

AWANTIKA SRIVASTAVA

AI Engineer | AI & Machine Learning Solution

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

PROFILE SUMMARY

AI Engineer with **2 years of experience** in developing, deploying, and operationalizing **machine learning and AI solutions**. Strong foundation in **machine learning, deep learning, and statistical analysis**, with hands-on experience across the **end-to-end ML lifecycle**—from problem understanding and data preparation to model development, evaluation, and deployment. Proven ability to support **proof-of-concepts (PoCs)**, collaborate with cross-functional teams, and deliver scalable ML solutions aligned with business objectives.

CORE TECHNICAL SKILLS

- **Programming Languages:** Python, Numpy, Pandas, C++, SQL.
- **Statistics & Mathematics:** EDA, Statistical Modeling, Descriptive Statistics, Hypothesis Testing, Probability, Sampling, Scenario Analysis.
- **Machine Learning:** Supervised & Unsupervised Learning, Regression, Classification, Clustering, Random Forest, Decision Trees, SVM, KNN, K-Means, XGBoost, Hyperparameter Tuning.
- **Deep Learning & AI:** Neural Networks, CNN, RNN, LSTM, Transformers (BERT), Transfer Learning, Model Fine-tuning, Time-Series forecasting, ARIMA, SRIMA, TensorFlow, Keras, PyTorch
- **Generative AI & LLMs:** LLMs (GPT, LLaMA), RAG, Prompt Engineering & Prompt Strategies, LLM Fine-Tuning (PEFT/LoRA), Embedding-based Semantic Search, Reranker Design, LLM Evaluation & Guardrails
- **Computer Vision:** Image Classification & Preprocessing, Object detection (YOLO, SSD, MobileNet, ResNet), Video Analytics, OpenCV
- **NLP:** Text preprocessing, Tokenization, chunking, Sentiment Analysis, Topic Modeling (LSA, LDA), Transformer-based models.
- **Azure & Cloud:** Azure OpenAI, Azure Cognitive Services, Azure DevOps (Sprint Planning), AWS EC2, S3, Cloud Deployment of LLM Applications
- **MLOps & Governance:** MLflow, Model Monitoring, Risk Mitigation & Deployment Planning, CI/CD Pipelines, Project Estimation & Sprint Planning
- **Tools:** LangChain, LangGraph, PyTorch, Docker, Git

EXPERIENCE

AI/ML Engineer | PPS International Pvt. Ltd.

January 2024-Present

- Designed and deployed **Generative AI and LLM-powered applications** aligned with business objectives.
- Implemented **end-to-end RAG pipelines** using embedding-based semantic retrieval and reranker logic.
- Fine-tuned transformer-based LLM models using **PEFT/LoRA techniques**.
- Integrated **OpenAI-based APIs** and cloud-hosted **LLM services** into production systems.
- Developed **Proof of Concepts (PoCs)** and demonstrated GenAI solutions to stakeholders.
- Worked in Agile environment using sprint planning, task estimation, and **Jira-based tracking**.
- Ensured risk mitigation and **performance monitoring** during **deployment cycles**.
- Collaborated with cross-functional teams for solution architecture and client engagement.

PROJECTS

Railway Driver Assistance System (RDAS) | Enterprise ML Project

- Designed and deployed a real-time computer vision-based ML system to detect unsafe driver behaviors from continuous video streams.
- Trained and optimized CNN-based object detection models (SSD MobileNet architecture) to perform real-time inference on video data.
- Implemented end-to-end ML pipelines covering data ingestion, preprocessing, model training, evaluation, and production inference.
- Achieved 20–25 FPS real-time processing with <150 ms inference latency by optimizing models for deployment.
- Built and integrated a **Flask-based web interface** to visualize detections and automatically recorded **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-ndzqo6r2mepjrtsdjmwxal.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

Bachelor of Technology (Electrical and electronics engineering)

September - 2020