Hiren Vaghela

hvaghela429@gmail.com | +91 7777941869| linkedIn/hirenvaghela | github/hiren-2911

EDUCATION

National Institute of Technology (NIT), Surat Surat, IN | June 2019 - May 2023
BTech in Electronics and Communications Engineering

WORK EXPERIENCE

VISA | Software Engineer(AI/ML) Bangalore, IN | April 2025 - Present

- Improved GenAI platform latency by **35%** by integrating **Claude, OpenAI, Gemini, and Grok** into a scalable **Model Service**, enabling seamless access across **6+ internal applications**.
- Boosted model serving performance by **40%** through **prompt caching, endpoint refactoring**, and **code tracking**—reducing average inference time from **1200ms to 900ms**.
- Streamlined observability and incident response by building **Splunk dashboards** for model usage, error rates, and system health—cutting **MTTD (mean time to detect) by 50%**.
- Designed and implemented an **agentic framework** to automate and streamline LLM model onboarding, enabling faster deployment and reducing manual intervention.
- Optimized integration testing pipeline by introducing multiprocessing and parallel execution strategies, reducing test execution time from **45 minutes to 6 minutes (7.5x improvement)**, thereby accelerating CI/CD workflows.

TVS CREDIT | Software Engineer(AI/ML) Noida, IN | July 2023 - March 2025

- **Built a production-grade KYC document verification system** using PyTorch and OCR, achieving real-time processing at 0.5s/image for Aadhaar, PAN, and face liveliness, improving accuracy and cutting latency by 60%.
- **Led delivery of 6 high-impact ML systems in under 12 months**, including real-time KYC verification and async API pipelines—contributing to □2.5 Cr in annual cost savings through automation and performance optimization.
- **Re-architected monolithic APIs into modular asynchronous services** using Celery and Redis, increasing throughput by 60% and improving fault isolation and system observability.

PUBLICATIONS AND ACHIEVEMENTS

- DCAN:DenseNet with Channel Attention Network for Super-resolution of WCE, EUVIP - 2023 (IEEE) 🔗
- Wireless Capsule Endoscopy Image Super-Resolution using Deep Learning, CVIP - 2023 (Springer) 🔗
- 5* on Codechef and Specialist on Codeforces

PROJECTS

VISIONVAULT ↗

- Developed and deployed an **end-to-end FastAPI microservice** for Aadhaar document automation using **YOLOv5** and **PaddleOCR**, achieving **95.6% field extraction accuracy** on **50k+ scanned documents**; optimized for **sub-1.8s latency** per 5MB file with **99.7% API uptime** over 6+ months.
- Built a **modular, multi-format pipeline** (PDF, TIFF, JPEG, PNG) with **fuzzy PIN/address resolution** and **layout generalization**, reducing **onboarding time by 70%** for new Aadhaar variants and supporting **50+ QPS** with **<2% error rate** in production.

LLM-POWERED HR ASSISTANT PLATFORM ↗ FastAPI, Redis, LangChain, OpenAI, JWT, Docker, K8s, RAG, ChromaDB.

- Designed and deployed a scalable, production-grade HR chatbot backend using **FastAPI**, with **JWT-based authentication** and role-based access control for secure, personalized interactions.
- Implemented **rate limiting using Redis** to prevent abuse and ensure fair usage across users, while maintaining system stability under high load.
- Achieved **150+ QPS** and supported **10k+ DAUs** by leveraging asynchronous APIs, Redis caching, **Docker**, and **Kubernetes** for high concurrency and seamless auto-scaling.
- Built robust **CI/CD pipelines** using GitHub Actions, enforced code quality through **Pytest** and **Postman**, and integrated **LLM fallback (OpenAI GPT)** to autonomously handle 95%+ of user queries.

SKILLS

Languages Python, C++, SQL
Core Engineering Data Structures & Algorithms, Git, CI/CD (GitHub Actions, Jenkins), Pytest
ML Frameworks PyTorch, TensorFlow, Scikit-learn, XGBoost, Hugging Face, MLflow
NLP & LLMs GPT, T5, LLaMA, LangChain, Prompt Engineering, RAG
MLOps & Infra Docker, Kubernetes, Redis, Celery, FastAPI, REST/gRPC
Cloud & Deployment AWS (SageMaker, Lambda), GCP (Vertex AI), Azure ML
System Design Model Serving, Monitoring (Splunk), Feature Stores, Microservices