Kaushal Maniyar

Portfolio: kaushal.maniyar.github.io
LinkedIn: linkedin.com/in/kaushal-maniyar
Email: maniyarkaushal111@gmail.com
Phone: +91 7359731209
GitHub: github.com/Kaushal-maniyar

PROFESSIONAL SUMMARY

AI/ML Developer and M.Tech graduate with hands-on experience in information retrieval, document ingestion, computer vision, deep learning, and deployed machine learning systems. Current work includes a tender-search pipeline built around BM25F ranking, custom word normalization, MongoDB retrieval, Playwright-based ingestion, and local LLM document processing with Qwen3-30B on DGX Spark. Research at SAC (ISRO) improved the floating-litter classification F1-score from the 0.50 MariNeXt baseline to 0.70 through model comparison and multispectral feature development.

EDUCATION

Nirma University, Ahmedabad, India **CGPA: 8.6/10**
Master of Technology - Data Science **2022-2024**

Relevant Coursework: Machine Learning, Deep Learning, Database Design, Big Data, Statistics for Data Science, Scalable Systems

Government Engineering College (GTU), Rajkot, India **CGPA: 8.6/10**
Bachelor of Engineering - Computer Engineering **2018-2022**

Relevant Coursework: Object-Oriented Programming, Data Structures and Algorithms, Operating Systems

TECHNICAL SKILLS

Programming: Python, SQL (PostgreSQL, MySQL), Java, C++, Bash
Machine Learning & AI: PyTorch, Scikit-learn, XGBoost, OpenCV, Transformers, Qwen3-30B, Qwen-7B, local LLMs, LLM APIs, AI agents, LangChain, LlamaIndex, RAG, PEFT/LoRA, QLoRA, feature engineering, hypothesis testing, statistical modeling
MLOps & Cloud: Docker, AWS (EC2, ECR, S3), NVIDIA DGX Spark, MLflow, FastAPI, Prometheus, Grafana, Git, Linux, CI/CD
Data Engineering & Retrieval: PySpark, PostgreSQL, MongoDB, Qdrant, Pinecone, Chroma, BM25F, custom word normalization, vector databases, Next.js, Playwright, Requests, BeautifulSoup, PyMuPDF, REST APIs, Webhooks, JSON, n8n
Analytics & Visualization: Pandas, NumPy, Matplotlib, Seaborn, Power BI

WORK EXPERIENCE

AI/ML Developer, Ishan Technologies, Ahmedabad, India **Jan 2026 - Present**

- Designed and implemented a hierarchical search pipeline in Next.js and MongoDB for more than 45,000 Government e-Marketplace (GeM) procurement tenders.
- Developed a BM25F multi-field ranking engine with lightweight plural stripping and suffix normalization to search top-level tender metadata alongside nested PDF text stored in MongoDB.
- Refined retrieval with MongoDB `$unionWith` aggregations, stemmed regex queries, and structured ID short-circuiting; these changes reduced document discovery latency by 65%.
- Developed web-scraping collectors with Playwright, Requests, and BeautifulSoup to extract tender and corrigendum data from more than 15 enterprise portals, including GeM, Alliance Air, Balmer Lawrie, and BSEDC.
- Parsed and normalized unstructured HTML tables and PDF documents with PyMuPDF, then persisted the structured records in MongoDB.
- Prepared the normalized tender data for downstream retrieval workflows and automated email notifications.
- Processed documents downloaded by the web-scraping collectors through a local RAG pipeline using Qwen3-30B deployed on NVIDIA DGX Spark.
- Created three document-type-specific system prompts to extract scope of work, eligibility criteria, contact details, and other relevant tender information from the collected documents.

Junior Research Fellow, SAC (ISRO) / Nirma University **Jun 2024 - Nov 2025**

- Conducted floating-litter detection research using more than 5GB of Sentinel-2 multispectral imagery from the MADOS dataset.
- Examined information across 12 multispectral bands and created spectral-signature features for the classification pipeline.
- Developed radiometric indices from the multispectral imagery to support feature selection and downstream classification.
- Implemented an SVM classifier that outperformed the MariNeXt benchmark, improving the F1-score from 0.50 to 0.70.

Software Engineering Intern - Test & Automation, Vantiva, Chennai, India **Jun 2023 - Jun 2024**

- Automated system-validation and data-verification workflows across more than 50 core software modules using Java and Selenium, reducing release validation time by 40%.
- Integrated automated test suites into nightly build and integration pipelines, helping development teams identify validation issues earlier in the release cycle.

KEY PROJECTS

Enterprise Hybrid Search & Automated RAG Benchmarking Engine

- **Tech Stack:** Python, PyTorch, Qdrant, BM25F, RRF, FastHTML, vLLM, Ragas, TruLens, Docker, AWS EC2
- Built a hybrid retrieval engine for unstructured enterprise documents, combining sparse BM25F keyword ranking with dense embedding search backed by Qdrant.
- Applied Reciprocal Rank Fusion to merge and rerank sparse and dense result sets, balancing exact keyword matches with semantically relevant document passages.
- Developed an automated RAG benchmarking framework with Ragas and TruLens to compare chunking strategies and embedding models through repeatable evaluation runs, increasing retrieval quality from 0.62 to 0.84 MRR@10.
- Containerized local model inference with Docker and deployed vLLM on AWS EC2 GPU instances; continuous batching reduced Time to First Token by 45%.

Distributed Multispectral Feature Pipeline & Real-Time Model Serving

- **Tech Stack:** PySpark, XGBoost, Scikit-learn, MLflow, FastAPI, Docker, Prometheus, Grafana
- Engineered a distributed PySpark ETL pipeline to process more than 20GB of 12-band multispectral satellite imagery with automated radiometric calibration and cloud-mask feature extraction.
- Tracked more than 50 XGBoost and SVM training runs with MLflow and implemented a CI/CD retraining trigger driven by data-drift checks.
- Deployed Dockerized FastAPI model endpoints with Prometheus and Grafana monitoring, achieving P95 inference latency below 120ms.

Qwen-7B Tender Schema Extraction Fine-Tuning Pipeline

- **Tech Stack:** Qwen-7B, Hugging Face PEFT, LoRA, QLoRA, Docker, vLLM
- Replaced zero-shot prompt-based extraction by fine-tuning a Qwen-7B base model with PEFT LoRA on more than 10,000 enterprise tender documents, reaching 94% structured-schema extraction accuracy.
- Applied 4-bit QLoRA with double quantization to reduce GPU compute costs by 65% during model retraining runs.
- Merged the trained LoRA adapter into the base-model weights and packaged the resulting model with Docker and vLLM for low-latency batch inference.

Website-Integrated AI Chatbot Automation Workflow

- **Tech Stack:** n8n, LLM APIs, AI Agents, Pinecone, REST APIs, Webhooks, JSON
- Developed an AI-powered chatbot for a live website, providing visitors with immediate, context-aware responses through the web interface.
- Built an event-driven n8n workflow to orchestrate communication among the website, Pinecone vector database, and commercial LLM APIs using REST endpoints, webhooks, and JSON payloads.
- Implemented a Retrieval-Augmented Generation pipeline that retrieves relevant content from indexed documents before response generation, improving factual grounding and reducing hallucination risk.