Animesh Kumar Singh

Github: animeshkr7

Email: animeshkrsc7@gmail.com
Mobile: +91-7462080100
LinkedIn: animeshkr7

Work Experience

- **Facctum** Pune
  Associate ML Engineer *Sept 2025 - Present*
  - **Sanctions Press Release Extraction System**: Integrated large language models to extract critical regulatory entities from complex HTML/PDF press releases, mapping raw data to a Common Data Format. Delivered an **audit-ready**, delta-tracking compliance system that improved data extraction accuracy by **83%** and accelerated downstream regulatory **SLAs** targets to Banks by **94%** .
  - **Trade Goods Screener**: Engineered an end-to-end intelligent Trade Goods Screener using **LangGraph** and **Bedrock's Claude Opus Model**, enabling HS-code classification, sanctions/license screening, risk scoring, and case routing to compliance officers. Built an orchestration layer with role-based memory and tool integrations, reducing manual review time by **70%** and removing coordination bottlenecks across trade compliance operations.
- **Volkswagen (VGDS - India)** Pune
  SDE Intern *Jan 2025 - July 2025*
  - **Agentic Service Center Automation** : Designed an end-to-end intelligent workflow for automotive service centers using **CrewAI** and **Bedrock's Llama-3B**, enabling autonomous task routing from UI intake to technician assignment, inventory checks, real-time communication, and payment processing. Implemented multi-agent orchestration with role-based memory using **tool-augmented agents**, reducing service resolution time by **45%** and eliminating manual coordination bottlenecks across operations.
  - **Hybrid Chatbot System for In-Drive Personalization**: Engineered a hybrid recommendation model integrating **LLama 3.1 LLM** with a dynamic **RAG** retrieval mechanism for real-time, in-vehicle assistance. The system encodes longitudinal user preference data (k-past interactions) into semantic vectors for retrieval, providing rich, contextual grounding for the generative model. Model efficacy was validated by a significant reduction in perplexity and almost **28%** improvement in Mean Reciprocal Rank **(MRR)** for Point-of-Interest suggestions, proving high-fidelity predictive accuracy.
- **Medecro AI**
  Machine Learning Engineer Intern *Nov 2024 - Jan 2025*
  - **Dental YOLOv8 Object Detector**: Developed a **YOLOv8**-based object detection model on Medecro.ai's annotated dental X-ray set (10k+ images) using Mosaic & CutMix enhancements. achieving **mAP@0.5 0.92**, and **25%** model size reduction via structured pruning & **8-bit quantization**, deployed as endpoint for real-time chairside diagnostics.

Projects

- **AI-Driven Cold Mail Generator**: Architected a generative outreach engine using Llama & **ChromaDB** with hybrid dense/sparse retrieval and FAISS indexing, **LoRA** fine-tuned LLMs, and optimized query pipelines for <100ms latency cutting email drafting time by **59%+** and driving a **30%+** lift in response rates.
- **LLaMA8B Domain-Adaptation Pipeline**: Built a Colab-optimized fine-tuning setup for **LLaMA-3 8B** with 4-bit bitsandbytes quantization, **LoRA r=16 adapters**, and gradient checkpointing. Trained 10 epochs on 2048-token sequences (batch size 8), achieving 300tokens/sec throughput, **85%** GPU memory savings, and <50ms inference latency
- **Credit Risk Assessment Engine**: Engineered an end-to-end credit risk prediction pipeline in Python, leveraging **Decision Tree, Random Forest, and XGBoost** classifiers on 10K+ customer profiles. Designed advanced feature engineering with financial ratio analysis, missing-value imputation, and hyperparameter optimization, achieving a **5%** lift in accuracy over industry baselines and enabling robust, real-time lending decisions.

Education

- **Bachelor of Technology - Computer Science and Engineering** 2021- 2025
  Silicon University (CGPA - 8.02) *Bhubaneswar, India*

Technical Skills

- **Core Competencies**: Data-Science, Machine-Learning, Agentic AI, Deep Learning, Computer Vision, Python, MLOps
- **Frameworks & Libraries**: TensorFlow, PyTorch, Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn, Keras, XGBoost, Random forest, LightGBM, FastAPI, Flask, Streamlit
- **AI & LLM Technologies**: OpenAI, Claude, Gemini, LLAMA2, BERT, Transformers, Langchain, HuggingFace, FLux, Mistral, Deepseek, Crewai, LangFlow
- **Vector Databases & Search**: FAISS, Pinecone, Weaviate
- **Data Engineering & Big Data**: PostgreSQL, MongoDB, MySQL, Oracle
- **Cloud & DevOps**: AWS (EC2, S3, Bedrock, Sagemaker, Lambda)
- **Software Engineering**: Git Version control, Docker

Achievements & Certifications

- **Competition Performance**: Volkswagen imobilothon 4.0 - TOP 5