AWANTIKA SRIVASTAVA

Gen AI Engineer | Generative AI Developer | LLMs | RAG | Machine Learning

+91-8920482037 | sawantika81@gmail.com | LINKEDIN | Github

PROFILE SUMMARY

Generative AI Engineer with **2+ years of hands-on experience** in **Generative AI, Large Language Models (LLMs), Retrieval Augmented Generation (RAG), Model Context Protocol (MCP) concepts, AI orchestration frameworks, and backend AI development**. Experienced in **designing, integrating, deploying, and managing end-to-end AI systems**, including **LLM-based services, MCP-style model context management, tool-based workflows, and production-grade AI applications**. Strong ability to bridge **AI capabilities with scalable backend systems**.

CORE TECHNICAL SKILLS

- **Programming & Data Analysis:** Python, C++, SQL, Flask, FastAPI, RESTful APIs, Jupyter Notebook
- **Statistics & Mathematics:** EDA, Statistical Modeling, Hypothesis Testing, Probability, Linear Optimization, Trend Analysis.
- **Machine Learning:** Supervised & Unsupervised Learning, Regression, Classification, Clustering, Random Forest, Decision Trees, SVM, KNN, K-Means, XGBoost, Hyperparameter Tuning.
- **Deep Learning:** Neural Networks, CNN, RNN, LSTM, BERT, Transfer Learning, TensorFlow, PyTorch
- **GenAI & LLMs:** LLMs, Generative AI, RAG, Prompt Engineering, Structured Outputs, Multi-step Reasoning, Context-aware Generation, Model Context Management, MCP (Model Context Protocol) Concepts, Tool-based LLM Workflows
- **LLM Orchestration & Agents:** LangChain, LangGraph , LLM Orchestration Pipelines, Agentic Workflows, Tool Calling, Context Injection
- **AI System Development:** End-to-End AI Pipelines, AI Model Integration, Production-grade AI Systems, Scalable AI Applications, AI Automation Workflows
- **Containerization & Deployment:** Docker, Containerized AI Deployments, Model Serving, Deployment Pipelines, AWS EC2, Production Deployment
- **Computer Vision:** Image Classification, Object Detection, OpenCV, YOLO, MobileNet, Image-based Decision Systems

EXPERIENCE

AI/ML Engineer | PPS International Pvt. Ltd. **January 2024-Present**

- Designed and deployed **end-to-end machine learning pipelines** covering data ingestion, preprocessing, training, evaluation, and production deployment.
- Built and optimized **ML models using Python, scikit-learn, TensorFlow, and PyTorch** for real-world analytical and predictive use cases.
- Implemented **LLM-based workflows** including prompt engineering, structured outputs, and **retrieval-augmented generation (RAG)** pipelines.
- Developed and evaluated **LLM-as-judge frameworks** using GenAI metrics such as **Recall@K, MRR, Faithfulness, and F1-score**.
- Performed **large-scale data analysis and EDA** to identify trends, anomalies, and data quality issues prior to modeling.
- Developed **RESTful ML inference services using FastAPI** for integration with downstream systems.
- Deployed and validated ML and GenAI solutions on **AWS**, ensuring scalability, reliability, and performance.
- Applied **MLOps best practices**, including experiment tracking (MLflow, W&B), model versioning, monitoring, and cross-functional collaboration.

PROJECTS

Railway Driver Assistance System (RDAS) | Enterprise ML Project

- Built and deployed a **production-grade machine learning system** to detect unsafe driver behavior, translating real-world safety requirements into AI-driven solutions.
- Designed and owned **end-to-end ML pipelines** covering data ingestion, preprocessing, feature engineering, model training, evaluation, and inference for real-world deployment.
- Trained **CNN-based computer vision models (SSD MobileNet)** using **TensorFlow**, and optimized inference using **TensorFlow Lite** for low-latency environments.
- Achieved **real-time inference performance (20–25 FPS, <150 ms latency)**, enabling near real-time decision-making in operational settings.
- Integrated model predictions into **downstream systems via REST APIs**, delivering **actionable insights** to support operational and safety-related decision

Chatbot using LLM & RAG | Applied ML Project

- Built a **production-grade Generative AI (GenAI) application** using **Large Language Models (LLMs)** and **Retrieval-Augmented Generation (RAG)** for context-aware question answering.
- Designed and implemented **document ingestion, embedding generation, vector retrieval, and LLM inference orchestration pipelines**.
- Applied **prompt engineering, response evaluation, and reliability tuning** to improve output accuracy, consistency, and relevance.
- Deployed the solution as an **observable LLM inference service**, aligning with modern **GenAI and LLM production workflows**.

YouTube Comments Sentiment Analyzer | link - https://youtube-ai-analyzer-ndzqo6r2mepjrtsdjmwaxl.streamlit.app/

- Developed an **end-to-end NLP modeling system** using **BERT / DistilBERT** for large-scale sentiment classification.
- Performed **EDA, text preprocessing, feature engineering, model training, and statistical evaluation** on real-world user-generated data.
- Conducted **baseline comparisons and experiment tracking** to optimize model performance.
- Deployed the model as a **REST API-based inference service** to deliver actionable sentiment insights to business stakeholders.

CERTIFICATION

- IBM Data Science & AI Certification
- AWS Generative AI with Large Language Models
- OpenCV Computer Vision Certification

EDUCATION

IMS Engineering College, Ghaziabad **September - 2020**
Bachelor of Technology (Electrical and electronics engineering)