RAMSI KALIA MBA, MSc

New Delhi, India • LinkedIn • Email • GitHub • Website
AI Engineer | Agentic AI • Multi-Agent Systems • Cloud Deployment • RAG
Open to relocation and remote opportunities

SUMMARY

AI Engineer specializing in agentic workflows, MCP tooling, and retrieval-augmented systems. Experienced in building multi-agent applications that combine LLMs, structured tools, memory, and real-time orchestration.

AWS Certified Cloud Practitioner and AI Practitioner, with 2 years of portfolio work across deep learning,; 1 year of CV/ML engineering experience; and 5+ years in industrial engineering roles. Strong documentation and cross-functional collaboration skills, translating research concepts into reliable, production-ready workflows.

Agentic AI design and orchestration	**RAG systems and embedding workflows**
MCP server tooling	**LLM reasoning and structured prompts**
LLMOps evaluation and reproducibility	**AI life cycle management**

Tools & Frameworks	Python, FastAPI, Docker, LangChain, LlamaIndex, CrewAI, Google ADK, OpenAI SDK, Langfuse, Prometheus, Grafana, AWS (S3, EC2, Lambda), DVC, MLflow
Certifications	Professional Scrum Master™ I (PSM I), AWS Certified Cloud Practitioner, AWS Certified AI Practitioner
Languages	English, Hindi, Korean (B2)

SELECTED PROJECTS (2024–PRESENT)

Yoga Assistant Knowledge RAG | Retrieval-Augmented System — Nov 2025

Developed an end-to-end RAG system for yoga pose, sequencing, and pranayama knowledge

Technologies: Python, BM25, sentence-transformers, hybrid search, Docker, Grafana, Hugging Face

- Evaluated BM25, vector, and hybrid retrieval, achieving 23% MRR improvement with a weighted hybrid model.
- Benchmarked multiple LLMs and prompt templates using LLM-as-a-judge scoring, selecting DeepSeek-V3 for 90% relevance.

Co-Playable Characters (CPC) | Korean Learning MUD Game — Aug 2025

Prototype of persona-driven CPCs (AhjummaGPT, AhjussiGPT, etc.) in a retro MUD-style Korean learning game.

Technologies: FastAPI, CrewAI multi-agent orchestration, OpenAI, retro HTML/CSS, Hugging Face

- Designed a multi-agent system with CrewAI to maintain distinct cultural personas and enforce safe role boundaries.
- Developed evaluation and memory handling for persona coherence, latency, and safety across extended conversations.

MCP Albumentations | Spec-Compliant Image Augmentation Tool — July – Sept 2025

MCP-compliant image augmentation server on PyPI, enabling teams to apply complex Albumentations pipelines from plain English prompts.

Technologies: Python, Albumentations, MCP (JSON-RPC), FastMCP

- Developed an MCP-compliant image augmentation server exposing Albumentations via structured JSON tools, with deterministic seeds and reproducible pipelines.
- Implemented a 7-stage hook system for metadata logging, batching, and preset selection, enabling controlled CV experimentation and semantic verification workflows.
- Packaged for PyPI/uv with cross-platform integration (Claude Desktop, Kiro IDE, CLI) and configurable operation without API keys to meet enterprise security constraints.

WORK EXPERIENCE

Data Scientist - Computer Vision R&D, CarScan — 2021 – 2022

Remote - Agile, Kanban | Insurance AI | Reporting to CTO

A firm at the intersection of AI and the automotive insurance industry with ~100 employees and clients like Telesure, GT Motive, Vieva , and Allianz, operating in South Africa, Nigeria, Kenya, Ghana, India, and the Middle East.

Technologies: PyTorch, TensorFlow, TensorFlowJS, Detectron2, OpenCV, TorchServe, Docker, Flask, AWS EC2/S3, DVC, WandB, Label Studio, SuperAnnotate

Skills: Image classification, object detection, segmentation, model compression/quantization, 3D vision, real-time inference, Agile facilitation, technical documentation

- Deployed a 10-class car image classifier (>96% accuracy, <300ms inference) in TensorFlowJS; reduced model size to ~3MB.
- Replaced U2-Net with DIS for salient object segmentation; cut training loss to 0.06 after 200k+ iterations and deployed via TorchServe/Flask.
- Built barcode segmentation model (Mobilenet v2/v3) compressed from 22MB → 1.2MB, achieving 96.5% IoU on live video frames.
- Created car color detection algorithm using Detectron2 + OpenCV; deployed as a microservice with Flask, Docker, and AWS.
- Scaled datasets from <800 → 50k+ images using a Selenium + AWS scraper and augmentation pipelines.
- Introduced reproducibility practices (WandB, DVC, Cookiecutter) and authored internal technical docs that became team-wide standards.

Project Manager – Engineering & Manufacturing, Mellcon Engineers — 2013 – 2019

Delhi NCR - JIT, Kaizen | Industrial Manufacturing | Reporting to MD

A Manufacturing firm with ~150 employees engaging in complete design and manufacture of industrial equipment including compressed air treatment systems and refrigeration equipment, with clients like Alstom, GE India, Indure and Bhabha Atomic Research Centre.

Technologies: ERP, CRM, mentoring, AutoCAD, SolidWorks

Skills: Project lifecycle management, cross-functional leadership, compliance & audits, Lean/5S, stakeholder communication, vendor coordination

- Managed 20-person cross-functional team across design, manufacturing, QA, and delivery; oversaw 10+ concurrent projects/quarter with portfolios up to $400K
- Owned full project lifecycle: P&ID analysis, BOM creation, mechanical design, vendor coordination, QA/QC testing, and client delivery
- Ensured compliance with ISO 9001 and BPVC standards; audited and validated production pipelines
- Spearheaded 5S implementation and lean manufacturing across shop floor and inventory system; generated $10K+ in cost savings within 3 months

EDUCATION

Master of Business Administration (MBA) | Quantic School Of Business and Technology — 2023

Dissertation: Burberry APAC Market Strategy — Digitalization & ESG for Gen Y Consumers

- Designed omnichannel strategy and segmentation plan for $20B Chinese luxury market, integrating ESG and cultural frameworks.

Master of Science in Mechanical Engineering | University of Exeter — 2017

Dissertation: Lean Manufacturing Techniques to Reduce Electrical Wastage in Food Processing SMEs

- Collaborated with a UK SME to pilot Lean interventions, cutting electrical waste through Kaizen and takt-time optimization frameworks.