UDAY SHANKAR PRASAD

PUNE, MAHARASTRA

+918789637214 # UDAY.PTRU@GMAIL.COM # LINKEDIN

Summary

Highly skilled and results-driven **Senior AI/ML Engineer** with over **5 years of experience** in designing, developing, and deploying AI-driven solutions using large language models (LLMs), **Multi-Agent systems**, and vector-based retrieval architectures. Proficient in **Python, PyTorch, TensorFlow**, and cloud platforms such as **AWS, GCP, and Azure**. Expertise includes building **RAG pipelines**, working with **open-source and fine-tuned LLMs** (e.g., LLaMA, Falcon, Mistral), developing **offline multimodal agents**, and implementing **PEFT techniques (LoRA, QLoRA)** for scalable, resource-efficient deployment. Demonstrated success in building AI-powered applications in healthcare, edtech, and document automation domains using **FastAPI, LangChain, LangGraph, and OCR tools**. Strong advocate MLOps and CI/CD, Docker, and Kubernetes for robust deployment pipelines.

Skills Summary

- **Natural Language Processing (NLP):**
  Experienced in text classification, named entity recognition, sentiment analysis, question answering, and vector search. Worked with Hugging Face Transformers, spaCy, NLTK, and fine-tuned LLMs on domain-specific corpora.
- **Generative AI (GenAI):**
  Proficient in building applications using open-source LLMs (LLaMA, Mistral, Falcon). Developed RAG pipelines integrating vector databases (Pinecone, FAISS), LangChain, Langgraph and multimodal models for text-image tasks.
- **Python Programming:**
  Expert in writing production-grade Python code for data pipelines, API integrations (FastAPI, Flask), automation scripts, and AI model deployment. Strong grasp of Python libraries like NumPy, Pandas, Scikit-learn, Matplotlib, and Pydantic.
- **Machine Learning & Deep Learning:**
  Built and deployed ML models for classification, regression, and time-series forecasting. Hands-on with XGBoost, LightGBM, TensorFlow, PyTorch. Experienced in model tuning (GridSearchCV, Optuna), evaluation, and deployment.
- **MLOps:**
  Experience deploying ML models on **Vertex AI**, managing data with **BigQuery**, orchestrating pipelines using **Cloud Functions**, **Cloud Run**, and **Cloud Storage**. Integrated GCP services into AI applications for scalability and monitoring.

Skill Highlights

- Linux and Python
- Cloud computing (AWS, GCP and Azure)
- Web frameworks (FastAPI)
- Machine learning frameworks (PyTorch and TensorFlow)
- Database management (SQL and MongoDB)
- Vector databases (Pinecone, FAISS, ChromaDB)
- Transformers and BERT
- Natural language processing (NLP)
- Generative AI
- OpenAI GPT-4o and OCR
- Langchain and LangGraphs
- RAG pipeline development
- MultiAI Agent design
- Knowledge Graph construction
- Large language models (LLMs)
- Model Context Protocol(MCP)
- MLOps, CI/CD Pipeline, Docker

Experience

Sr. AI/ML Engineer

iStreet Technologies Pvt. Ltd., Pune

Sep 2024 – Present

Project Description:

Development of a Multi-Agent Health Analysis Pipeline capable of calculating a patient’s health index using multiple open-source LLMs and domain-specific models. Built translation systems, offline chatbots, and advanced RAG systems integrated with vector databases.

Tools, Software, Skills:

- LLMs: LLaMA, Mistral, Falcon, Gemma
- PEFT: LoRA, QLoRA
- Frameworks: FastAPI, LangChain, LangGraph
- Databases: FAISS, Pinecone, ChromaDB
- OCR: pypdf, fitz, python-docx
- Containerization: Docker, CI/CD
- MLOps: MCP Server, Kubernetes
- Natural Language Processing (NLP)
- Generative AI (GenAI)
- Python Programming
- Machine Learning & Deep Learning
- Google Cloud Platform (GCP)

Role and Responsibilities:

- Build **Multi-Agent Pipeline** for different types of **Health System** like Cardio, Pulmo, etc. and **fine-tuned** to build **distilled model** with **medical domain dataset**.

Also, Agent will calculate the Health Index of patient. Integrate the Automation for calculating the Index part when ever database get updated with new records. Used Model Context Protocol(**MCP Server**) for different type of agents tools.

- Worked on Medical Domain for converting medical data into Vector Embedding and store into vector databases (**FAISS, Pinecone, ChromaDB**). Implemented **RAG Pipeline** for Query Based Hybrid search and Semantic search.
- Experience with **PEFT** Techniques to fine-tuning open-source LLMs using **LoRA, QloRA**. LLM (e.g., Open-source models like **LLama, Falcon, Mistral, Gemma**, etc) for various AI-driven applications.
- Developed and implemented **LangChain Agent** in my project using a Multi-Agent Conversation Framework, that can evaluate the result.
- Develop and maintain AI-powered **ChatBot**, virtual assistants, and content generation pipelines using open-soure LLM models like **Whisper**, Llama, mistral, Falcon, etc. It can be run offline with any data sharing. I also conclude **Multimodal LLM** for **Image Generation** and **Speech to Text**. Implemented PEFT techniques to run in any system without GPU.
- Keep myself up to date with emerging AI/ML research and contribute to the company's AI strategy.
- Knowledge of **MLOps, Docker, Kubernetes,** and **CI/CD pipelines** for scalable AI deployment.

Project Description: *Agentic SQL Execution & Feedback Automation System*
Built a **multi-agent system** integrated with **Groq LLM** that fetches SQL scripts from GitHub, sequentially executes them in environment-specific Microsoft SQL Server instances, captures execution logs/screenshots, and generates **AI-driven root cause feedback** for failed queries.

Tools, Software, Skills:

- Python
- FastAPI
- LangGraph
- OpenAI
- pyodbc
- GitPython, Pillow,
- SMTP, SQL Server

Role and Responsibilities:

- Designed and developed an **agentic AI workflow** using **LangGraph** for automating SQL deployment (precheck, deployment, postcheck).
- Integrated **Groq LLM** to generate intelligent feedback and root-cause analysis for failed SQL executions.
- Built **FastAPI-based REST service** for orchestrating execution, logging, and email

notifications with visual screenshots.

- Automated **multi-environment SQL execution pipeline** (DEV/SIT/UAT/PROD) with rollback and error handling.
- Implemented **end-to-end CI-style automation**, reducing manual debugging effort and improving deployment efficiency by 60%.

AI/ML Engineer
DreamsAI Innovations Pvt. Ltd., Mumbai
Sep 2023 – Jul 2024
Project Description:
Designed and deployed OCR pipelines and education-focused chatbots using OpenAI GPT-4 Vision, Groq inference engine, and Azure-based LLMs. Built hybrid search systems powered by LangChain and integrated knowledge graphs.
Tools, Software, Skills:

- LLMs: GPT-4 Vision, Falcon-7B, Mistral
- Frameworks: LangChain, LangGraph
- Vector Stores: Pinecone, Neo4j, FAISS
- Cloud: GCP Vision API, Azure OpenAI
- Deployment: FastAPI, Docker, Kubernetes
- Evaluation: Groq, QLoRA, Cosine Similarity

Role and Responsibilities:

- Designed and implemented a robust system utilizing Google Cloud Vision API, Pytesseract and Azure OpenAI's GPT-4 Vision for precise OCR text extraction from PDFs. Enhanced content accessibility by integrating GPT-4 for generating concise summaries and facilitating interactive Q&A chat functionalities. Deployed Models into production environments, using API model calls FastApi, Containerization tools like Docker and Orchestration platforms like Kubernetes to ensure seamless integration.
- Designed an advanced AI Chatbot for educational assistance. Utilized Cosine Similarity, Hybrid Search and Vector Embedding techniques. Implemented the advance RAG Pipeline, LangChain, Azure OpenAI Embedding Model and Groq a Fast AI Inference framework to enhance functionality.
- Leveraged Pinecone for efficient embedding storage to fetch the contextual contents to improve accuracy of responses. Implemented Reference Link Recommendations for the Chatbot capable of dynamically recommendation of reference links related to user's Query, significantly improving the learning resources' depth and relevance.
- Enhanced Chatbot with Knowledge Graphs DB like neo4j. Implemented Hybrid Search with Keyword, Semantic and Knowledge Graphs Search. Used Cypher Query Language to represent and visualize Graphs. Integrated libraries like Lanchain-community, Langchain-Groq to generate Nodes, Relationships and Properties &

Values between different entities.
- Worked on Falcon-7b and Mistral LLM models by implementing PEFT techiniques like QLoRa, fine- tuned them on specialized datasets to significantly improve performance and efficiency.
- Developed and implemented LangGraph in my project using a Multi-Agent Conversation Framework, enabling the coordination of multiple actors across cyclic computational steps.

Machine Learning Engineer
Instoried Research Labs, Bangalore
Nov 2020 – Mar 2023
Project Description:
Worked on AI applications in content generation, customer support bots, sentiment analysis, and LLM-based document processors using GPT-3 and BERT.
Tools, Software, Skills:

- LLMs: OpenAI GPT-3, BERT
- OCR: OpenAI, Tesseract
- Vector DBs: Pinecone, ChromaDB
- Frameworks: TensorFlow, PyTorch
- Deployment: CI/CD, LangChain

Role and Responsibilities:

- In my diverse experience, I've showcased strong expertise in Machine Learning and NLP. I leveraged Bert Transformers for sentiment analysis, optimized deep learning models with GPU frameworks like TensorFlow and PyTorch, and used various NLP tools.
- I excelled in content generation with GPT-3, fine-tuning for specific styles, and streamlined deployment with CI/CD. I designed customer support chatbots using large language models, enhancing satisfaction through CRM integration.
- In image recognition, I utilized OpenAI GPT3 for text extraction and content generation. Moreover, I built Document-Based AI Chatbots, fine-tuned LLM models, and employed vector databases like Chroma and PineCone for scalable embeddings and fast queries.
- Model I have used is OpenAI GPT3. Tool have the features like extract text from image with the help of OCR(Optical Character Recognition) and then it provide suitable Headline, Ad content, Product Descriptions for Ad Agencies. Contributed to the development of LLM and Langchain frameworks and implemented them in various projects, gaining a strong understanding of these technologies.

Machine Learning Engineer (Intern)
DQ Labs, Bangalore
Jul 2018 – Dec 2018
Project Description:

Developed predictive models to identify high-ranking students based on test data using ML techniques.

Tools, Software, Skills:

- ML Libraries: Scikit-learn, TensorFlow
- Algorithms: Regression, Classification
- Visualization: Matplotlib, Seaborn
- Data Processing: Pandas, NumPy

Role and Responsibilities:

- Built regression and classification models to forecast student performance.
- Preprocessed and cleaned test datasets for model input and validation.
- Designed and validated various ML algorithms for accuracy and bias.
- Used Python libraries for EDA and statistical analysis of learning trends.
- Documented the entire pipeline and collaborated with faculty to align models with education outcomes.

Education

Masters of Computer Application - 2018
VIT University, Chennai

Bachelor of Computer Application – 2016
Naraina Institute Of Technology, Kanpur