Komal Kumavat

kkumavat437@gmail.com ❖ 9004600238 ❖ Mumbai, IN ❖ LinkedIn

PROFILE

Innovative and results-driven AI/ML Engineer with 4.5+ years of experience designing, optimizing, and deploying production-grade AI systems across healthcare, SaaS, and education domains.Specialized in deep learning, MLOps, computer vision, and AI agent systems, with proven success reducing inference latency, improving product engagement, and enabling scalable AI infrastructure. A strong believer in building from first principles — translating cutting-edge AI research into real-world, high-impact features with ownership, speed, and precision..

SKILLS

Machine Learning & Deep Learning: CNNs, RNNs, Transformers, LLM Fine-tuning, Model Optimization, Hyperparameter Tuning
AI & MLOps Tools: TensorFlow, PyTorch, scikit-learn, LangChain, Hugging Face, MLflow, Docker, FastAPI, Airflow, DVC, Kubeflow
Data Engineering: ETL Pipelines, SQL/NoSQL, Pandas, NumPy, PySpark, Feature Stores, Vector Databases (FAISS, Pinecone)
Cloud & Deployment: AWS (EC2, S3, SageMaker, Lambda), GCP (Vertex AI, BigQuery), CI/CD, REST APIs, Streamlit, Gradio
System Design: Async Microservices, gRPC, Server-Side Events (SSE), Redis, Celery, ONNX Runtime, TensorRT Optimization
Analytics & Tools: Power BI, Tableau, Matplotlib, Jira, GitHub, Confluence, Notion

EXPERIENCE

Neve Jewels 01/08/2024 -Present
Python/AI/ML Engineer, *Mumbai*

▪ Designed AI-powered pricing and recommendation models using Python + XGBoost, improving pricing accuracy by 22 %.
▪ Automated model retraining via Airflow + Docker on AWS (S3 + EC2), cutting manual updates by 70 %.
▪ Integrated OpenAI GPT models with LangChain to summarize competitor data, reducing research time by 60 %.
▪ Deployed REST APIs (FastAPI) for inference and integrated with Power BI dashboards for live KPI monitoring.
▪ Built end-to-end ML pipelines for sales prediction and anomaly detection; reduced forecast error by 30 %.
▪ Integrated OpenAI GPT models via LangChain for summarizing and analyzing competitor insights, cutting analysis time by 70%.

Techiesys IT Solutions 05/1/2023 - 15/07/2024
Remote ML Engineer, *Bangalore*

- Developed end-to-end ML pipelines for sales prediction and anomaly detection; reduced forecasting error by 30%.
- Deployed models with Flask and Docker, managed retraining workflows via Airflow and automated evaluation metrics.
- Used AWS SageMaker for distributed training and integrated model endpoints for live business dashboards.
- Applied SHAP and LIME for model explainability, improving stakeholder trust and decision transparency.

Freelance AI/ML Developer 05/1/2021-15/12/2022
AI-Enhanced Learning Management System (LMS)
Clients: Silica Institute & Edit System Pvt Ltd | **Role:** Python / AI Engineer | **Duration:** Jul 2023 – Jan 2024
Stack: Python, FastAPI, PyTorch, LangChain, OpenAI APIs, PostgreSQL, Redis, Docker, AWS, SSE, gRPC
Problem:
Manual tutoring and grading restricted scalability and personalization for >10 K learners.
Solutions:
- Developed AI personalization module automating tutoring, grading, and adaptive assessments.
- Architected LangChain + FAISS RAG pipeline grounding GPT outputs in verified course data.
- Built LLM tutoring agents using async FastAPI + Redis + SSE for live sessions.
- Deployed gRPC microservices for scalable inference; implemented evaluation metrics (BLEU, precision@k).

Impact:
- Instructor workload ↓ 40 %, student engagement ↑ 22 %, model latency ↓ 37 %.
- Shipped E2E AI module with zero downtime.

PROJECTS

Brain Tumor Detection & Segmentation System

Client: St. Jude Children's Research Hospital, Tennessee,USA | **Role:** Lead AI Research Consultant | **Division:** Drug Discovery / Diagnostics
Problem:
Manual MRI analysis required hours per patient, delaying tumor diagnosis and research throughput.
Solutions:
- Built U-Net + ResNet-50 Attention hybrid model (96.3 % accuracy).
- Handled imbalance via Focal Loss + GAN augmentation; reduced false positives with attention

gating.

▪ Optimized inference 45 s → 8 s via TensorRT FP16; compressed Docker image 8 GB → 2.1 GB.

▪ Integrated FHIR APIs for EMR sync and added Grad-CAM interpretability.

Impact:

▪ Diagnosis time ↓ 2–3 h → 12 min | 150 + scans/week | throughput ↑ 40 %.

▪ Strengthened clinical confidence through explainable AI.

Instafy – AI Image Transformation SaaS Platform

Startup: Digital Advantage Media | **Role:** Python / AI Engineer | **Duration:** Jan 2024 – May 2024

Stack: Python, FastAPI, OpenCV, TensorFlow, ONNX Runtime, Celery, Redis, Docker, AWS (EC2, S3), MongoDB, Cloudinary, Stripe, Next.js

Problem:

Creative clients needed a scalable AI platform for real-time image restoration, recoloring, and generative fill. Existing tools were slow and expensive.

Solutions:

▪ Built **asynchronous microservice architecture** with **Celery + Redis**, parallelizing GPU jobs.

▪ Converted models to **ONNX Runtime**, cutting inference **2.8 s → 1.1 s**.

▪ Integrated **U²-Net + Meta SAM**, achieving **92 % IoU** accuracy.

▪ Built **credit-based billing system (Stripe Webhooks)** with real-time tracking.

▪ Deployed **Dockerized services** on **AWS Auto-Scaling EC2**, maintaining >99 % uptime.

Impact:

▪ 60 % latency reduction | 25 % cloud-cost savings | 22 % conversion lift | 35 % higher retention.

▪ Delivered reusable **AI SaaS MVP** powering multiple Qureshi Holdings ventures.

ADDITIONAL PROJECTS:

▪ **Predictive Maintenance (IoT + LSTM):** Built Kafka → Airflow → FastAPI pipeline; deployed on AWS EC2 for real-time monitoring.

▪ **Conversational AI Chatbot (LLM + RAG):** Implemented GPT-4 + LangChain chatbot with FAISS retrieval and custom vector store.

▪ **Image Quality Assessment (E-Commerce):** Built CNN-based photo-quality validator improving catalog consistency by 80 %.

EDUCATION

Bachelor's of Computer Science, *Mumbai University* 12/08/2021-02/05/2024

Bachelor's of Data Science and AI , *Mumbai University (Distance Learning)* 12/08/2021-09/05/2024