SAI THANMAI

AI Engineer · Data Scientist · ML Engineer
Virginia, USA | thanmai.sp@gmail.com | linkedin.com/in/sai-thanmai-peddader-pally-8110721b6 | github.com/Thanmai-22

SUMMARY

AI Engineer with 5+ years building production LLM systems, agentic workflows, and ML pipelines. Architected enterprise RAG pipelines processing 22M+ medical documents with LangChain, FAISS, and Pinecone. Built multi-model evaluation frameworks across GPT, Claude, and Llama. Designed prompt engineering strategies and RLHF techniques driving 30% accuracy gains. Deployed ensemble ML models at 4M+ subscriber scale enabling $10M+ revenue impact. Deep expertise in Python, Transformers, AWS, and end-to-end AI operationalization.

EXPERIENCE

Senior Data Scientist | Pycube, Inc · Virginia, USA *Jan 2023 – Present*

- Architected enterprise RAG pipeline using LangChain, FAISS, Pinecone, and AWS SageMaker for document Q&A across 22M+ medical documents, achieving 92% retrieval accuracy and securing multi-year client engagement through executive presentations.
- Built end-to-end Generative AI evaluation framework with Hugging Face Transformers to assess hallucination, coherence, and relevance across GPT, Mistral, and Claude, reducing model failure rates by 30% and establishing deployment standards across 8+ AI projects.
- Designed advanced prompt engineering strategies (few-shot, Chain-of-Thought, RLHF) using OpenAI API and Claude, boosting NLP summarization and QA accuracy by 30% and establishing team-wide LLM optimization standards.
- Deployed AutoML annotation infrastructure on AWS Lambda with active learning loops, orchestrating 50K+ labeled examples, reducing inter-rater error by 25% and accelerating model release cycles from 14 to 4 days.

Data Scientist | ECIL · Hyderabad, India *Jun 2020 – Jan 2022*

- Spearheaded ensemble ML models (XGBoost, Random Forest, LSTM) on 4M+ telecom subscriber sequences achieving 0.91 AUC, improving churn prediction precision by 18% and enabling $10M+ retention revenue.
- Deployed CNNs and LSTMs for real-time customer intent classification from 500K+ monthly voice/text interactions, reducing misrouting by 22%.
- Engineered NLP pipeline with BERT and Hugging Face to extract structured insights from 500K+ unstructured responses, reducing manual analysis time by 60%.

PROJECTS

Real-Time Infrastructure Monitoring Platform [Kafka · TimescaleDB · FastAPI · scikit-learn · Docker] github.com/Thanmai-22/Real-Time-Infrastructure-Monitoring-and-Analytics-Platform
Full-stack observability platform: 12-service event streaming through Kafka, 4-strategy ML anomaly detection (Isolation Forest), WebSocket live dashboard refreshing every 2s. Single Docker Compose deployment.

Mini Container Orchestration Simulator [Python · Docker · Scheduling Algorithms] github.com/Thanmai-22/Mini-Container-Orchestration-Simulator
Kubernetes-style simulator with bin-packing scheduling, resource-aware placement, pod lifecycle management, health checks, restart policies, and eviction logic across simulated multi-node clusters.

Multi-Model LLM Evaluation Framework [PyTorch · Hugging Face · Claude · Llama · AWS Lambda]
End-to-end evaluation pipeline assessing hallucination, coherence, and relevance across GPT, Claude, and Llama with custom metrics, deployed as serverless pipelines.

TECHNICAL SKILLS

AI & LLMs: GPT, Claude, Llama, LangChain, RAG, Prompt Engineering, RLHF, Hugging Face, Transformers, OpenAI API, Agents
ML & DL: PyTorch, TensorFlow, scikit-learn, XGBoost, LSTM, CNN, BERT, NLP, Active Learning
Data & Infra: Apache Kafka, TimescaleDB, PostgreSQL, MongoDB, FAISS, Pinecone, Snowflake, Redis
Cloud & DevOps: AWS (SageMaker, Lambda, EC2, S3, Glue), Docker, Kubernetes, FastAPI, Git
Languages: Python, SQL, Go, Bash, Java, C/C++

EDUCATION

M.S. in Data Science | University of Maryland, Baltimore County *Aug 2022 – Dec 2023*
B.Tech in ECE | SCSVMV University, Kanchipuram, India *Jul 2018 – Jun 2022*