THIRUMANODHAM YOGENDRA MANI SAI

tymsai6076@gmail.com | LinkedIn | Bangalore | +91 9302433994

Professional Summary

Professional Software Engineer - AI/ML with **1.8 year** of experience in Gen AI, Machine Learning, and production-grade AI systems. **Expertise** in generative AI, ML fine-tuning, and integrating AI models into full-stack applications.

Skills

Generative AI: RAG, Autogen, **AWS Bedrock**, AWS Lambda, AWS Step Functions, **LangChain**, LlamaIndex, **Agentic AI,** LangGraph, Ollama, QLora, FastMCP

Core Programming: Data Structures, Operating Systems, OOP, **Python**, SQL

Machine Learning: Keras, PyTorch, TensorFlow, **XgBoost,** Scikit-learn, LightGBM, CNN**,** Classification, Regression, Neural Networks**,** Optuna, Knowledge Distillation

Databases: AWS RDS, AWS S3, MySQL, Redis, Neo4j, PgVector

Frameworks & Platforms: GitLab, GitHub, Docker, Azure VMs, Matplotlib, FastAPI, Flask, Streamlit, Pandas, Numpy

Experience

Kadel Labs Private Limited, *Software Engineer - AI/ML, Bangalore* December 2024 – present

- Designed agentic AI workflows, used LangGraph for multi-step reasoning tasks, implemented **tool-calling** using FastMCP, **state management** to orchestrate complex LLM interactions.
- Implemented **hybrid retrieval** combining dense vector similarity and sparse retrieval, followed by **re-ranking** to achieve 94% in retrieval accuracy.
- Modified ResNet18 classifier using Qlora, applied transfer learning on waste segregation images, fine-tuned the pretrained ResNet18 model and **replaced the final fully connected layer** for binary classification and achieving 95% accuracy.

Kadel Labs Private Limited, *Associate Software Engineer Intern, Bengaluru* June 2024 – November 2024

- Integrated AI RAG chatbot into full-stack application used ChromaDB for semantic search for query-document matching, enabling accurate natural language responses to user queries.
- Used LightGBM Regressor to predict the numerical values in excel, **which significantly reduced the OpenAI calls by 95%**.
- Fine-tuned YOLO models for object detection, achieving over 95% accuracy. These models process 360-degree panoramic images, ensuring robust detection from all perspectives.

Projects

AutoML Pipeline

- Built an autonomous ML pipeline using **LangGraph agents** and FastMCP that leverages Qwen Coder to intelligently plan data cleaning, feature engineering, and model selection, with **conditional workflow logic** that ensures optimized model is delivered.
- Implemented 15+ production-ready FastMCP tools for end-to-end ML operations including schema extraction, adaptive data cleaning, feature engineering, hyper-parameter tuning using optuna, and **automated model training**/evaluation.

Advanced RAG

- Engineered multistage retrieval Pipeline with 40% relevance improvement, Implemented hybrid search combining vector similarity and keyword matching(**TF-IDF**). Used LLM based re-ranking, reducing irrelevant results.
- Developed intelligent document chunking with **sentence-aware segmentation and overlapping tokens**, processing multiple formats documents with **metadata enrichment** and MD5 based de-duplication.

Certificates

Data Analytics Professional Coursera ↗ , Machine Learning Coursera ↗ , Python ↗ , Gen AI Engineer Professional ↗

Education

Shri Shankaracharya Technical Campus, 2020 – 2024

B.Tech Computer Science Engineering

CPI: 8.8 (Honors) , GPA: 3.62 and qualified GATE CS Examination AIR - 9965 ↗ (92%ile)