Hrishikesh Keswani

857-398-5289 | keswani.hrishikesh7@gmail.com | linkedin.com/in/hrishikeshkeswani | github.com/hrishikeshkeswani

Education

Northeastern University — Boston, MA
Master of Science in Data Analytics Engineering — *Sep 2023 – Dec 2025*

Experience

Machine Learning Engineer — Jan 2025 – July 2025
Boehringer Ingelheim — *Athens, GA*

- Built an internal HR tool using RAG system with Llama-3, GPT-4o, and Nova Lite to analyze 10K+ employee surveys, enabling HR to surface sentiment trends and retrieve targeted feedback via natural language queries
- Increased LLM inference throughput from 1.2K to 2.0K tokens/sec using vLLM batching and multi-GPU Ray, enabling faster and concurrent HR query handling
- Orchestrated pipeline for survey ingestion, chunking, embedding generation, and FAISS index refresh
- Deployed monitoring with Grafana, and Prometheus, optimizing p95 latency, GPU utilization and retrieval time

Data Scientist — June 2022 – July 2023
Yunometa — *Mumbai, India*

- Developed a multilingual document text detection model using Faster R-CNN (ResNet-50) in PyTorch, improving extraction recall by 30% across loan application forms, accelerating review for loan processing teams
- Built OpenCV preprocessing (deskew, resize, normalize) to improve detection robustness across noisy scans
- Architected event-driven AWS pipeline (S3, Lambda, SageMaker) for PDF ingestion and scalable inference
- Converted detected regions into structured JSON stored in PostgreSQL to enable downstream processing

Data Scientist — Oct 2021 – Mar 2022
Kritexco — *Mumbai, India*

- Analyzed 50K+ NFT marketplace transactions from OpenSea using Python to identify suspicious trading behavior and monitor secondary market activity for client NFT collections
- Engineered transaction features including wallet trading frequency, trade velocity, and price deviation from floor price, and addressed class imbalance using SMOTE for a dataset with ∼2% fraudulent transactions
- Developed fraud detection models using Logistic Regression and Random Forest, achieving PR-AUC of 0.82 and recall of 0.78 on highly imbalanced fraud transactions

Projects

AI Analyst Copilot | *FastAPI, pgvector, Groq, React, PostgreSQL, sentence-transformers*

- Engineered RAG-based schema linking from scratch using pgvector and sentence-transformers, achieving 80% NL2SQL execution accuracy on a 20-query benchmark
- Built end-to-end NL2SQL agent over 50K+ row PostgreSQL database with automated insight generation, root cause analysis, and sub-1.5s average query latency via Groq/Llama-3
- Implemented SQL correction feedback loop storing query-correction pairs for future fine-tuning, deployed as full-stack system with FastAPI backend and React dashboard

LinkedIn Search Optimizer | *Python, LangChain, FAISS, LLMs, GCP*

- Built a RAG-based search engine (LangChain, FAISS, LLMs) analyzing 124K+ LinkedIn posts to deliver contextual job search via natural language queries
- Built distributed Airflow pipeline on GCP with TTL-based vector refresh to sustain search relevance across LinkedIn, Indeed, HackerNews, and Reddit
- Deployed production search system using Docker, FastAPI, and Railway with real-time logging, achieving 0.8 Recall@10 on labeled query-job pairs

Technical Skills

Gen AI: Finetuning, LoRA, QLoRA, HuggingFace, LLMs, CUDA, vLLM, MLFlow, RAG, PyTorch
Machine Learning / Statisitcs: Regression models, Tree-based/Boosted Models, Neural Networks, A/B Testing
Programming: Python, SQL, Java, C++, Golang, Scala
Frontend Full-Stack : React, Next.js, HTML, CSS, JavaScript, TypeScript
Databases: Kafka, Redis, PostgreSQL, MongoDB, Druid, Solr
Cloud DevOps: AWS (S3, Lambda, EC2, EKS, EMR), Docker, Kubernetes, OpenShift, CI/CD, Git
APIs Frameworks: FastAPI, Django, GraphQL, TensorFlow/Keras, Springboot