Sudheer Vadlamudi

Dallas, TX | ☎ +1 331 246 6088 | ✉ vamsu2ee@gmail.com | LinkedIn

SUMMARY

AI Engineer with 4+ years of experience designing, deploying, and optimizing Generative AI, LLMs, and ML-powered systems across healthcare, banking, and retail domains. Proven expertise in LLM agent frameworks (LangGraph, PydanticAI, Google ADK), prompt engineering, RAG pipelines, and fine-tuning (LoRA/QLoRA). Strong background in FastAPI, MCP integrations, MLOps, model observability, and vector databases (FAISS, Pinecone, pgvector). Adept at blending backend engineering (Python, FastAPI, Spring Boot) with front-end development (React, Next.js, TypeScript) to deliver AI-powered applications at scale.

EXPERIENCE

AI Engineer (Gen AI)

MediqCare Systems – Dallas, TX | Feb 2024 – Present

- Architected and deployed LLM-powered healthcare agents leveraging LangGraph, React planner-executor chains, and RAG pipelines to process EMR data.
- Integrated vector databases (FAISS + Neo4j) to support chatbot memory, contextual retrieval, and patient query resolution.
- Developed and maintained FastAPI services with MCP integration for secure tool/plugin access and context-aware workflows.
- Designed prompt strategies with safety guardrails to reduce hallucinations and increase accuracy in patient analytics dashboards.
- Built evaluation tooling for trace analysis, model debugging, and drift detection using custom logging pipelines in Elasticsearch & Grafana.
- Fine-tuned LLaMA-2 models with LoRA adapters for healthcare classification tasks, deployed in production using Docker + Kubernetes.
- Collaborated with front-end team to integrate Next.js dashboards for patient insights, ensuring seamless API ↔ UI interactions.

AI Engineer (Gen AI)

Everwest Bankcorp – Remote, USA | Jan 2023 – Nov 2023

- Implemented credit-risk AI agents combining RAG + anomaly detection pipelines with Kafka streaming for real-time fraud detection.
- Wrapped Scikit-learn and PyTorch models into FastAPI endpoints, integrated with loan processing microservices.

- Developed async orchestration for agent workflows, leveraging LangChain & PydanticAI for structured reasoning.
- Benchmarked orchestration frameworks, comparing LangGraph vs. custom ReAct agents, improving latency by 18%.
- Built observability dashboards with Prometheus + Grafana for monitoring prediction drift and hallucination frequency.
- Led Kubernetes Helm deployments of LLM-powered fraud detection services, with secure role-based model access (Keycloak).
- Deployed LoRA-fine-tuned LLMs to classify suspicious transactions, achieving a 25% reduction in false positives.

Data Scientist

V-Mart Retail – Hyderabad, India | Oct 2020 – Jul 2022

- Developed AI-driven demand forecasting and product recommendation engines using Python, Prophet, and XGBoost, integrated into POS workflows.
- Created GenAI assistants for store managers by combining LangChain + Pinecone vector DB + FastAPI, enabling policy/document Q&A.
- Implemented ETL pipelines in Airflow for preparing training datasets from loyalty & transactional systems.
- Built clustering and segmentation models (K-Means, LightGBM) to optimize targeted campaigns, increasing retention by 12%.
- Automated OCR (Tesseract) + NLP (spaCy) pipelines to extract structured data from invoices and vendor forms, reducing manual processing.
- Partnered with BI teams to publish insights via Tableau & Power BI dashboards, visualizing sales and churn metrics.

EDUCATION

- Master of Science in Computer Science – Western Illinois University | Dec 2023
- Bachelor of Technology in Computer Science – Vignan University | Jul 2021

CERTIFICATIONS

- AWS Certified Developer – Associate (2023)
- TensorFlow Developer Certificate (2024)
- Oracle Certified Java SE 11 Developer (2023)

SKILLS

Generative AI & LLMs: LangGraph, PydanticAI, Google ADK, LangChain, Prompt Engineering, RAG, ReAct, CoT, LoRA/QLoRA Fine-tuning, HuggingFace, Guardrails AI, Trulens
Programming: Python (Pandas, NumPy, PyTorch, TensorFlow, Scikit-learn), Java, SQL, TypeScript, Shell

Databases & Retrieval: FAISS, Pinecone, pgvector, ChromaDB, Neo4j, MySQL, PostgreSQL, MongoDB
Backend & APIs: FastAPI, Spring Boot, Node.js, Express.js, REST, gRPC, Protobuf, WebSockets
MLOps & Deployment: Docker, Kubernetes, Airflow, MLflow, DVC, Prometheus, Grafana, Model Observability, CI/CD (Jenkins, GitHub Actions, Azure DevOps)
Frontend & UX: React, Next.js, Angular, Material UI, Streamlit, Power BI, Tableau
Cloud: AWS (ECS, S3, Lambda, CloudWatch), GCP (BigQuery, Cloud Run), Azure (AKS, Azure ML, Cognitive Search)

PROJECTS

- LLM-Powered Patient Risk Engine (Healthcare): Built an agentic pipeline using RAG + LLaMA-2 LoRA fine-tuning to score patient risk in real time, integrated into EMR dashboards.
- Fraud Detection AI Agent (Banking): Architected LangGraph-based orchestration for fraud alerts, combining streaming + LLM reasoning to detect anomalous loan patterns.
- GenAI Retail Q&A Assistant: Developed LangChain + Pinecone powered chatbot for retail policy queries, enabling store managers to interact with AI for inventory and compliance FAQs.