ABHAY A

AI Engineer — Agentic AI

abhayanil77@gmail.com
Linkedin
GitHub
+918139082885
Kerala, India
Portfolio

PROFESSIONAL SUMMARY

AI Engineer with 3+ years building production AI across agentic automation, LLM systems and edge computer vision. Designs agentic RPA, MCP tool servers and LLM gateways that connect AI agents to enterprise systems, validated with AI evals. Combines business-analyst skills (process discovery, automation gap analysis, scope and ROI) with hands-on architecture, taking solutions from discovery to production.

PROFESSIONAL EXPERIENCE

AI/ML Engineer JUN 2026 - PRESENT

NGXP Technologies, Kochi India

- Leading AI and automation discovery for a large enterprise as AI business analyst: assessed 600+ processes and shortlisted 100+ process cycles, defining scope and ROI.
- Designing an agentic RPA platform that pairs custom LLM agents with a plugin on the open-source Cua framework, cutting the licensing cost of tools like UiPath.
- Architecting MCP tool servers for ERP/CRM behind an LLM gateway (routing, fallback, cost control) with human-in-the-loop approvals; platform onboarding underway.
- Defining the AI evals and LangSmith tracing plan to verify agent accuracy before rollout.

Artificial Intelligence Developer SEPT 2025 - APR 2026

Acerite

- Designed and deployed a humanized LLaMA-based review response system using Hugging Face, Ollama, Docker, and RunPod serverless, cutting LLM inference cost by ~80%.
- Architected a 3-stage fine-tuned LLM pipeline with 4-bit quantization, achieving 98% high-quality contextual outputs.
- Implemented multilingual generation (NLLB) with automated web scraping and summarization for scalable, real-time deployment.

Associate AI Engineer AUG 2023 - AUG 2025

Safetywhat (Facilig Intelligent Solutions Pvt Ltd), Kochi India

- Built and deployed production-grade vision AI (ANPR, facial recognition, intrusion and posture analytics) with custom and fine-tuned models reaching up to 95% accuracy in real time.
- Engineered edge pipelines with DeepStream, OpenVINO, YOLO, TensorRT and NVIDIA Jetson, applying knowledge distillation to run real-time AI on CPU-only hardware.
- Automated data annotation with Vision-Language Models (VLMs), accelerating dataset preparation and client POC-to-production delivery.

Data Science Intern NOV 2022 - JULY 2023

Luminar Technolab ,Kochi India

- Built end-to-end projects on ML/DL/NLP and Big Data using LSTM, SQL, Hadoop, Spark, and Sqoop. Delivered solutions from data collection to deployment with impactful visualizations in PowerBI and Tableau.

ACADEMIC QUALIFICATION & CERTIFICATIONS

Bachelor of Science: Chemistry JUN 2019 - MAY 2022
Kerala University, Thiruvananthapuram
Actively participated in NCC, C Certificate holder

Professional Development in Big Data Analytics and Data Science (Certification) NOV 2022 - MAY 2023
National Council for Technology and Training

PROJECTS

- 3D Ergonomic Assessment: Developed a 3D posture risk analysis system using SMPL, REBA/RULA frameworks, and MeshLab for ergonomic risk classification.
- Hivefloor: Built a local-first multi-agent desktop office where a boss agent routes tasks to AI coding agents in real terminals, with shared memory, sandboxing and approval gates.
- Fashion Industry RAG Assistant: Engineered a production-grade RAG chatbot (WhatsApp/Instagram) enabling real-time product intelligence via LLM and vector search.
- SpatialVLM: Added a grounding head to 4-bit Qwen2-VL-7B for pixel-precise object localization, reaching ~58% Acc@0.5 on RefCOCO on an 8GB GPU.
- WebRTC Video Calling: Built P2P browser video calls with Socket.io signalling, plus a GStreamer pipeline exposing calls as RTSP/YouTube Live for YOLO-based vision.

TECHNICAL SKILLS

Programming: Python
ML/DL: PyTorch, TensorFlow, Keras, Scikit-learn, Transformers, Knowledge Distillation (KD), TensorRT, CPU/GPU Optimization
GenAI: LLaMA, Hugging Face, Ollama, RAG, Prompt Engineering, 4-bit Quantization, NLLB
Agentic AI & LLMOps: MCP Servers, Agentic RPA (Cua), Computer-Use Agents, Code Mode, OpenCode, LangChain, LangGraph, LLM Gateway, AI Evals, LangSmith
Computer Vision: YOLO, OpenCV, DeepStream, OpenVINO, GStreamer, ResNet, Object Detection, Segmentation, Pose Estimation, Real-Time AI, Edge AI
Deployment: Docker, AWS, RunPod (Serverless), Jetson, RTSP, WebRTC, Kafka
Data & DB: NumPy, Pandas, Matplotlib, CuPy, MongoDB, SQLite, FAISS, ChromaDB, Milvus, Qdrant

DECLARATION

I hereby declare that the information provided above is true and correct to the best of my knowledge and belief.

Date: 01/11/2026 Abhay.A