Abdul Hakkeem P A

◇ abdulhakkeempa2002@gmail.com ◇ +91-9995385225 ◇ Location: Kochi, India
◇ https://www.linkedin.com/in/abdul-hakkeem-pa ◇ https://github.com/abdulhakkeempa
◇ https://leetcode.com/u/jiEmgu4WpU

Education

Cochin University of Science & Technology July 2021 - July 2026
M. Sc (5 year integrated) in Computer Science (Artificial Intelligence & Data Science) CGPA (Current): 8.4

Technical Skills

Proficient in Python, PyTorch, Tensorflow, Numpy, Pandas, Scikit-Learn, Matplotlib, Docker, C++, JavaScript, Object Oriented Programming, Data Structures and Algorithms, Unix/Linux, FastAPI, Node.js, PostgreSQL, MySQL, AWS.

Professional Experience

Nav Technologies January 2024 - May 2025
AI Engineer Intern

- Designed and deployed an **AI agent for conversational search** on a core banking platform serving microfinance banks and NBFCs, uses a hybrid architecture with **RAG** and **few-shot learning** to generate accurate **SQL queries** for real-time data retrieval.
- Implemented **evaluation metrics** and **self-learning** mechanisms to continuously improve system accuracy and performance.
- Built the entire application stack using **Python**, **FastAPI**, **Next.js**, **PostgreSQL**, **AWS EC2**, and **Nginx**, with end-to-end CI/CD workflows and data pipelines.
- Developed a **payment gateway** for a major financial organization to orchestrate **transactions between multiple banks** for their users. Designed a **queue-based asynchronous architecture** for **scalability**. The system is built using **Node.js, Kafka, and PostgreSQL**, incorporating **standard encryption, logging, and authentication techniques**. Developed the whole system including **REST APIs** and **webhook endpoints** for the dashboard and payment status checking.

Projects

A Comparative Study on MLOps Architectures for Scalability, Latency, and Cost Efficiency GitHub
Skills: MLOps, Model Optimization, AWS EC2, AWS Fargate, AWS Lambda, AWS ECS

* Conducted a study on **MLOps**, researching optimal **cloud architectures** for deploying **high-performance AI models** with a focus on **scalability**, **low latency**, and **cost efficiency**.
* Evaluated AI model performance across domains such as **LLMs, computer vision, and machine learning** by comparing **VMs, serverless architecture, serverless containers, containerized deployments**, and **Kubernetes-based solutions**.
* Explored **model optimization techniques** like **pruning, quantization, and knowledge distillation** to enhance **efficiency, reduce inference costs**, and improve **deployment feasibility** on cloud infrastructure.

Fine Tuning Gemma 2 on Malayalam Language - [Kaggle Competition] Link
Skills: LLM Fine-Tuning, LoRA, Instruction-Tuning

* **Fine-tuned** the **Gemma:2 2B** pretrained model on an **instruction-tuned Malayalam dataset** using **LoRA** for efficient adaptation.

Generative AI-Based Crime Analysis Platform GitHub
Skills: RAG, FastAPI, OpenAI, Embeddings, Semantic Search Python, ChromaDB

* **Developed** a **information retrieval system** using **Retrieval-Augmented Generation (RAG)** and **vector embeddings** to retrieve **contextually similar police complaints**.
* Incorporated **modus operandi analysis** to detect **crime linkages** based on **victim profiles**, **crime geography**, and **suspect behavior patterns**, enhancing **crime pattern detection**.

Distributed Object Detection on Edge Devices using tinyML - [Academic] Link
Skills: Python, Yolo, Object Detection, Raspberry PI, MQTT, tinyML

* Implemented **distributed computing** on edge devices to **reduce central server load** and **latency**, enhancing real-time processing capabilities.
* Developed an **object detection system** that identifies cars from camera streams and compares them to a target image, facilitating vehicle tracking across **distributed devices** with **edge-based processing**.

Research Papers

1. Research paper titled **"Applying Monolithic to Microservices Strategy for Elastic Container Deployment of AI Applications"** accepted and presented at the **2025 IEEE 6th Annual World AI IoT Congress (AIIoT), Seattle, USA**.
2. Research paper titled **"MLOps Challenges in Deploying High-Performance Vision Models: An Empirical Analysis"** presented at the **IEEE CONECCT 2025, Bangalore**.

Open Source Contributions

1. **Smolagents (Hugging Face)** - Contributed to Hugging Face's Smolagents library (22k+ stars) by fixing a critical method signature bug in core agent initialization, enabling proper class instantiation.[PR #1462]
2. **Agents Course (Hugging Face)** - Improved documentation and corrected code samples in LangChain examples for Hugging Face's Agents Course (22k+ stars).[PR #561]
3. **Infisical** - Enhanced user experience in Infisical's secrets management platform (19k+ stars) by resolving UX issues and implementing reliable modal cancel handling [PR #1406]

Certifications

1. AI Agents Fundamentals - Hugging Face Agents Course

Blogs

1. Moving from Request-Response to Async: Engineering Scalable Notification for Better Performance - Medium
2. From EC2 to RDS: Mastering Database Migration with AWS DMS - Medium

Achievements

1. Ranked Top **15 finalists out of 600 teams** at South Indian Bank's Fin-Tech Hackathon at IIT Delhi. Achieved this by presenting a **technical presentation** and building a **personalized recommendation system** integrable with existing banking apps.
2. Ranked Top **30 teams** at **IEEE Global Generative AI Challenge** at **International Level** for developing an **AI Agent for Automated UI Design** on Figma.
3. Ranked Top **15 finalists out of 250 teams** at CareHack 2025 conducted by CareStack & CareRevenue.