Manideep Sriperambudhuru

Hyderabad, India ● manideepsp.16@gmail.com ● +91 6300-519-475

https://www.linkedin.com/in/manideepsp ● https://manideepsp.github.io/Portfolio-ManideepSP

APPLIED AI / ML ENGINEER

Applied ML / AI Engineer with nearly three years of experience building ML-driven and GenAI systems, including RAG pipelines, agentic chatbots, and personalization workflows. Strong data foundations with hands-on ownership of internal POCs, model evaluation, and iterative experimentation.

CORE SKILLS

Machine Learning: Regression, clustering, scikit-learn, feature engineering, experimentation and evaluation

GenAI / LLMs: RAG, LangChain, LangGraph, LangSmith, vector similarity search, relevance and faithfulness evaluation (RAGAS-style), BLEU, prompt optimization, LLM latency and cost optimization

Engineering & Data: Python, SQL, Medallion Architecture, Azure (ADF, Databricks), Docker

EXPERIENCE

Cognine Technologies — Hyderabad, India

Applied ML / AI Engineer — Apr 2023 - Present

○ Built a hyper-personalized AI tutoring system using LLMs and RAG pipelines to generate adaptive student performance reports used internally by educators.

○ Reworked report generation using concurrent, section-wise LLM execution, reducing end-to-end latency from **4-5 minutes to 35-40 seconds**.

○ Used token reduction and prompt optimization techniques to balance response quality with cost and latency constraints.

○ Designed persona and mood tracking to dynamically adapt tutoring responses during student interactions.

○ Built an agentic chatbot using LangGraph to support stateful, multi-step conversational flows.

○ Introduced rule-based content moderation and acceptance layers to prevent misuse and avoid unnecessary LLM calls.

○ Integrated ML and GenAI components with existing backend services via REST APIs.

○ Applied ML evaluation practices including cross-validation, metric selection, and model tradeoff analysis to guide algorithm choices.

PROJECTS

Mental Health Relapse Prediction (Internal POC)

○ Built an ML pipeline using PHQ-9, DSM-5 questionnaires, patient journals, and clinician notes to estimate relapse risk.

○ Extracted features using sentiment analysis, frequency analysis, and contextual signals from text data.

○ Modeled relapse risk as a continuous score and evaluated multiple approaches using cross-validated error metrics.

○ Selected LSTM-based models where temporal patterns improved predictive performance while maintaining interpretability constraints.

○ Added a RAG-based intervention layer to surface relevant clinical guidance.

○ Deployed a Flask + React application using Docker with real-time database synchronization.

Diabetes Risk Prediction System (Internal POC)

○ Developed a questionnaire-driven chatbot to collect KPIs and predict diabetes risk using ML models.

○ Implemented scheduled model retraining with checkpointing and best-model selection for inference.

○ Built a containerized two-tier Flask architecture with isolated database services.

○ Integrated a RAG-based intervention layer to provide contextual health recommendations.

LLM-Based Resume Screening System (POC)

○ Built a resume screening system using vector databases and RAG pipelines to rank candidates by relevance.

○ Designed similarity-based scoring to automate shortlisting and reduce manual screening effort.

EDUCATION

CMR Institute of Technology, Hyderabad

B.Tech in Mechanical Engineering — GPA: 8.46/10 — 2019 – 2023

CERTIFICATIONS & RECOGNITION

○ Databricks Generative AI Fundamentals ● Star of the Quarter, Cognine Technologies