition; rapid POCs directly secured new business.

## Software Engineer, Machine Learning (Internship)

**ApplicateAI** | Bengaluru, India | Mar 2019 – May 2019

- Built conversational-AI assistants with Rasa and spaCy NLP (custom tokenization, intent and entity extraction), delivering production natural-language understanding ahead of the LLM era.
- Developed demand-forecasting models using deep regression networks.

## TECHNICAL KNOWLEDGE

---

- **GenAI & LLM Engineering** RAG architecture (chunking, vector search/IVF/FAISS), LLM APIs (Claude, GPT), Prompt Engineering, Structured Extraction, Evaluation (RAGAS, hallucination monitoring), Image Generation (Stable Diffusion SDXL, ControlNet).
- **Problem Solving** Independently solve complex problems across diverse fields; adapt quickly to new environments using logical reasoning with strong attention to detail.

- **Theoretical Foundation** Solid grasp of the mathematics underpinning CV and ML (linear algebra, calculus, probability) and software-engineering fundamentals (OS, Linux, design patterns, SOLID).
- **Algorithm Design** Efficient algorithm design with low execution time, greedy, divide-and-conquer, and dynamic programming approaches.
- **MLOps** Build and maintain robust MLOps pipelines with GCP, AWS, Azure DevOps, Kubeflow, MLflow, Airflow, Kubernetes, Docker, and TFX; model monitoring and automated retraining in production.
- **Deep Learning, Internals** Activation functions (Swish, Sigmoid, ReLU, Leaky ReLU, Softmax, TanH) and their derivatives; regularization (L1/L2, dropout, early stopping, data augmentation, batch norm).
- **Optimization** Gradient-descent variants (Momentum, AdaGrad, RMSprop, Adam), learning-rate scheduling, hyperparameter tuning (grid/random/Bayesian), and loss functions (MSE, Cross-Entropy, Focal, Dice).
- **Model Evaluation** Evaluation metrics (mAP, Precision, Recall, F1), train/validation/test data distribution, and comprehensive error analysis.
- **Computer Vision** Edge detection (Canny, Sobel), Hough transforms, interpolation, feature extraction (SIFT, SURF, ORB), camera calibration, RANSAC, essential/fundamental matrix estimation, object localization, sliding window, anchor boxes, region proposals.
- **CV Applications** Face recognition (Siamese networks, Triplet loss), image captioning (CNN-RNN), object detection (YOLO), multi-object tracking (SORT, custom trackers), segmentation, ControlNet, Stable Diffusion (SDXL).
- **NLP & Transformers** LSTM, word embeddings (Word2Vec, GloVe), transformer models (BERT, GPT, LLaMA); pre-training and fine-tuning, attention mechanisms; LangChain/LangGraph, RAG, embeddings & vector search, structured extraction, RASA, spaCy.

## EDUCATION

---

- B.Tech, Computer Science, Jain University (SET), Bengaluru, 2018
- Higher Secondary (XII), Science, Kerala State Board
- Secondary Education (X), CBSE, Delhi