Rishabh Bipin Shukla

Mumbai, India
rishabhshukla092@gmail.com | linkedin.com/in/rishabhbshukla

Professional Summary

Associate Data Scientist with hands-on experience in **Generative AI, Large Language Models (LLMs), and NLP systems**. Proven ability to architect, fine-tune, evaluate, and deploy AI solutions including **Retrieval-Augmented Generation (RAG)**, text analytics pipelines, and scalable inference APIs. Strong grounding in **machine learning workflows, statistics, and model performance optimization**. Experienced in cross-functional collaboration with product, data, and engineering teams. Currently pursuing advanced academic research in deep learning and real-time data processing.

Core Technical Skills

Generative AI & NLP: LLMs (LLaMA, Mistral, Phi-3.5), GPT-style models, BERT, T5, Prompt Engineering, Text Generation, Text Classification, Named Entity Recognition (NER), Semantic Search, Embeddings
Machine Learning: Supervised & Unsupervised Learning, Feature Engineering, Model Fine-Tuning, Bias Evaluation, Inference Optimization, Validation Metrics
Frameworks & Tools: PyTorch, Hugging Face Transformers, FastAPI (Async), TensorFlow (foundational), Matplotlib, Seaborn
Data & Platforms: Python, SQL, NoSQL fundamentals, Docker, AWS Lambda, REST APIs

Applied AI & Data Science Projects

Deep Learning for Real-Time Music Visualization (MCA Thesis)
Designed a deep learning system converting high-dimensional acoustic features into real-time visual representations. Implemented CNN-based architectures for audio feature extraction and temporal pattern recognition. Optimized inference latency and throughput using asynchronous FastAPI pipelines for real-time processing.

Local Privacy-First Knowledge Engine (RAG)
Architected an offline document question-answering system using Phi-3.5 LLM. Built complete NLP workflows including document ingestion, text preprocessing, chunking, embedding generation, and vector similarity search. Improved factual accuracy and reduced hallucinations using context-aware RAG and prompt optimization.

Charity Suno – AI-Powered Study Assistant
Developed an LLM-driven conversational agent supporting academic reasoning, content generation, and contextual Q&A. Integrated Gemini API and open-source models. Deployed scalable AI-backed APIs enabling real-time text and speech-based interactions.

Professional Experience

Freelance Application Developer

- Delivered multiple production-grade applications incorporating AI-enabled features and data-driven backends.
- Designed, deployed, and scaled RESTful APIs and microservices using FastAPI, Docker, and AWS Lambda.
- Collaborated with cross-functional stakeholders to translate business requirements into technical AI solutions.
- Ensured error-free deliverables through model validation, backend testing, monitoring, and performance tuning.

Education

Master of Computer Applications (MCA)
MET Institute of Computer Science, Mumbai

Bachelor of Computer Applications (BCA)
K. P. B. Hinduja College, Mumbai