

Manthan Hirani

AI/ML Engineer | Agentic AI | LLM | IIITV M.Tech Silver Medalist
Junagadh, Gujarat, India — mnhirani212@gmail.com — +91 9898484241
GitHub — LinkedIn — Portfolio

Generative AI Engineer experienced in designing LLM applications, RAG pipelines, Agentic AI workflows, and backend AI services. Skilled in Python, LangChain, LangGraph, FastAPI, SQL, Docker, and AWS.

Experience

AI & Data Consultant Jun 2025 – Present

Birlanest (CKA Birla Healthcare), Gurugram

- Developed production-grade LLM applications including **WhatsApp chatbot**, **Voice AI** assistant, **OCR-powered document intelligence**, and **Text-to-SQL** systems using **OpenAI**, **LangChain**, and **LangGraph**.
- Built **Retrieval-Augmented Generation (RAG)** pipelines with document ingestion, embeddings, retrieval optimization, **tool calling**, memory, and multi-step reasoning to improve factual accuracy and reduce hallucinations.
- Implemented **prompt engineering** techniques including prompt chaining, context optimization, structured outputs, and function/tool calling for enterprise healthcare workflows.
- Developed **backend AI services** using **Python**, **FastAPI**, **Streamlit**, **MongoDB**, and **PostgreSQL** while integrating enterprise healthcare systems and **structured/unstructured data sources**.
- Designed **ETL pipelines** transferring conversational data from **MongoDB** to **PostgreSQL** and created **Apache Superset dashboards** for AI performance monitoring and business analytics.
- Containerized AI services using **Docker** and deployed scalable applications on **AWS EC2** while monitoring production infrastructure and resolving critical synchronization issues.

Teaching Assistant

IIIT Vadodara, Gandhinagar

Aug 2024 – May 2025

- Mentored 200+ students in Algorithms and Cryptography, improving problem-solving and system design thinking.

Technical Consultant

Mudrika AI (Remote)

Jul 2023 – Jul 2024

- Developed AI Proof-of-Concept applications using Streamlit for image generation, enhancement, and computer vision use cases.
- Implemented backend features in Django applications and integrated AI-powered image processing pipelines.
- Developed OCR pipelines for extracting structured information from unstructured documents.
- Collaborated on cloud-based AI deployments, debugging, testing, and production integration.

Odoo Trainee Engineer

Aspire Softserv, Ahmedabad

Jan 2023 – Jul 2023

- Developed custom ERP modules and automated reporting workflows, improving operational efficiency.
- Built data-driven dashboards supporting business decision-making processes.

Skills

Programming: Python, SQL

LLM & Agentic AI: OpenAI GPT, Claude, RAG, Prompt Engineering, LangChain, LangGraph, Tool Calling, Multi-Agent Workflows, Conversational AI, Text-to-SQL

Backend Development: FastAPI, Flask, Streamlit, REST APIs

Data Engineering: ETL Pipelines, PostgreSQL, MongoDB, Apache Airflow, Apache Superset, Data Warehousing

Machine Learning & NLP: PyTorch, TensorFlow, Transformers, BERT, GPT, NLP, Deep Learning

Document Intelligence: OCR, Vision-Language Pipelines, Multimodal AI

Cloud & DevOps: AWS EC2, Docker, Git, Linux

Concepts: LLM Evaluation, Retrieval Optimization, Hallucination Reduction, AI Guardrails, Vector Embeddings

Education

M.Tech Computer Science (AI), IIIT Vadodara — CPI: 8.84

2024–2026

B.E. Information Technology, Shantilal Shah Engineering College — CGPA: 8.01

2019–2023

Publication

Dynamic Risk Stratification in ICU Sepsis: A Deep Reinforcement Learning Approach Using Synthetic Data — IEEE Xplore

International Conference on Recent Advances in Engineering and Computational Sciences (REACS 2025)

- Proposed D3QN framework for sequential decision-making in ICU patient outcome prediction.
- Trained on 2,164 ICU admissions achieving 75% temporal prediction accuracy.

Projects

NeuralMT-LLM

- Built a Transformer-based Neural Machine Translation system implementing RoPE, RMSNorm, GQA, Beam Search, and SentencePiece tokenization.
- Implemented BERT-style pretraining, GPT-style decoder pretraining, cross-attention, and efficient sequence generation using PyTorch.
- Designed scalable training and evaluation pipelines for multilingual NLP under constrained compute resources.

AI-Generated Image Detection using Lightweight CNNs

- Designed lightweight CNN architecture using Separable Convolutions and SE Blocks for AI-generated image detection.
- Optimized model accuracy and computational efficiency achieving 95% accuracy using only 55K parameters.

Personal Portfolio Website

- Built and deployed a responsive portfolio website showcasing projects, publications, technical skills, and professional experience.
- Integrated project links and a modern UI to improve online visibility and recruiter accessibility.

National Statistics Hackathon (MPCE Analysis)

- Analyzed socio-economic dataset to extract consumption patterns and insights.
- Built predictive models using SVM and ensemble methods to support data-driven conclusions.

Certifications

Advance Course of Emerging Technologies — SAP & Edunet

Fundamentals of Deep Learning — NVIDIA

Achievements

Awarded Silver Medal for Academic Excellence in M.Tech (AI), IIIT Vadodara

Qualified GATE (Data Science & AI), 2024

Qualified GATE (Computer Science), 2024

Qualified UGC-NET (Computer Science), 2024