



Lahiru Munasinghe

AI Engineer · Agentic Systems · Machine Learning

lahirumm92@gmail.com

linkedin.com/in/lahiru-munasinghe

Sri Lanka · Open to remote

lahirumunasinghe.com

github.com/lahirumanulanka

PROFILE

AI Engineer with **8+ years** of software engineering experience and **1.5+ years** shipping production multi-agent systems, RAG pipelines, and LLM-powered backends. MSc Data Science (Merit), with a research thesis on multilingual multi-agent AI. I work end to end, from agent orchestration and retrieval through model training and explainability to deployment. Years as a mobile technical lead for airlines, banks, and telcos taught me the engineering discipline I now bring to AI products.

CORE SKILLS

Agent frameworks	CrewAI, LangGraph, LangChain, AutoGen, PydanticAI, Model Context Protocol (MCP)
LLMs & GenAI	OpenAI, Anthropic Claude, Gemini, Ollama, function calling, structured output, LoRA/QLoRA, Stable Diffusion
RAG & vector DBs	Qdrant, ChromaDB, pgvector, FAISS, LlamaIndex, Sentence-Transformers, hybrid search, re-ranking
Machine learning	PyTorch, scikit-learn, XGBoost, LightGBM, transfer learning, time series (SARIMA, Prophet, LSTM), SHAP, LIME, Grad-CAM
MLOps & data	MLflow, ONNX, Hugging Face, Kafka, Flink, Spark, Hadoop, Neo4j, Grafana
Backend & infra	Python, FastAPI, Docker, Redis, PostgreSQL, AWS (Lambda, EC2, Bedrock), Railway, CI/CD, WebRTC, Twilio

EXPERIENCE

AI Engineer / Software Consultant · Softvil Technologies	Aug 2024 – Present
- Design and ship multi-agent systems, RAG pipelines, and LLM-powered backends for enterprise robotics and SaaS clients. - Built the LLM inference layer and human-in-the-loop agent workflows for an AI-powered remote robotic presence platform used by enterprise clients worldwide. - Engineered sub-100ms WebRTC control pipelines and real-time services on FastAPI, Redis, PostgreSQL, and AWS. - Integrated 8+ LLM APIs into production services using structured output and function calling.	
Associate Technical Lead (Mobile) · Axiata Digital Labs	2022 – 2024
Led mobile engineering for digital banking and media products: Boost Digital Bank, Stage11.	
Associate Technical Lead (Mobile) · Arimac Lanka	2019 – 2022
Led iOS delivery for airline and telecom apps: Etihad Airways, Air Astana, Dhiraagu.	
Software Engineer, Full Stack · Happiness Mountain Inc. (USA, remote)	2017 – 2020

SELECTED AI PROJECTS

HireAgent: multilingual multi-agent job platform (MSc research)

Informal employers in Sri Lanka post jobs by voice or text in Sinhala, Tamil, English, or Singlish. An intent gateway extracts a structured posting, then a 6-stage CrewAI pipeline (validation, enrichment, classification, moderation, content, image prompt) publishes it. Built as 4 repositories (7 services on Railway) with real-time search and background AI processing kept on separate paths, plus an MCP server for agent provider search. **Results:** extraction F1 0.83, 433ms real-time search, scam recall 1.00 with zero false positives, 92.5% intent routing accuracy.

CrewAI · OpenAI · MCP · FastAPI · Twilio Voice · PostgreSQL/pgvector · Redis · Railway

Batta: personal AI voice assistant (macOS · iOS · CarPlay)

Real-time voice assistant on the Gemini Live API with interruption, on-device "Hey Batta" wake word, and shared encrypted memory across devices. 30+ self-describing tools (Gmail, iMessage, WhatsApp, calendar, maps, Mac control via Accessibility API), filtered per platform. Irreversible actions need a human tap and reversible ones support undo. No API keys in apps (short-lived tokens). Apple granted the CarPlay conversational-app entitlement. Built solo: ~40K lines of Swift and 6.6K lines of Python.

[Gemini Live](#) · [Swift 6](#) · [SwiftUI](#) · [App Intents](#) · [CarPlay](#) · [FastAPI](#) · [PostgreSQL](#) · [Railway](#)

Strategy Sync AI: RAG strategy alignment (live on Hugging Face)

Scores how well action plans support strategic goals using sentence embeddings, ChromaDB retrieval, and RAG recommendations, evaluated with Precision@K, MAP, and NDCG.

[Sentence-Transformers](#) · [ChromaDB](#) · [RAG](#) · [OpenAI](#) · [Streamlit](#)

Visual Emotion Recognition

CNN/ViT transfer learning on 43.7K images (6 classes), Stable Diffusion synthetic data for class balance, Grad-CAM/SHAP/LIME explainability, and ONNX export to a real-time API and mobile app.

[PyTorch](#) · [timm](#) · [Stable Diffusion](#) · [ONNX](#)

KASI · Kiku

AI-powered expense tracker for Sri Lankan users; bilingual Sinhala/English AI play companion for children aged 4–7.

[FastAPI](#) · [LLM APIs](#) · [Speech](#) · [Flutter](#)

MACHINE LEARNING & DATA SCIENCE PROJECTS

Bank Term-Deposit Prediction

Production ML pipeline on 86K+ records: SMOTE, six models tracked in MLflow, threshold tuning, and a deployed Gradio app and FastAPI service. LightGBM ROC-AUC 0.93, 78% precision, 67% recall.

[LightGBM](#) · [XGBoost](#) · [SMOTE](#) · [MLflow](#) · [FastAPI](#)

Household Power Forecasting

Compared SARIMA, Prophet, LSTM, XGBoost, and LightGBM for daily demand. Best was XGBoost (test RMSE $\approx$ 0.0125).

[Prophet](#) · [SARIMAX](#) · [LSTM](#) · [XGBoost](#)

Optimization & Big Data

GA vs MIP sprint task allocation; Spark vs Hadoop graph processing at LiveJournal scale; Kafka + Flink streaming analytics with Grafana.

[DEAP](#) · [PuLP](#) · [PySpark](#) · [Kafka](#) · [Flink](#) · [Neo4j](#)

EDUCATION

MSc Data Science, Merit

Coventry University UK · NIBM

Thesis: A Multi-Agent Multilingual Voice-and-Text Job Posting Platform for Inclusive Digital Recruitment

BSc (Hons) Information Technology

University of Moratuwa