

YUVAL MEHTA

+91 70459 83199 — yuvalmehta.728@gmail.com — linkedin.com/in/yuvalmehta728 — github.com/Yuval728

SUMMARY

Generative AI Engineer specializing in agentic systems, LLM infrastructure, and production AI platforms. Experienced in LangGraph, distributed systems, MLOps, and end-to-end delivery of AI solutions in regulated environments.

EDUCATION

NMIMS - Mukesh Patel School of Technology, Management and Engineering

06/2025

B.Tech Computer Engineering | CGPA: 3.84 / 4.0

Mumbai, India

EXPERIENCE

Generative AI Engineer

06/2025 – Present

xLM - Continuous Intelligence

Mumbai, Maharashtra

- Owned architecture and delivery of **cEMS** (Environmental Monitoring System) from scratch, building distributed systems, ML pipelines, AI agents, CI/CD workflows, and GxP-compliant infrastructure.
- Reduced **cTM** (Temperature Mapping System) infrastructure costs by **5x** through end-to-end pipeline redesign, scalability improvements, and performance optimization.
- Built a production-grade **audit trail system** capturing actor actions and contextual metadata, enabling reusable GxP-compliant traceability logging across multiple company products.
- Designed **LangGraph** multi-agent systems and delivered key **cIV** (Intelligent Validation Platform) components, improving execution time by **30%**, task success rate by **40%**, and workflow coverage by **65%**.

AI/ML Intern

01/2025 – 05/2025

xLM - Continuous Intelligence

Mumbai, Maharashtra

- Built a traceability matrix generator that cut manual overhead by **60%** and boosted workflow consistency by **45%**.
- Prototyped 3 AI-driven document intelligence solutions, cutting internal validation cycle time by **50%**.

Machine Learning Intern

07/2024 – 05/2025

IIT Kharagpur

Kharagpur, West Bengal

- Built a video feature extraction pipeline using autoencoders and graph neural networks for spatiotemporal representation learning, improving frame processing efficiency by **30%**.

PROJECTS

Qwen3.5 from Scratch | Python, PyTorch, GQA, SwiGLU, RoPE, RMSNorm

- Implemented core Qwen3.5 architecture components (GQA, SwiGLU, RoPE, RMSNorm) from scratch in PyTorch and validated outputs against the official implementation.

Research Pilot | Python, Supabase, LangGraph, LiteLLM, RAG, FastAPI, React

- Built an AI research copilot that converts papers into structured extractions, summaries, architecture diagrams, and PyTorch code using a 10-stage LangGraph workflow.

AQI Prediction | Python, PyTorch, GCP, FastAPI, MLOps, WandB

- Designed a dual-view attention model forecasting AQI and 6 pollutants, achieving **93% R^2** accuracy; reduced cloud training costs by **35%** through targeted optimization and automated hyperparameter sweeps.

AI Therapist | Python, LangGraph, LLMs, Supabase, Vector DB, FastAPI

- Architected an LLM-powered therapy agent with emotion-aware control flows and persistent memory, improving session consistency by **70%** across simulated conversation loops.

TECHNICAL SKILLS

Programming Languages: Python, JavaScript, SQL, C / C++

AI & ML: PyTorch, TensorFlow, LangGraph, LangChain, HuggingFace Transformers, LiteLLM, Scikit-learn, Weights & Biases, MLflow, MCP, A2A, CrewAI, OpenCV, PySpark, Langfuse, Arize Phoenix

Backend & Infrastructure: FastAPI, Docker, GitHub Actions, DVC, REST APIs, Git

Databases: PostgreSQL, MySQL, MongoDB, Redis, Supabase, FAISS, Qdrant, ChromaDB

Cloud Platforms: Azure, AWS, GCP

PUBLICATIONS AND AWARDS

Placed **4th out of 250+ teams** at GenAI Week 2025 Hackathon, Silicon Valley; built a GxP audit chatbot end-to-end.

Published at IEEE InCoWoCo 2025: *Estimating Ground-Level Air Quality Index from Satellite Imagery*.

Achieved **274th out of 74,824 participants** (Top 1%) in the Amazon ML Challenge 2024.

Published at IEEE APCIT 2024: *Examining ML Approaches for Early Diabetes Prediction*.