AWANTIKA SRIVASTAVA

AI/ML Engineer | LLMs & AWS

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

PROFILE SUMMARY

AI/ML Engineer with 2+ years of experience building scalable Generative AI systems using LLMs, RAG architectures, and agentic workflows. Hands-on expertise in multi-agent orchestration, tool-calling patterns, embeddings, vector search, and backend microservices development. Experienced in AWS cloud services and deploying production-grade AI systems using Docker and CI/CD pipelines. Strong background in Python development, retrieval pipelines, and performance optimization.

CORE TECHNICAL SKILLS

- **Programming & Backend:** Python, C++, SQL., FastAPI, Flask, REST APIs, Microservices Architecture
- **Statistics & Mathematics:** EDA, Statistical Modeling, Descriptive & Inferential Statistics, Bayesian Statistics, Hypothesis Testing, Probability, A/B Testing
- **Machine Learning:** Supervised & Unsupervised Learning, Regression, Classification, Clustering, Random Forest, Decision Trees, SVM, KNN, K-Means, XGBoost, LightGBM, Model Evaluation Metrics (Accuracy, Precision, Recall, F1-score, ROC-AUC).
- **Deep Learning:** ANN, CNN, RNN, LSTM, BERT, Transfer Learning, Self-Attention, multi-head Attention, Tensorflow, PyTorch, Keras
- **NLP:** Text preprocessing, Tokenization, Chunking, Sentiment Analysis, Transformer-based models, NLTK, TextBlob, SpaCy, Time-Series Forecasting, ARIMA, SARIMA
- **Generative AI & LLMs:** RAG, Multi-Agent Orchestration, Tool-Calling Workflows, Agent Reasoning & State Management, Embeddings & Vector Search, Prompt Engineering, LLM APIs (OpenAI / Claude – API based)
- **Computer Vision:** Image Classification & Preprocessing, Object detection (YOLO, SSD, MobileNet, ResNet), Video Analytics, OpenCV
- **Frameworks & Tools:** LangChain, LangGraph, Hugging Face APIs, FAISS / Chroma, Git, GitHub, Docker, Streamlit
- **MLOps & Deployment:** MLflow, REST API, AWS (EC2, S3), CI/CD Pipelines, DynamoDB,

EXPERIENCE

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

- Developed **predictive modeling solutions** using regression, decision trees, random forest, gradient boosting, and deep learning models for datasets.
- Built **scalable ML pipelines** using PySpark and Spark SQL for large-scale distributed data processing and model training.
- Designed and optimized **classifiers using feature selection, hyperparameter tuning, and ensemble methods**, improving model performance by 15–20%.
- Applied **Bayesian statistics and advanced evaluation metrics** to enhance model robustness and decision accuracy.
- Built **LLM-powered search and automation systems** using **Retrieval-Augmented Generation (RAG)** pipelines with embedding-based semantic retrieval.
- Fine-tuned **LLM models using PEFT/LoRA techniques** for domain-specific query understanding and improved contextual response generation.
- Designed and implemented **reranker models (cross-encoder based reranker design)** to improve search relevance and ranking quality.
- Integrated **Elasticsearch / OpenSearch** for semantic indexing, hybrid search, and vector search capabilities.
- Developed **scalable data systems in collaboration with product and engineering teams** to enhance user experience and search quality.
- Managed model lifecycle using **MLflow**, and deployed **real-time inference APIs on AWS**, ensuring low-latency production systems.

PROJECTS

Railway Driver Assistance System (RDAS) | Enterprise ML Project

- **Designed and deployed a** real-time computer vision–based ML system **for unsafe** driver behavior detection using **CNN-based SSD MobileNet** model.
- Trained and optimized models on large-scale video datasets, achieving **20–25 FPS** real-time processing with **<150 ms inference latency**.
- Implemented **end-to-end ML pipelines** for data ingestion, preprocessing, model training, evaluation, and production inference.
- Deployed optimized models using **TensorFlow Lite** on edge/production environments for continuous monitoring.
- Built a **Flask-based web dashboard** to visualize detections and automatically record **30-second event clips**, reducing manual review effort.

Chatbot Using LLM & RAG | Applied ML Project

- Built a Lightweight **LLM-Powered chatbot** using **TinyLLaMA** to answer user queries over content.
- Implemented a **Retrieval-Augmented Generation (RAG) pipeline** to retrieve relevant resume sections for contextual question answering.
- Selected **TinyLLaMA** to ensure **low memory footprint and fast inference**, making the solution suitable for resource-constrained environments.
- Applied **prompt engineering techniques** to improve response relevance and consistency.
- Deployed the chatbot as an interactive **Streamlit web application** for real-time user interaction.

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

- Developed **LSTM-based time-series forecasting** model trained on 5+ years historical stock data reducing **RMSE** by 18**%.**
- Engineered lag **features**, **rolling statistics**, and volatility indicators improving prediction stability.
- **Implemented** custom **BPTT training** loop with adaptive learning rate scheduling for **stable convergence**.
- Applied **regularization techniques** reducing overfitting gap by **20**%.

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