Sushant Twayana

Sallaghari, Bhaktapur

sushanttwayana1@gmail.com | +977 9861488545| Linked IN | Github | Blog

AI/ML Engineer specializing in production-grade LLM systems, **MLOps**, and **scalable AI solutions**. Expertise in deploying enterprise applications using LangChain, LangGraph, RAG, Kubernetes, and AWS infrastructure with proven track record of optimizing model performance and reducing latency.

PROFESSIONAL EXPERIENCE

Junior AI/ML Engineer | VANILLA TRANSTECHOR *(July 2025 - Present)*

Core Technical Achievements:

- Architected end-to-end multilingual speech processing pipeline integrating Whisper transcription, custom translation model & Rasa intent classification, improving **ASR** accuracy & reducing latency
- **Fine-tuned Whisper** models using **LoRA adaptation** for domain-specific optimization and benchmarked against AWS speech recognition services
- Working on research and production of **OCR for documents** in fintech domain.
- Collaboratively implemented real-time **fraud detection algorithms** for fintech transaction data using collaborative ML approaches

Associate Data Scientist | DLytica – Data Analytics and AI *(Dec 2024 - Apr 2025)*

- Deployed enterprise **MLOps pipelines** using Kubernetes and Docker with automated CI/CD for model versioning and deployment on **KubeFlow.**
- Built intelligent **conversational AI systems** leveraging LangChain, RAG architecture, vector databases, and autonomous agent frameworks
- Engineered real-**time banking recommendation system** and **customer segmentation solution**.
- Engineered customer segmentation solution processing large-scale data with **PySpark** and **Apache Superset**

Training Assistant: Served as Training Assistant for DLytica Academy's **Data Science programs**, contributing to curriculum development and providing technical mentorship to aspiring data scientists.

AI/ML Engineer Intern | Comfort Yantra *(July 2024 - October 2024)*

- Developed **pre-production-ready conversational AI** using RAG architecture with LangChain, FastAPI, GPT-3.5 Turbo, and Chroma vector store
- Built hybrid automation solutions combining LLM-powered and regex-based methods for e-commerce product categorization and **NL-to-SQL translation**
- Integrated Azure OpenAI and Meta LLaMA-3-8B for semantic parsing and multitool agent architectures.

KEY TECHNICAL PROJECTS HIGHLIGHTS

Production RAG System & Intelligent Chatbots | LangChain, Vector DBs: (GitHub) | (git)

- Built end-to-end RAG systems integrating OpenAI and Ollama models with LCEL, advanced document processing, conversation history management, and persistent memory components for production GenAI applications.

Multi-Agent Blog Generation System (LangGraph): (GitHub)

- Engineered production-grade multi-agent system with autonomous workflow orchestration, state management, and conditional routing for automated multilingual content creation. Implemented custom nodes and graph builders for complex multi-step reasoning.

Autonomous Chatbot with Tool Calling: (GitHub)

- Deployed end-to-end agentic chatbot leveraging tool calling, dynamic decision-making, and autonomous agent architectures

AI-Powered Expense Tracking Application | (MCP) Integration: (GitHub)

- Developed financial management application using MCP for standardized AI-to-system communication with automated budget monitoring and intelligent financial analysis across three architectural patterns.

Distributed Restaurant Order System (FastAPI + gRPC): (GitHub)

- Architected microservices-based system with FastAPI gateway and gRPC inter-service communication. Implemented three distributed services with Protocol Buffers, containerized deployment, and service orchestration

Insurance Premium Prediction API: (GitHub)

- Built FastAPI-based ML inference service with Pydantic validation, Streamlit frontend, and RESTful endpoints with simple Random Forest Model for prediction.

Scalable MNIST Classification Pipeline (Kubeflow + KServe): (GitHub)

- Deployed MNIST classification pipeline with Kubeflow orchestration, KServe for scalable inference, and automated CI/CD workflows for model versioning and deployment.

Fixed Deposit Recommendation System & Customer Segmentation | Banking Domain

- Engineered system using PySpark for distributed processing, ML models with hyperparameter optimization, and Apache Superset for feature engineering and visualization.

Text-to-Face Generation System (DC GAN & DF GAN): (GitHub)

- Implemented generative AI system using PyTorch for human face generation from text descriptions

Food Recognition & Recipe Generation System | ResNet50, YOLOv8, BLIP, T5: (GitHub) | (link)

- Developed end-to-end computer vision pipeline combining ResNet50 and YOLOv8 for Newari food detection, integrated with BLIP Transformer for visual understanding and T5 for automated recipe generation and ingredient extraction.

Newari Food Assistant Chatbot (RAG + Dialogflow): (GitHub)

- Architected production-grade chatbot using RAG, LangChain, vector databases, and autonomous agents and integrated dialogflow for natural conversation and intent recognition

Content-Based Movie Recommender: (GitHub)

- Built recommendation system using NLP embeddings and cosine similarity algorithms

LEADERSHIP & ACHIEVEMENTS

- **Winner:** ProtoBytes Hackathon at ACEM 2024: Developed innovative AI solution under competitive deadline
- **Best Presenter Award -** KU IT FEST 2024
- **Event Organizer -** Hult Prize Event 2022/23: Led team in executing large-scale entrepreneurship event
- **Associate** - ITSNP: Active contributor to technical and community-building initiatives (2022)

EDUCATION

Bachelor's in computer engineering, **Khwopa Engineering College |** ***(Oct 2019 – Aug 2024)***
CGPA: 3.50/4.0

TECHNICAL SKILLS

- **LLM & GenAI**: LangChain, LangGraph, RAG, OpenAI GPT, Ollama, Vector Databases (Chroma, Pinecone), MCP
- **MLOps & Cloud**: Kubernetes, Docker, Kubeflow, KServe, AWS EC2, CI/CD Pipelines
- **ML/DL Frameworks**: PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, PySpark
- **Specialized Models**: Whisper, Rasa, YOLO, BERT, T5, BLIP, GANs
- **Development**: Python, FastAPI, gRPC, SQL, Git

CERTIFICATIONS

- **Complete Agentic AI Bootcamp With LangGraph and Langchain**
- **Associate Data Scientist in Python Certification** (DataCamp)
- **PyTorch and Keras (TensorFlow) Full Course**
- Advanced Python and Data Science Certifications (DataCamp)