# Soumyabrata Banerjee

soumyajee12345@gmail.com | +91-6290551330 | linkedin.com/in/soumyabratabanerjee1234 | github.com/soumyajee

## Summary

AI/ML Engineer with 3+ years of experience building Generative AI applications, LLM fine-tuning, and scalable ML pipelines. Skilled in RAG, vector databases, computer vision, and deploying deep learning models on cloud (AWS, GCP, Azure). Currently working on AI Agents with FastAPI, Redis, PostgreSQL, and Gemma-9b-it model integration.

## Experience

**AI Engineer** Jul 2025 – Present
Barinium Infotech Pvt Ltd

- Designed and implemented **Content Sourcing Agent** and **Assessment Agent** pipelines for automated content generation and evaluation.
- Built scalable backend services using **FastAPI server**, integrated with **Redis CLI** for event handling and **PostgreSQL** for persistent storage.
- Integrated AI agent workflows with the **Gemma-9b-it model from Groq**, improving response quality and reducing latency for inference.
- Contributed to modular, production-ready architecture enabling faster development cycles and robust agent orchestration.

**AI/ML Developer** May 2022 – Jul 2025
CDAC Hyderabad, Telangana

- Developed offline desktop app (Python, OpenCV) for deepfake detection, reducing inference latency by 40%.
- Trained ViT and MC3-18 achieving 98% accuracy on violence detection (UBI Fights dataset).
- Built Django-based *DeepfakeCheck.in* portal used by 500+ testers to detect deepfake images, videos, audio.
- Benchmarked models achieving 97% (UBI Fights) and 89% cross-dataset validation.

**AI Full Stack Developer** Nov 2021 – Apr 2022
Graylogic Technologies, Hyderabad

- Developed full-stack image processing apps with DL models for classification.
- Optimized FastAPI backend, reducing model inference response time by 45%.
- Collaborated with cross-functional teams to deploy AI-driven product features.

**Data Science Intern** Oct 2021 – Jul 2022
AI Varient, Bengaluru

- Built real-time ML projects deployable via Streamlit and cloud platforms.
- Designed scalable ML pipelines with Streamlit APIs for seamless integration.

## Projects

**AI-Powered Educational Platform (AI Agents, FastAPI, Redis, PostgreSQL)** 2025

- Designed and implemented an educational platform powered by multiple AI Agents: **Assessment Agent**, **Feedback Agent**, and **Content Sourcing Agent**.
- Integrated backend using **FastAPI server**, **Redis CLI**, and **PostgreSQL**, enabling scalable agent orchestration.
- Leveraged **Gemma-9b-it model from Groq** for AI reasoning and content generation.

- Enhanced learning experience by automating assessment generation, feedback delivery, and curriculum-aligned content sourcing.

**AI Weather & RAG Agent (GPT-3.5, LangChain, Qdrant)** Sep 2024

- Built Streamlit UI integrating OpenWeather API + RAG for PDF Q&A, embeddings stored in Qdrant, tested with LangGraph/LangSmith.

**End-to-End Medical Chatbot (RAG, GPT-3.5, Pinecone)** May 2023

- Developed diabetic chatbot using RAG, deployed with CI/CD on AWS EC2, ECR, Docker.

**Disastrous Tweets Classification (AWS, Docker)** Jun 2023

- Built NLP pipeline for tweet classification with Streamlit API, deployed on AWS with CI/CD.

**FakeCheck Deepfake App (PyTorch, OpenCV)** Mar 2023

- Offline deepfake detection app packaged with PyInstaller; leveraged EfficientNetAutoAttB4 for heatmap-based video analysis.

## Technical Skills

**Languages:** Python, SQL, R
**ML/DL:** PyTorch, TensorFlow, Scikit-learn, ViT, BERT, T5, DistilBERT
**GenAI:** LangChain, LangGraph, Hugging Face, LlamaIndex, Fine-tuning, Prompt Engineering, RAG, Azure OpenAI
**Vector DBs:** Pinecone, Qdrant, Faiss, Chroma
**MLOps/Deployment:** Docker, Kubernetes, CI/CD, DVC, FastAPI, Streamlit, AWS, GCP
**Core CS:** DSA, OOP, Algorithms

## Education

**B.E., Electrical and Electronics Engineering** Jul 2012 – Apr 2016
Anna University, Chennai CGPA: 8.3/10