

Bhuwan Jain

er.bhuwanjain@gmail.com | +91 78917 23145 | [LinkedIn](#) | [GitHub](#) | [Portfolio](#) | [LeetCode](#) (138)

SUMMARY

Machine Learning Engineer specializing in LLM applications, agentic systems, RAG pipelines, and production AI deployment. Shipped production systems including a multi-tenant Voice AI SaaS serving 3+ business clients, a hybrid-retrieval RAG drafting agent over a 50+ document corpus, and a deepfake detection model at 88.4% test accuracy. Experienced across the full ML lifecycle — data pipelines, modeling, deployment, and monitoring — with a focus on scalable, reliable systems.

SKILLS

Languages: Python, C++, TypeScript, JavaScript

AI/ML: PyTorch, Scikit-Learn, Hugging Face, LangChain, LangGraph, LlamaIndex, RAG, embeddings, semantic search, OpenCV

LLMs & Inference: GPT, Claude, Llama, Mistral (OpenAI / Anthropic APIs, Ollama)

LLMOps & Observability: LangSmith, Langfuse

Infrastructure & Tools: FastAPI, Flask, Docker, Kubernetes, Qdrant, Kafka, LiveKit, AWS (EC2), Git, Postman

EXPERIENCE

Junior AI Developer — Chittanshai Feb 2026 – Jun 2026

- Architected and launched a production multi-tenant Voice AI SaaS serving 3+ business clients on a FastAPI + Next.js stack integrated with the Bolna telephony API, cutting new-tenant onboarding to under 2 hours.
- Engineered LLM prompt pipelines powering customer-service chatbots and AI assistants across 5+ use cases, reducing fallback handoff rate ~30% through structured outputs and few-shot tuning; instrumented LangSmith and Langfuse tracing for LLM evaluation and production monitoring.
- Built the Oil India PQ Drafting Agent, a production RAG system that auto-drafts cited replies to Parliament Questions on the company's E&P operations from a 50+ document ministry corpus, with a LangGraph drafter (GPT/Claude) and citation validation.
- Delivered an MVP AI stock-market insights platform with real-time market analysis and a conversational order-placement interface, built end-to-end in 6 weeks on LangGraph and FastAPI.

Summer Intern — Pune Knowledge Cluster (PKC) May 2024 – Aug 2024

- Built a MERN-stack quiz platform that crowdsourced 5,000+ galaxy image classifications for downstream ML training; designed the MongoDB schema and a SQL Server to MongoDB ingestion pipeline.
- Reduced image-retrieval latency 40% with Redis caching and containerized the full stack with Docker for one-command deployment; implemented JWT auth, an admin dashboard, and randomization capping each image at 100 quizzes.

PROJECTS

Deepfake Detection — Master's Thesis | *PyTorch, OpenCV, AWS EC2, Flask, React, TypeScript*

- Built a ResNeXt-50 + LSTM hybrid on the FaceForensics++ and DFDC datasets, achieving 88.4% test accuracy and outperforming pure-CNN baselines on temporally manipulated clips.
- Containerized and deployed on AWS EC2 behind a Flask API with a React/TypeScript UI for real-time video analysis.

Oil India PQ Drafting Agent — RAG over Ministry Corpus | *LangGraph, Qdrant, FlashRank, FastAPI, Next.js*

- Engineered the hybrid-retrieval core over a 50+ document corpus (DOCX + XLSX): Qdrant dense-embedding (semantic) search fused with BM25 keyword search via RRF, FlashRank cross-encoder reranking, and token streaming to a Next.js client over SSE.

Autonomous Multi-Agent AI System | *Python, LangChain, LlamaIndex, Qdrant, FastAPI, Docker*

- Designed a Planner–Executor–Critic multi-agent system that decomposes complex goals, performs dynamic tool calling (function calling), and self-validates through long-term memory; shipped as Dockerized FastAPI microservices for research, data-analysis, and decision-making workflows.

Voice AI Platform — Multi-tenant SaaS | *FastAPI, PostgreSQL, Next.js, TypeScript, LiveKit, JWT*

- Designed the multi-tenant architecture — workspace isolation, JWT auth, webhook event handling, and a provider-agnostic telephony abstraction layer that lets the voice engine be swapped with no client-side changes.

EDUCATION

Birla Institute of Technology, Mesra, Ranchi Aug 2020 – Aug 2025

Integrated M.Sc. in Mathematics and Computing | CGPA: 7.6/10