THEEPAN KUMAR GANDHI

+1(312)545-5215 | LinkedIn| tgandhi1107@gmail.com| GitHub | My-Portfolio

SUMMARY

AI/ML Engineer with 3 years of applied experience building and deploying production AI systems including LLM-based agents, multimodal search, recommendation systems, and ML pipelines on AWS and Kubernetes. Experience with RAG, multi-agent orchestration, and classical ML including model evaluation, fine-tuning, monitoring, deployment and performance optimization. Strong background in API development and cloud-based deployment. Authorized to work on F-1 OPT without sponsorship for three years.

EXPERIENCE

Data Science Intern **Jan 2025 – May 2025, Chicago, USA**

Pure Platform

- Engineered and served a multimodal hybrid product search engine supporting simultaneous text and image input, leveraging FAISS for low-latency vector similarity search and achieving 92% top-k retrieval accuracy, helping users find the right product faster
- Fused OpenAI CLIP text and image embeddings as unified vectors with BM25 to improve multimodal relevance and user satisfaction
- Optimized search with FAISS IndexHNSW, achieving 20% faster retrieval and 2x concurrent searches without additional costs
- Structured a flexible search pipeline supporting image uploads, product URLs, and keyword queries, enabling API integration for real-time retrieval and boosting relevance by 25% via weighted alpha blending, improving product discovery for users

Machine Learning Engineer **June 2021-Nov 2021, Coimbatore, India**

Optisol Business Solutions

- Contributed to INNOVOEDGE, a system for processing and extracting insights from real-time network device telemetry data
- Architected and implemented a data prefetcher for efficient acquisition and preparation of diverse real-time data streams, developed and deployed a high-performance prediction engine using Prophet to deliver accurate, timely forecasts
- Designed and implemented API endpoints to efficiently integrate with enterprise systems and external applications
- Optimized data workflows, improving retrieval speed by 30% and enabling scalable analysis in a high-performance database

PROJECTS

End-to-End Multi-Agent Orchestration Platform Using LangGraph & FastAPI

- Engineered production multi agent orchestration system with LangGraph, FastAPI, PostgreSQL deploying 10 specialized AI agents
- Built RAG pipeline with ChromaDB vector store, LlamaGuard moderation, and supervisor routing across specialized sub-agents
- Implemented dual-layer persistence with PostgreSQL checkpointer and Store enabling stateful sessions across conversations
- Containerized multi-service platform with Docker Compose orchestrating FastAPI backend, Streamlit UI, PostgreSQL database

Finance Document Assistant - RAG & Agent System with AWS EKS Deployment

- Achieved 92% Hit@5 accuracy by deploying LlamaIndex hybrid RAG with BM25 and Elasticsearch to AWS EKS via Docker for finance QA
- Reduced deployment time by 40% building CI/CD pipeline with GitHub Actions and Kubernetes deploying Streamlit RAG app to EKS
- Reduced query latency by 23% engineering LangChain agent with LlamaIndex retriever and GPT-4o-mini on EKS with RRF optimization
- Implemented Grafana evaluation pipeline with MRR and Hit@K metrics enabling production model health monitoring for finance RAG

Two-Tower Recommender Platform with MLflow, Airflow, and Monitoring

- Built two-tower retrieval in TensorFlow Recommenders with FAISS index and FastAPI on AWS EKS, achieving 0.81 top 10 accuracy
- Implemented TensorFlow Ranking pointwise model, achieving validation NDCG of 0.904 and tracking runs and models in MLflow
- Orchestrated end-to-end training and release pipeline using Airflow, DVC, Docker, DagsHub, and GitHub Actions, deploying to AWS EKS
- Implemented FastAPI service with Prometheus metrics, Grafana dashboards, A/B logging in Postgres to track latency and model lift

End-to-End Malicious URL Detection using XGBoost

- Deployed XGBoost URL classifier with 92 percent F1 score as FastAPI service using Docker on AWS EC2 for real time API scoring
- Integrated MLflow tracking logging precision recall and ROC AUC across model versions to support reliable deployment decisions
- Orchestrated Airflow DAGs for automated training and batch inference pipelines reducing manual retraining effort by 75 percent
- Implemented GitHub Actions CI CD building Docker images and pushing to AWS ECR for consistent production deployments

SKILLS

Tools and Libraries: Python, PyTorch, TensorFlow, scikit-learn, LangChain, LangGraph, LlamaIndex, Hugging Face Transformers, FastAPI, Docker, Kubernetes, AWS, PostgreSQL, ChromaDB, Elasticsearch, OpenAI APIs, Anthropic APIs, MLflow, GitHub Actions, Git, Airflow, DVC, DagsHub, FAISS, BM25, Grafana, Prometheus, MongoDB

Data Techniques: Gen AI, Multi-Agent Orchestration, RAG, Prompt Engineering, API Integration, Model Evaluation, Model Monitoring, LLM-Fine Tuning, Agentic AI, Containerization, CI/CD Automation, Model Deployment, Vector Search, A/B Testing

EDUCATION

Master of Applied Science, Data Science **December 2025 | Chicago, Illinois**

Illinois Institute of Technology

- Coursework: Machine Learning, Deep Learning, Big Data Technologies, Regression, Statistical Learning, Information Retrieval