Anupam Kumar

+91-6202907787 anupraj1620@gmail.com LinkedIn GitHub LeetCode

Education

Indian Institute of Information Technology and Management, Gwalior **2021 - 2025**
Bachelor of Technology in Computer Science

Experience

AI Engineer – Applied AI Systems **Apr 2025 – Nov 2025**
Yahweh Innovations — AI Solutions and Consulting *Remote*

- Designed a production-grade chatbot and knowledge retrieval service using Python and vector search, serving **10,000+ monthly requests** across 5 enterprise clients and reducing query resolution time from **45 sec** to **29 sec** .
- Built scalable backend workflows for customer support, onboarding, and audit automation using Python and PyTorch, processing **500+ tasks daily** and reducing manual effort by **55%** through reliable background job execution and service-based architecture.
- Developed a distributed B2B SaaS platform using FastAPI microservices, Docker containers, and CI/CD-enabled cloud deployment, enabling parallel processing, fault isolation, and improving overall workflow performance by **2**×.
- Delivered a **real estate audit automation platform** using OCR and machine learning to extract and validate data from **400+ monthly documents**, deployed on **AWS EC2**, achieving **95%+ accuracy** and reducing processing time from **several days to 15 minutes**.

Technical Skills

Languages: Python, C++, JavaScript
Software & Systems: Data Structures & Algorithms, Distributed Systems, Microservices, REST APIs, Concurrency, Backend Development
Frameworks: FastAPI, PyTorch, Transformers, scikit-learn
Databases & Storage: PostgreSQL, Redis, MongoDB, Vector Search (Pinecone/FAISS)
Cloud & DevOps: Docker, CI/CD, Git, AWS EC2, Cloud Deployment, Model Evaluation & Monitoring
AI/ML: Machine Learning, Deep Learning, NLP, RAG, Langgraph, LLM Applications, Model Serving, MLOps

Projects

AyurGenix — Intelligent Knowledge Retrieval System — GitHub — *PyTorch, Pinecone, PostgreSQL, Docker*

- Built an **LLM-powered retrieval solution** over **10,000+ documents**, validated across **500+ evaluation queries** using RAGAS-based metrics, delivering context-grounded responses with real-time latency..
- Enhanced semantic search using vector embeddings and Pinecone, achieving **sub-second latency** while improving similarity ranking, search precision, and context relevance for RAG workflows.
- Implemented hybrid retrieval with intent-aware reranking to improve answer relevance, citation accuracy, and response reliability across multi-query and conversational interactions.
- Exposed secure FastAPI-based inference endpoints supporting streaming responses and concurrent workloads, maintaining **5-second response time** in distributed deployments.

LeadBoost SaaS — Lead Generation Platform — GitHub — *FastAPI, Celery, Redis, Docker, React*

- Architected a **multi-tenant SaaS application** with secure authentication and role-based access control for multiple clients.
- Engineered a **distributed task queue** using Celery and Redis to execute large-scale background jobs with retries and parallel workers.
- Automated lead discovery, enrichment, and scoring workflows in Python, enabling AI-assisted data extraction and structured insights that reduced manual research effort by up to 80%.
- Integrated containerized services with monitoring dashboards, ensuring scalable and reliable production operations.

Achievements & Leadership

- Solved **500+ Data Structures and Algorithms problems** on **LeetCode** (**Knight Badge**).
- Secured **State Rank 11** in **NTSE** Stage 1 and cleared the **KVPY** Aptitude Test (Stage 1).
- Fest **Coordinator** at **Aurora '24**, Central India's largest cultural and technical festival.
- Mentored **200+ students** in Generative AI and applied machine learning workflows.
- Founder of **Codegen Alpha**, a technical community with **500+ members**.
- Core organizing team member, **ARIC '23** (Incubation and Entrepreneurship Conclave).