Nikhil C R

Thrissur, Kerala | nikhilcramakrishnan@gmail.com | +91 9539356872 | nikhilcr.com

linkedin.com/in/nikhilcr01 | github.com/nikhilcramakrishnan

Professional Summary

AI/ML Engineer with 1.5+ years of experience building production-grade **Generative AI, agentic, and multimodal AI systems**. Experienced in **LangGraph, RAG, LLM fine-tuning, voice AI, and computer vision**, with hands-on experience developing and deploying AI systems across **agri-tech, med-tech, and industrial inspection** domains.

Experience

Associate AI/ML Engineer, Revin Krishi — Nov 2025 – Present

- Developed an AI-powered agronomy platform using **LangGraph**, implementing stateful agent workflows, multi-step tool calling, and dynamic access to soil, weather, crop, and field-specific data.
- Architected **RAG pipelines** for AI applications, designing end-to-end workflows spanning document ingestion, preprocessing, embedding generation, vector retrieval, context construction, and LLM-based generation.
- Architected and engineered a **real-time multilingual voice AI system**, integrating STT, TTS, and agentic workflows to enable natural voice-based interactions in regional languages.
- Optimized **LLM inference and prompting strategies** to improve model efficiency, response quality, and runtime performance across AI workflows.
- Designed **AI evaluation and guardrail workflows** to assess response quality, retrieval relevance, hallucination behavior, and tool-calling reliability across conversational AI systems.
- Architected and implemented **production-grade AI services** with asynchronous processing, error handling, logging, and **Langfuse** observability for tracing agent execution, LLM calls, latency, and token usage.

Junior AI Engineer, Ombrulla — May 2025 – Nov 2025

- Developed industrial computer vision systems for automated visual inspection using **YOLO, Mask R-CNN, Mask2Former, and Grounding DINO**, improving inference efficiency by **30%** through model optimization and pipeline tuning.
- Designed and implemented a **RAG-based document intelligence pipeline** for extracting structured insights from large technical documents.
- Built multi-stage vision pipelines combining **Grounding DINO, SAM, and YOLO** for open-world detection and segmentation workflows.
- Integrated **VLMs** into visual inspection workflows for image understanding, defect analysis, and automated reporting.
- Optimized **computer vision inference pipelines** to improve processing efficiency and reduce end-to-end workflow latency.

Intern, Image and Vision Computing Lab, Digital University Kerala — Feb 2024 – Jul 2024

- Proposed a few-shot road damage detection framework by modifying the **Region Proposal Network (RPN)** of **Faster R-CNN** to improve rare-class proposal quality through integrated **channel attention**.
- Applied a **two-stage fine-tuning (TFA)** approach using **Detectron2** with a **ResNet-101** backbone, training on abundant base damage classes before fine-tuning on as few as **3–10 annotated samples** of rare damage categories.
- Achieved improved performance on rare few-shot classes across **3-shot, 5-shot, and 10-shot** configurations, evaluated using **mAP@50**, with model training and optimization tracked through **Weights & Biases**.

Education

- **Kerala University of Digital Sciences, Innovation and Technology** — 2022 – 2024
  Master of Science in Computer Science with specialization in Machine Intelligence
- **University of Calicut** — 2019 – 2022
  Bachelor of Science in Computer Science

Projects

- **AI Web Research Agent** *GitHub*
  - Designed a **multi-node LangGraph agent** with shared state and custom reducers, supporting **parallel web and news search**, conditional routing, retry limits, and loop-safe fallback logic.
  - Implemented a **self-evaluating retrieval loop** where the agent assesses search snippets and dynamically decides when to scrape deeper content via **Crawl4AI** with LLM-guided extraction.
  - Built a full-stack research platform with a **FastAPI backend** and **React frontend**, producing confidence-aware structured reports.
- **Interview Agent** *GitHub*
  - Developed an AI-driven interview simulator leveraging **agentic reasoning** for dynamic question generation, response analysis, and feedback generation.
  - Designed a **backend API architecture** for seamless real-time interaction with the frontend.
- **Llama-based Mental Health Counseling Assistant** *GitHub*
  - Fine-tuned a **Llama 8B** model using **SFTTrainer and QLoRA**, implementing **4-bit quantization** for parameter-efficient training.
  - Curated multi-turn conversational datasets using **instruction-tuned chat templates** and token filtering to support consistent, empathetic responses.
- **Aviation Regulations RAG System** *GitHub*
  - Built a **domain-specific RAG system** over live US federal aviation regulations by integrating directly with the official **eCFR XML API**, enabling current regulatory querying without manual data uploads.
  - Implemented **structure-aware document parsing** that preserves regulatory hierarchy as rich chunk metadata, enabling traceable, citation-backed answers at the section level.
  - Applied a **two-stage retrieval pipeline** combining semantic search over **ChromaDB** with **Cohere reranking** for dense regulatory text.

Awards and Recognitions

- **1st Place, AI Filmmaking & Short Film Competitions** Dec 2024 & Jun 2025
  100xEngineers Generative AI Buildathons
- **Winner, INNOFEST Hackathon** Sep 2023
  Kerala University of Digital Sciences, Innovation and Technology
- **Best Contribution to Corporate Life Award** Mar 2022
  Prajyoti Niketan College

Certifications

- **Certified Blockchain Associate** — Kerala Blockchain Academy

Skills

- **AI & Machine Learning:** Deep Learning, Computer Vision, Generative AI, Vision-Language Models (VLMs), LLMs
- **LLM & Agentic Systems:** LangGraph, LangChain, RAG, LLM Fine-Tuning (LoRA/QLoRA, PEFT), Prompt Engineering, Tool Calling
- **ML Frameworks:** PyTorch, Hugging Face Transformers, OpenCV, Detectron2, MMDetection, Ultralytics YOLO
- **Inference & Deployment:** TensorRT, ONNX Runtime, FastAPI, Docker, REST APIs, CI/CD
- **Vector Databases:** ChromaDB, FAISS, Milvus, Neo4j
- **MLOps & Observability:** Langfuse, MLflow, Weights & Biases (W&B)
- **Programming & Data:** Python, SQL, NumPy, Pandas, Git