Muth Ramkumar S

AI / ML Engineer

+91 6379974977 | muthuramlap262003@gmail.com | linkedin | Portfolio

Experience

AI / Machine Learning Engineer — Kore.ai, Inc. — Aug 2024 - Present

LLM-powered Investment Intelligence Platform

- **Designed and deployed** a production-grade **LLM document intelligence pipeline** for financial PDFs and Excel files, incorporating document parsing, contextual chunking, embedding generation, and indexing
- Built a **RAG pipeline for extracting financial insights** and key metrics, achieving **90% accuracy** in information retrieval.
- Architected and shipped a production **multi-agent AI chatbot using AutoGen**, orchestrating multiple agents and tool calls for contextual financial reasoning with end-to-end latency under **15 seconds**.
- Integrated **OpenTelemetry**-based **distributed tracing** across all AI services in the pipeline, enabling end-to-end observability, faster debugging, and identification of performance bottlenecks using **SigNoz**.

LLM-based Support Ticket Automation System

- **Contributed to a fully automated first-level ticket routing pipeline**, reducing manual triage time by **~40%** and improving routing consistency.
- **Analyzed historical support ticket data** to resolve assignment conflicts by clustering similar tickets and aligning predictions with majority assignment groups.
- **Fine-tuned open-source LLMs (Mistral, Phi-3, LLaMA)** using **LoRA**, QLoRA, and **DPO-based feedback learning**; applied model quantisation to reduce inference cost and latency while achieving **85% prediction accuracy**.

AI Assistant Tools for Enterprise Support

- **Architected and developed an asynchronous system of 8 AI tools** to examine tickets, generate summaries, and recommend next actions, boosting agent productivity by **~50%** using **GPT and Claude models**.
- **Integrated AI automation with Zendesk, JIRA, Bitbucket, and Google Chat**, embedding intelligence into existing workflows and reducing response time by **~30%**.
- **Contributed to a GraphRAG-based ticket chat assistant** using vector databases and Neo4j, leveraging historical resolutions to suggest first-level resolution steps for new tickets automatically.

AI Research & Prototyping

- **Fine-tuned open-source LLMs** for intent classification using PEFT techniques and deployed optimised inference pipelines using vLLM, achieving Time-To-First-Token (TTFT) under 200 ms and significantly improving throughput.
- Built a **layout-aware chunking pipeline** using open-source OCR models and Docling to preserve structural elements like tables and headers.

Machine Learning Intern — Kore.ai, Inc. — Feb 2024 - Aug 2024

- Built Food Assist, an agentic AI application using LangGraph and LangChain with local LLM inference via Ollama.
- **Developed an AI-powered PR review system** using LangGraph to evaluate pull requests and suggest code improvements and best practices, integrated with the bitbucket.

Education

Bachelor of Engineering, Computer Science — **Nov 2020 - Apr 2024**

Francis Xavier Engineering College, Tamil Nadu

CGPA: 9.40 — **Awards:** Academic topper

Skills

Languages: Python, SQL, JavaScript
GenAI/Machine Learning: Natural Language Processing (NLP), Retrieval Augmented Generation (RAG), Large Language Model (LLM), Agentic AI
Frameworks: LangChain, LangGraph, FastAPI, Vector DBs (Weaviate, Pinecone), Pytorch, Pandas, Keras, Tensorflow, Docling
MLOps: Docker, Kubernetes, vLLM, OLLMA