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. Delivered production-ready prototypes including an MCP augmentation server, a Korean CPC multi-agent system, and a validated RAG pipeline with measurable ranking improvements. Skilled in FastAPI, Docker, AWS, and reproducible evaluation for LLM behavior, retrieval quality, and agent coherence.

AWS Certified Cloud Practitioner and AI Practitioner, with 2 years of portfolio work across deep learning, LLMOps, and MCP deployment. Strong documentation and cross-functional collaboration skills, converting research concepts into reliable, production-ready workflows.

Agentic AI	Designing agent reasoning pipelines, function-calling and tool use, multi-agent orchestration, retrieval-augmented generation (RAG), LlamaIndex and LangChain, prompt engineering
MCP & Tooling	Custom server development, tool integration, cross-platform agent interoperability, hooks and testing
Deployment	Docker, FastAPI, TorchServe, real-time/streaming inference, microservices architecture
MLOps & Cloud	DVC, MLflow, WandB, Git, GitHub, AWS (SageMaker, EC2, S3), CI/CD pipelines, certified Cloud Practitioner + AI Practitioner
LLM & GenAI	LLM experimentation, model selection, embedding workflows, semantic search, tuning and evaluation for accuracy, latency, and cost

SELECTED PROJECTS (2024–PRESENT)

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

AI Engineer

Technologies: Python, BM25, sentence-transformers, hybrid search, DeepSeek-V3, Qwen2.5-72B, Llama-3.1-70B, Jupyter, Docker, Grafana Hugging Face

Skills: RAG pipelines, retrieval evaluation (Hit Rate/MRR), hybrid scoring, LLM benchmarking, prompt experimentation, LLM-as-a-Judge evaluation, reproducible research, performance analysis

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

***Impact**: Demonstrated production-ready retrieval configuration with measurable ranking gains and a validated model-selection strategy*

- Evaluated 3 retrieval methods (BM25, Vector, Hybrid) and achieved a 23% MRR improvement with a Weighted Product hybrid model (alpha=0.4).
- Benchmarked 5 LLMs and 3 prompt templates using LLM-as-a-Judge scoring; selected DeepSeek-V3 + structured prompting for 90% relevance.
- Conducted extensive experiments on query rewriting, re-ranking, and hybrid fusion strategies, validating what improves vs degrades performance.

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

AI Strategist / Prototyper

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

Skills: Persona design, multi-agent orchestration, evaluation & benchmarking (coherence/latency/safety), RAG pipelines, prompt engineering

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

***Impact**: Delivered a working CPC demo showing how persona-driven companions increase immersion, retention, and cultural authenticity.*

- Designed and implemented multi-agent orchestration with CrewAI to maintain distinct cultural personas and safe role boundaries.
- Built interactive rooms (kitchen, garden, study, etc.) linking language practice to cultural themes and objects.
- Logged and evaluated conversations for persona coherence, latency, and safety, establishing a repeatable CPC evaluation process.
- Implemented constrained role boundaries, guardrails, and memory handling to ensure stable agent behavior across long conversations.

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

AI Engineer

Technologies: Python, Albumentations, MCP (JSON-RPC), CLI, PyPI, FastMCP
Skills: Computer vision (image augmentation, transforms, segmentation), MCP toolchain design, schema validation, hooks & testing, reproducibility, metadata logging, prompt parsing, batching, deterministic seeding, MLOps practices, cross-platform integration

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

***Impact**: Standardized augmentation as MCP tools, creating reproducible pipelines that can scale across agents, IDEs, and R&D workflows.*

- Built an MCP-compliant image augmentation server exposing Albumentations via structured JSON tools (augment_image, presets, reproducible seeds).
- Integrated Gemini 2.5 Flash Image (Nano Banana) VLM into augmentation server for semantic testing and visual verification
- Implemented a 7-stage hook system with metadata logging, deterministic seeding, batching, and preset pipelines for repeatable CV experiments.
- Designed for cross-platform integration (Claude Desktop, Kiro IDE, CLI demos) and packaged for PyPI/uv for zero-friction testing.
- Configurable without exposing API keys, addressing enterprise security concerns.

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.
- Advocated and implemented WandB + DVC for reproducibility; introduced Cookiecutter + Sphinx for repo hygiene.
- Authored internal technical papers and experiment reports to standardize methods and share findings across R&D.
- First in company to achieve Agile certification; mentored team on sprint planning, retros, and story estimation.
- Proposed and led internal process upgrades (experiment tracking, version control, documentation standards) adopted team-wide.

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.

ADDITIONAL INFORMATION

- **Languages:** English, Korean-B2
- **Dissertation Links:**
  - Enterprise Deployment Study — AWS Multi-Agent Architecture (Design Study), **2025**
  - Turn Detection in AI Language Learning, **2025**
  - Lean Manufacturing Techniques To Reduce Electrical Wastage In Food Processing SMES, **2017**
- **Certifications:**
  - Professional Scrum Master™ I (PSM I)
  - AWS Certified Cloud Practitioner
  - AWS Certified AI Practitioner
- **Community:**
  - Language Cafe Discord Server Dev Team and STEM channel waiter (15k members) | Language learning community
  - 100 Days Of Cloud Discord Server Mod and Dev Team(6.5k members) | Cloud learning community