

AWANTIKA SRIVASTAVA

AI / Machine Learning Engineer | LLM | Generative AI | Chatbot

+91-8920482037 | sawantika81@gmail.com | [LINKEDIN](#) | [Github](#)

PROFILE SUMMARY

AI Engineer with **2+ years of experience** architecting, developing, and deploying **scalable, resilient, and ethical AI solutions**. Strong hands-on expertise in **Large Language Models (LLMs)**, **Generative AI**, **open-source frameworks**, **end-to-end AI pipelines**, and **automation**. Experienced in building **LLM integrations**, **RAG-based systems**, **guardrails**, **backend APIs**, and **enterprise-ready ML workflows**. Proven ability to collaborate with **product**, **data science**, **UX**, and **engineering teams** while adhering to **Responsible AI principles**, **security**, and **performance standards**.

CORE TECHNICAL SKILLS

- **Programming & Data Science:** Python (Pandas, NumPy, Scikit-learn), C++, REST APIs, FastAPIs, Flask, Git, GitHub
- **Statistics & Mathematics:** Statistical Modeling, Hypothesis Testing, Confidence Intervals, A/B Testing, Bias & Variance Analysis, Scenario Analysis.
- **Machine Learning:** Supervised & Unsupervised Learning, Regression, Classification, Clustering, Random Forest, Decision Trees, SVM, KNN, K-Means, XGBoost, Model Evaluation Metrics, Hyperparameter tuning, Model Evaluation
- **Deep Learning :** Neural Networks, CNN, RNN, LSTM, Transformers (BERT), Transfer Learning, Model Fine-tuning , PyTorch, TensorFlow, Keras
- **Computer Vision:** Image Classification & Preprocessing, Object detection (YOLO, SSD, MobileNet, ResNet), Video Analytics, OpenCV
- **NLP:** Text preprocessing, Tokenization, Chunkings, Sentiment Analysis, Transformer-based models, NLTK, Spacy
- **Generative AI & LLMs:** LLMs, OpenAI APIs, Hugging Face, LangChain, Prompt Engineering, Context Management, LLM Integration
- **Automation & Testing:** Selenium, Automated Test Suites, UI & API Testing, Workflow Automation, Reliability Testing
- **Cloud & Enterprise Platforms:** AWS EC2, AWS SageMaker, AWS Lambda, AWS ECR, Model Deployment, Secure API Integration
- **Cloud & MLOps:** AWS EC2, AWS SageMaker, AWS Lambda, AWS ECR, Model Deployment, ML Pipelines, Experiment Tracking

EXPERIENCE

AI/ML Engineer | PPS International Pvt. Ltd.

January 2024-Present

- Architected and delivered **end-to-end AI solutions**, covering data preprocessing, model training, evaluation, and deployment.
- Developed **scalable ML and LLM-enabled backend services** using Python and REST APIs.
- Built and optimized **AI pipelines** to ensure reliability, performance, and maintainability.
- Integrated **automation and testing workflows** (Selenium-based) to validate AI-driven features.
- Worked with **large datasets** while ensuring data integrity and consistency.
- Collaborated with **cross-functional teams** (product, UX, engineering) to gather requirements and deliver enterprise-ready solutions.
- Followed **secure coding practices, documentation standards, and Responsible AI principles**.
- Maintained technical documentation and followed **software development best practices**.

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/>

- Deployed transformer-based **NLP models** as **production-ready** services with **REST APIs**.
- **Fine-tuned** and served a **DistilBERT**-based **sentiment classification model** for large-scale text inference.
- **Built** and deployed an interactive **streamlit web application** to perform **real-time sentiment analysis** on YouTube comments.
- **Processed** high-volume text **data** with sub-second inference **latency** for real-time sentiment analysis.

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)