Manideep Sriperambudhuru

Hyderabad, India • manideepsp.16@gmail.com • +91 6300-519-475
https://www.linkedin.com/in/manideepsp • https://manideepsp.github.io/Portfolio-ManideepSP

AI/ML Engineer - LLM Systems & Agentic Workflows

AI Engineer with 3 years building production LLM systems, multi-agent workflows, and ML pipelines. Experienced in RAG, memory design, model routing, and latency optimization. Focused on shipping reliable AI systems with measurable gains.

Core Skills

AI & Deep Learning: Tensorflow, PyTorch, transformers, CNNs, LSTM, LoRA/QLoRA (PEFT, 4-bit), model evaluation, inference optimization
GenAI & Agentic Systems: RAG (Qdrant, FAISS, Chroma, PGVector),Langchain, LangGraph multi-agent workflows, ReAct reasoning, session memory, groundedness validation, prompt optimization
Backend & Cloud: Python, FastAPI (async), Flask, REST APIs, Docker, Redis, SQL, Azure (ADF, Databricks), AWS basics

Experience

Cognine Technologies Hyderabad, India
Applied ML / AI Engineer Apr 2023 – Dec 2025

- Built an AI tutoring agent for a US-based edtech startup platform using LLMs + RAG with LangGraph state and Postgres memory, enabling personalized explanations and learning assistance.
- Designed multi-agent workflows (mood, knowledge, reasoning agents) for adaptive tutoring.
- Reduced report latency from **5–6 minutes to 35–40 seconds** via async execution and parallel LLM calls.
- Optimized token usage and chunking to lower cost while preserving output quality.
- Implemented hallucination mitigation using groundedness checks and structured validation.
- Added moderation and prompt-injection safeguards with controlled tool routing.
- Deployed containerized ReactJS, AstroJS, FastAPI/Flask services integrated with production data pipelines.
- Fine-tuned small LLMs using QLoRA (PEFT) for conversational adaptation.

Projects

QLoRA Fine-Tuning for Empathetic Conversational Agents

- Applied **QLoRA (4-bit PEFT)** on Qwen2.5-3B-Instruct using CUDA.
- Trained on ESConv + GoEmotions for emotion-aware dialogue modeling.
- Achieved **a 33% reduction in MAE** and **a 25% improvement in empathy correlation** on EQ bench 3 .
- Reached **a 95% acknowledgement** and **90% emotion-naming accuracy**.

TensorflowJS-ReactivAI — Client-Side Emotion & Voice Analysis

- Built a fully client-side multimodal AI system in TensorFlow.js with real-time WebGL inference (no backend).
- Implemented facial emotion recognition using MediaPipe (468 landmarks) + custom CNN (FER-2013).
- Fused facial and speech energy signals (Web Audio API) into a temporally smoothed engagement score.

Mental Health Relapse Prediction System

- Built NLP-based relapse risk model using PHQ-9, DSM-5, journals, and clinical notes.
- Engineered contextual features and applied LSTM for temporal risk modeling.
- Integrated RAG-based intervention retrieval; deployed via Dockerized Flask + React.

Diabetes Risk Prediction System

- Developed questionnaire-driven ML chatbot with scheduled retraining.
- Implemented checkpoint-based best-model selection.
- Built containerized Flask architecture with RAG-powered recommendations.

Education

CMR Institute of Technology, Hyderabad
B.Tech in Mechanical Engineering GPA: 8.47/10 2019–2023

Certifications & Recognition

- Star of the Quarter, Cognine Technologies
- Databricks Generative AI Fundamentals