MEDHAJA H L

www.linkedin.com/in/medhaja| medhajahlbhat@gmail.com |
https://github.com/medhaja | medhaja.me

EDUCATION

Manipal Institute of Technology Manipal, India
B.Tech in Computer Science and Engineering (Minors: Big Data)

PROFESSIONAL EXPERIENCE

Biz-Metric India Pvt Ltd Bengaluru
GenAI Technical Lead Feb 2026 – Present

- Led the development of an enterprise-scale agentic transition accelerator for ADF-to-PySpark modernization, incorporating deterministic translation, LLM-based semantic validation, and automated repair workflows, resulting in 90% reduction in manual effort and 90% operational cost savings.
- Architected a sophisticated multi-model orchestration framework utilizing LangGraph and LiteLLM, facilitating adaptive routing across Claude and OpenAI ecosystems to maximize semantic precision while minimizing expenditure

Infinite Computer Solutions Bengaluru
Senior Software Engineer Oct 2024 – Feb 2026

- Deployed Mistral-7b model through VLLM in Google Kubernetes with speculative decoding(llama 3.1 as draft model)
- Developed a proof-of-concept Agentic Bot leveraging Autogen for multi-agent orchestration and FastAPI for deployment, enabling automated document entity extraction.
- Developed a Document Quality Scoring system combining LLMs and OpenCV, quantifying scan quality for compliance workflows

SourceFuse Technologies India Private Limited Remote
Software Engineer-AI/ML April 2023 – April 2024

- Developed RAG applications tailored for various scenarios using LLMs like OpenAI, Mistral, Llama2,AWS Bedrock, Langchain, and FAISS Vector Database. This project also included Prompt engineering and creation of API through FastAPI.
- Led the strategic design and execution of a robust ETL pipeline using Python, significantly enhancing data integration and analytics capabilities. This initiative increased accuracy by 28% in the result.

EXL Services Remote
Analyst June 2021 – Dec 2022

- Performed Exploratory Data Analysis on the client data, data append from different sources, and calculation of Credit scores using PySpark, automating the process with GitHub Actions, containerizing them using Docker. Implemented Hyper Parameter Tuning using Kubeflow.
- Managed a team of two to create a SaaS Automation tool using Python.
- Created NLP Question Answer System as a tool for Model Validation which reduced the turnaround time by a factor of 26x. Deployed using the Flask framework.

PyImageSearch Remote
Machine learning Engineer (Freelancer) June 2020 – Mar 2021

- Converted Python projects into cloud-based notebooks to leverage the scalability and flexibility of cloud computing.
- Wrote AI blog posts to share insights and best practices for converting Python projects into cloud-based notebooks, including tips for version control, bug fixing, and optimizing for performance.

Harman International (India) Private Ltd Bangalore, India
Machine learning Engineer Intern Sept 2019 – June 2020

- Developed end-to-end implementation of Anomaly Detection using Deep Generative models for real time data from 5G Telematics device. Implemented Deep Learning based Autoencoder algorithm in Sagemaker, Data collection from AWS S3 bucket, Data Preparation(scikit-learn and Numpy), AutoML and Neural Architecture Search for hyper parameter tuning(Scikit learn).

SELECTED PROJECTS

Isocline — Open-Source Visual Platform for Agentic AI Workflows · GitHub · Case study

- Built an open-source, self-hosted platform for designing and running multi-agent workflows on a visual canvas, with agents, tools and human-approval steps.
- Designed a durable runtime on PostgreSQL with parallel DAG execution, crash recovery, checkpoints and rerun from any node; v3.3.0 passed 233 tests and 10/10 tested durability scenarios.
- Implemented per-agent model and tool configuration across OpenAI, Anthropic, Gemini, OpenRouter, Ollama and OpenAI-compatible servers.
- Built run inspection showing each node's output, token usage, latency, errors and estimated cost.
- Shipped a Windows installer and an end-to-end tutorial covering setup, provider configuration, workflow execution and human approval.
- Stack: FastAPI, Celery, PostgreSQL + pgvector, Redis, Next.js Tutorial

L1 Interviewer Bot — voice-first technical screening agent

- Designed a voice-first conversational agent that screens technical candidates, using RAG to probe specific technical domains.
- Built automated grading that produces technical and communication scores, reducing manual effort in L1 screening rounds.
- Optimized voice-response latency for natural, human-like conversation. Demo

PROFESSIONAL DEVELOPMENT

DeepLearning.AI
MLOps Specialization
Duke University
Large Language Model Operations (LLMOps) Specialization

TECHNICAL SKILLS

- **Languages:** Python, SQL, C, C++, Java
- **GenAI:** LangGraph, LangChain, LiteLLM, AutoGen, RAG, vLLM, FAISS, ChromaDB, pgvector, OpenAI, Anthropic, Amazon Bedrock
- **Backend and infrastructure:** FastAPI, Flask, Celery, PostgreSQL, Redis, Docker, Kubernetes (GKE), GitHub Actions, Next.js
- **ML and MLOps:** PyTorch, TensorFlow, Keras, scikit-learn, MLflow, Kubeflow, Amazon SageMaker, Airflow, PySpark, Pandas, NumPy, NLTK
- **Cloud:** AWS, GCP, Azure