VIJAY KUSHWAHA

AI/ML (MLOPS) ENGINEER

CONTACT

+919898864226

vijaysinghkushwaha3737@gmail.com

blog.hugging.space | prediction.works

Ahmedabad, Gujarat, India

linkedin.com/in/h3110Fr13nd

github.com/h3110Fr13nd

x.com/h3yFr13nd

SKILLS

Generative AI	5/5
Machine Learning (MLOps)	5/5
AI/ML Development	5/5
Docker, Cloud Tech & DevOps	4/5
Fine-tuning (LLM, Image Gen)	4/5
Web Development	4/5
Data Engineering	4/5
Scraping & Automation	4/5
Frontend (React, Next)	4/5

EDUCATION

PhD - AI/ML

Pandit Deendayal Energy University

Jan '26 - Present

B.E. Computer Engineering

LJ Institute of Engineering & Technology

Nov '21 - May '25

COURSES

DeepLearning - Specialization

DeepLearningAI — *Dec '24 - Jan '25*

Architecting with Google Compute Engine - Specialization

Google Cloud — *Aug '22 - Oct '22*

Meta Back-End Developer

Meta — *Apr '23 - Jun '23*

IBM Data Science & Data Analysis

IBM | Coursera — *Aug '23*

PROFILE

Experienced AI/ML engineer and PhD Scholar specializing in GenAI, self-hosted inference, and MLOps. Open-source contributor @Hugging Face. Proficient in PyTorch, Transformers, Scikit-Learn, and backend tech like FastAPI, Django, Node.js, and cloud platforms (AWS, Azure, GCP). Active in tech communities, driving AI and MLOps solutions.

PROJECTS AND EXPERIENCE

AI/ML Engineer (SDE-2) - eSparkBiz

Apr '25 - Present

Credo AI - AI Governance Platform — *Jan '25 - Present*

- Designed an agent-driven governance platform verifying compliance for AI vendors.
- Orchestrated multi-agent to extract and validate critical security policies.
- Continuous benchmarking with deepeval & improvements using DSPy.

K-One Kelios — *Apr '26 - May '26*

- Built React Native app for enterprise network routing, enhancing setup and debugging.
- Integrated dynamic RAG based screen-aware agent for guiding K-One device installation.

SmackDab AI - AI-Enhanced CRM Workflows Integration — *Apr '25 - Dec '25*

- Deployed context-aware AI Assistants handling CRM queries and workflows.
- Automated insights, summaries, and next-action suggestions using LangGraph.
- Streamlined contacts, meetings, emails, and records via AI extraction.

AI/ML Engineer (SDE-1) - PragetX

Mar '24 - Mar '25

Seamless Fabric Pattern Generator Toolkit — *Feb '25 - Mar '25*

- Built AI textile pattern generator with SDXL, ComfyUI, and text + image input support.
- Used Circular VAE, IPAdapter, and K-Means for style transfer and color quantization.
- Integrated OpenCV and Pillow for masking, separation, and image manipulation.

Spanish TTS - StyleTTS2 Finetuning — *Oct '24 - Jan '25*

- Trained StyleTTS2 for TTS + Voice Cloning on VoxPopuli Dataset with A100 on Colab
- Improved Noise Reduction, Emotion/Style Capturing, Long text-to-speech inference.
- Deployed on AWS accelerated Instance(g4dn.xlarge) for fast inference(<2s res time)

VoAgents - OrionSolutions — *Jul '24 - Sep '24*

- Built End to End real-time AI call assistants for recruitment, support, education etc.
- Using Bolna (open source project) integrated Telephony Providers(Twilio, Plivo), Transcribers(deepgram, whisper), LLMs(OpenAI, LLaMa, Cohere), Synthesizer(AzureTTS, AWS Polly, Deepgram), backend(FastAPI, Uvicorn).
- SetUp CI/CD pipelines using Jenkins, Deployed using docker compose on EC2.
- Optimized the Docker image size from 18GB to 1.2GB (open source contribution) and further sped up docker compose builds by 5x utilising cache mounts.

Company Classifier (dynamic CSV to SQL) - ListenBravo — *Jul '24*

- Streamlines MLOps for CSV-based semi-structured data classification and PostgreSQL insertion with adaptive schema using Gemini, SQLCoder, and LangChain.
- Built with FastAPI, Pandas, and SQLAlchemy. Supports TPU-powered vLLM inference on Modal.

Real Estate Data Analysis & Property Suggestion — *Mar '24 - Jun '24*

- Scraping propertyfinder.ae with Scrapy, Preprocessing and EDA & KNN clustering
- Setup data pipeline to ElasticSearch Indexes, Integration with Kibana Dashboards.
- Containerized using Docker and Docker Compose

Co-founder / CTO - QuantumReach

Nov '23 - Jun '24

AutoBlogging System (VC Grant Received) — *Feb '24 - May '24*

- Built an AI-driven blogging system automating content creation workflows.
- Used GPT-4, SDXL, Qdrant, Flask, Quartz SSG and AWS CI/CD deployment.

Research - Current Trends in Chromatographic Prediction using AI/ML

LJ Institute of Pharmacy - Research Project — *Dec '22 - Apr '23*

- Created ML models for chromatographic drug mobile phase prediction in RP-HPLC.
- Published in Royal Society of Chemistry, patent filed. Research Paper

Personal Projects / Open Source Contributions

HuggingFace - Open Source Contribution — *Sep '24 - Nov '23*

- Transformers - Modular Phi model, solving failing tests, docs update, etc.
- Optimum - added support of image-to-image pipeline for transformer based model (Swin2SR), deps/docs update.

Chat with Me - Llama 2 Fine-tuned on WhatsApp Data — *Oct '23 - Dec '23*

- Finetuned LLaMa 2 7B using Transformers, PEFT, QLORA, TRL, Gradio.

Serverless LLM Inference — *Jan '24*

- Deployed Serverless StableLM-2 1.6B LLM Inference leveraging Lambda, SAM, ECR.