Muhamed Adil Edavana Valappil

+971 50 7701549 | muhamedadil03@gmail.com | linkedin.com/in/muhamedadil | github.com/adilzubair

Summary

Applied AI Engineer with hands-on experience building, fine-tuning, and deploying **LLM-powered systems** end to end. Specialized in **Retrieval-Augmented Generation (RAG)**, **Agentic Workflows**, and **Parameter-Efficient Fine-Tuning (LoRA/QLoRA)**. Proven ability to take models from research to **production-grade APIs** using **FastAPI, Python, Golang, and Docker**. Experienced in **local model deployment (Ollama)** for data privacy and security. Available for on-site roles in the UAE.

Experience

Software Developer, Applied AI Aug 2025 – Present
Greentruth Technology *Dubai, UAE*

- Designed and implemented **multimodal AI systems** spanning **classical ML**, **NLP**, and **Generative AI**, focusing on production-grade reliability and scalability.
- Spearheaded the transition to **local LLM execution (Ollama)**, reducing external API dependency and ensuring **data privacy** for sensitive enterprise data.
- Developed rigorous **evaluation pipelines** using **BLEU, ROUGE**, and **human qualitative analysis** to monitor model factuality and significantly reduce hallucinations.
- Integrated complex **LLM inference logic** into production APIs via **FastAPI**, implementing **JWT authentication**, **rate limiting**, and **structured output validation**.
- Fine-tuned transformer models such as **BERT, GPT-2, and Phi-3** on domain-specific datasets using **LoRA and QLoRA** for memory-efficient training on consumer GPUs.
- Developed **multimodal generative workflows** including **Stable Diffusion-based image generation** and vision-enabled chatbot integrations.

Software Engineer May 2025 – July 2025
Cooee *Remote*

- Developed **backend microservices** in **Golang and Python** to support a consumer platform targeting **100K+ users**.
- Optimized **asynchronous workflows** and **message queues** using **RabbitMQ**, reducing end-to-end notification latency by **90 percent**.
- Resolved **production bottlenecks** affecting **1,000+ users** by profiling **API bottlenecks** and **database query paths**.
- Implemented **JWT-based authentication**, **OAuth flows**, and **secure audit logging** for sensitive user operations.

Industrial Trainee, GenAI Edge Deployment May 2024 – July 2024
Intel Corporation *Bangalore, India*

- Optimized **LLM inference pipelines** on **Intel CPUs** using **OpenVINO** for **quantization** and **graph optimization**.
- Automated **model conversion** and **benchmarking workflows** to evaluate **latency** and **throughput** across edge hardware.
- Built an interactive **Gradio interface** for real-time inference with configurable **decoding** and **performance parameters**.

Projects

Local Agentic RAG (Ollama & Private Data)
LangChain, Ollama, ChromaDB, LangGraph, Python

- Built a secure, **local RAG system** using **Ollama** to run models like **Mistral** and **Phi-3** on private enterprise data, ensuring offline capability.
- Implemented strict **agentic logic** using **LangGraph** to ensure responses were 100% grounded in retrieved context with no hallucinations.
- Engineered a data ingestion pipeline with efficient **chunking** and **vector embeddings** for specialized knowledge retrieval.

PRISM, Full Stack Multimodal Chatbot
FastAPI, React, TypeScript, OpenAI, Gemini

- Developed a **multi-conversation platform** featuring **streaming responses**, **multimodal vision support**, and **conversation memory**.
- Implemented a comprehensive **usage tracking system** to monitor **token consumption** and calculate **real-time API costs** per user.
- Designed a modern frontend using **React, Radix UI, and Tailwind CSS**, integrating **JWT-based security** for conversation isolation.

LLM Fine-Tuning & Quantization
Hugging Face PEFT, QLoRA, PyTorch, bitsandbytes

- Fine-tuned **Phi-3-mini** and **GPT-2 Medium** models using **QLoRA**, achieving **99% reduction** in trainable parameters.
- Applied **4-bit quantization** to enable high-performance model training and deployment on **consumer-grade GPUs**.
- Conducted comparative evaluation between **full fine-tuning** and **PEFT** across memory usage and convergence metrics.

Transformer Architecture from Scratch
PyTorch

- Implemented a complete **GPT-style transformer** with **multi-head self-attention**, **positional embeddings**, and **residual connections**.
- Developed **character-level tokenization** and autoregressive decoding to master the fundamentals of modern LLM logic.

Technical Skills

Languages: Python, Go, JavaScript, SQL, C
AI and ML: PyTorch, Transformers, LoRA, QLoRA, PEFT, RAG, Agentic Workflows, LangChain, LangGraph
Tools & Frameworks: Ollama, OpenVINO, FastAPI, Node.js, RabbitMQ, LangChain, React, Gradio
Databases: PostgreSQL, SQLite, ChromaDB, MongoDB, SQLAlchemy
DevOps: Docker, GitHub Actions, CI/CD, Linux, Git
Cloud & Security: AWS (Elastic Beanstalk, WAF, Shield), JWT Authentication

Education

Saintgits College of Engineering 2021 – 2025
Bachelor of Technology in Computer Science *Kerala, India*

Achievements

- **Winner**, Yukthi 2024 Hackathon for building a GPT-powered automation bot.
- **Runner-up**, FrostCode Hackathon for AI chatbot development.
- **Research Intern**, National Chung Cheng University, Taiwan — Published **two peer-reviewed papers**.
- **Finalist**, EY CTF Challenge combining cybersecurity and machine learning.