Jayant Kothari

AI/ML Engineer | NLP | Generative AI
+91-6367413575 | kotharijayant73@gmail.com | linkedin.com/in/jayant-kothari | github.com/jay12676 | leetcode.com/u/jayantkothariabcd

Education

Indian Institute of Information Technology, Kota — Kota, India
B.Tech in Computer Science and Engineering, GPA: 8.15/10 — *2023 – 2027*

Experience

Machine Learning Intern — Jan 2026 – Mar 2026
Gyanama — *Remote*

- Developed an end-to-end analytics platform using XGBoost to predict academic performance, attendance patterns, and institutional health scores across 15+ partner schools, enabling data-driven intervention decisions
- Constructed feature pipelines over 10,000+ records (marks, attendance, grades, demographics) using Pandas and Scikit-learn, achieving 87% classification accuracy and 0.91 ROC-AUC
- Designed automated evaluation dashboards aggregating per-learner and per-institution metrics, reducing manual analysis effort by 60% and cutting report turnaround from 3 days to 4 hours
- Deployed model inference as FastAPI microservices integrated with Gyanama's backend, serving real-time predictions with p95 latency under 200ms

Projects

DocuMind – Multi-PDF RAG System | *Python, FastAPI, LangGraph, FAISS, Celery, Docker* GitHub | Live

- Architected an asynchronous multi-stage RAG pipeline (FastAPI + Celery + LangGraph) with 4-level hierarchical chunking and hybrid retrieval (FAISS + BM25 + reciprocal rank fusion), handling 1,000+ page PDFs with zero memory spikes
- Implemented 5 LangGraph reasoning workflows – contradiction detection, knowledge-gap discovery, question generation, timeline evolution, and cross-document insight synthesis – with Redis TTL caching that cut redundant LLM calls by 40% and query latency by 55%
- Containerized the full stack via Docker Compose across 5 services (API, workers, frontend, cache, database), adding startup reconciliation, health checks, and SSE streaming to deliver tokens in under 800ms

Credit Card Fraud Detection System | *Python, XGBoost, TensorFlow, SHAP, FastAPI, Docker* GitHub | Live

- Built a real-time fraud-decisioning microservice fusing a supervised XGBoost classifier (PR-AUC 0.88, ROC-AUC 0.98) with an unsupervised deep autoencoder to catch both known and novel/zero-day fraud across 284,807 transactions
- Designed a 3-tier graded decision engine (APPROVE / VERIFY_OTP / BLOCK) with a simulated OTP challenge for medium-risk transactions, separating risk decisioning from payment enforcement
- Integrated SHAP feature attribution with a generative-AI layer (Gemini 2.5 Flash-Lite, template fallback) to explain every decision in plain English, returning predictions in ∼1.25s
- Served via a Dockerized FastAPI endpoint (/predict, /predict/batch) with a live Streamlit dashboard featuring a color-coded risk gauge and dual-model contribution view; tested with pytest for leakage, ensemble weighting, and tier boundaries

YouTube Smart Chatbot | *Python, Streamlit, FastAPI, Groq Llama 3.3 70B, Deepgram, yt-dlp* GitHub

- Built a multilingual conversational RAG system over YouTube transcripts with timestamp-synced captions, supporting videos up to 12+ hours via parallel 300-second Deepgram Nova-2 chunks
- Engineered a dual-source ingestion flow (native captions with speech-to-text fallback) preserving word-level timing across 20+ languages, improving transcript coverage by 35% on low-resource videos
- Optimized prompt templates with temperature tuning (0.3–0.4) enforcing grounded responses within video context, reaching sub-3-second answers via Groq inference with 10-turn stateful memory

Technical Skills

Languages: Python, C++, SQL, Java
AI/ML: Generative AI, NLP, LLM Fine-tuning, RAG, Prompt Engineering, Deep Learning, XGBoost
Frameworks: LangChain, LangGraph, Hugging Face, Sentence-Transformers, PyTorch, Scikit-learn, Pandas, NumPy
Retrieval: FAISS, BM25, Cross-Encoder Re-ranking, Hybrid Search, Embeddings
Backend & Databases: FastAPI, Streamlit, Gradio, REST APIs, Celery, Redis, PostgreSQL, SQLite, SQLAlchemy
DevOps & Tools: Git, GitHub, Docker, Jupyter, Google Colab
Fundamentals: DSA, OOP, System Design

Achievements

- Solved 600+ DSA problems on LeetCode
- 3-star rating on CodeChef