Anshul Pandey

☎ 74669-97889 — ✉ anshulpandey0077@gmail.com — linkedin.com/in/anshulpandeyyy — github.com/anshul-20

Summary — Jr. Data Scientist specializing in applied Generative AI and LLM systems for newsroom automation. Experienced in building RAG pipelines, embedding-based similarity systems, and GPU-optimized open-source model deployments. Focused on scalable AI infrastructure, cost-efficient inference, and production-ready AI applications.

Technical Skills

Programming: Python, SQL

ML/DL: Scikit-learn, TensorFlow, Pandas, NumPy

Generative AI: Hugging Face, LangChain, LangGraph, RAG, AI Agents, Ollama

Vector DB: FAISS, ChromaDB

Deployment: REST APIs, GPU Deployment, Linux

Tools: Git, Jupyter Notebook, Postman

Education

Chandigarh University **2023 - 2025**
MCA – Artificial Intelligence and Machine Learning
Minors: Deep learning

MJP Rohilkhand University **2019 - 2022**
Bachelors of Computer Applications
Minors: Web Development

Projects

Hindi Voice Cloning System

- Deployed open-source Chatterbox voice cloning model locally on GPU infrastructure to ensure full data privacy and zero API dependency
- Optimized large-model inference using quantization and efficient GPU memory handling to enable real-time synthetic speech generation
- Managed KV cache and token streaming strategies to improve generation latency for long audio synthesis tasks
- Engineered pipeline integrating text preprocessing, phoneme alignment, and model inference for high-quality Hindi speech output
- Reduced voice generation cost to near-zero compared to commercial TTS APIs

Feed Gap Analysis Engine

- Built automated RSS ingestion and clustering pipeline processing multi-source news feeds for topic-level similarity detection
- Generated dense embeddings for competitor and internal articles and indexed them using FAISS for high-speed semantic similarity search
- Designed coverage scoring framework to identify under-covered topics using threshold-based embedding similarity matching
- Optimized embedding and similarity workloads using GPU acceleration and batch processing to reduce processing time for large news datasets
- Built evaluation metrics to benchmark clustering quality and topic coherence across newsrooms

Work Experience

Amar Ujala Web Services **August 2025 - Present**
Jr. Data Scientist

- Designed and deployed LLM-powered meta-content generation pipelines for automated news summaries, headlines, and structured fact extraction using open-source models
- Built embedding-based feed gap analysis system by clustering competitor RSS feeds and identifying under-covered topics using similarity search
- Optimized inference cost by experimenting with quantized models (4-bit/8-bit), KTransformers, and GPU batching, reducing infrastructure dependency on paid APIs
- Engineered scalable RSS ingestion and clustering pipelines handling large-scale news streams with automated topic grouping and similarity matching
- Researched and implemented Hindi voice cloning workflows (Chatterbox + open-source TTS models) enabling low-cost synthetic media generation with full data privacy