

Rohit Lakhani

AI / ML Engineer | RAG, LLM Agents & Evaluation | Python, FastAPI, AWS

Vadodara, Gujarat, India | Open to Ahmedabad, Pune, Bengaluru and remote | +91 99130 45163
rohitlakhani8888@gmail.com | linkedin.com/in/rohit-lakhani-68710a192 | github.com/rohitlakhani207

SUMMARY

AI / ML engineer with nearly 3 years of production software experience, now focused on LLM applications. Builds RAG pipelines and LLM agents and measures them: an open-source legal RAG system with 98.6% retrieval recall, and an agent guardrail that missed zero dangerous actions across 150 labelled tool calls. Brings the engineering to ship AI to real users: FastAPI and ASP.NET Core backends, PostgreSQL, Docker, AWS, and two live apps on Google Play and the App Store.

TECHNICAL SKILLS

AI & LLMs: Retrieval-Augmented Generation (RAG), LLM agents, tool calling, LangGraph, LangChain, prompt engineering, embeddings, vector search (pgvector), hybrid search, rerankers, LLM evaluation, guardrails

Machine Learning: scikit-learn, TensorFlow, Keras, Transformers, NLP, NumPy, Pandas

Models & Serving: Ollama, llama.cpp, Hugging Face Hub, ONNX (fastembed), Groq, Cloudflare Workers AI

Languages: Python, SQL, C#, Dart, Kotlin, Java

Backend & Cloud: FastAPI, ASP.NET Core, REST APIs, WebSockets, SignalR, PostgreSQL, SQLite, Docker, AWS (EC2, ECS, RDS, S3), GitHub Actions (CI/CD), pytest

Mobile & Tools: Flutter, Android, WebRTC, Clean Architecture, MVVM, Git, Jira, Agile

AI PROJECTS

Legal RAG + Evaluation | github.com/rohitlakhani207/legal-rag Oct 2026

Python, FastAPI, PostgreSQL + pgvector, Ollama, ONNX embeddings and reranker, Flutter, Docker, GitHub Actions

- Built a RAG assistant over the EU GDPR and India's DPDP Act that cites the article or section behind every sentence and abstains when the law doesn't cover the question (5 of 5 out-of-scope questions).
- Benchmarked vector, hybrid (pgvector + full-text, Reciprocal Rank Fusion) and reranked retrieval on 40 labelled questions; the reranker raised Recall@5 from 90.0% to 98.6% and citation accuracy from 80.0% to 94.3%.
- Selected the reranker from four cross-encoders and added a document-frequency cutoff that lifted hybrid recall from 85.7% to 90.0%; runs on CPU with no paid APIs, with CI against a real pgvector database.

AgentGate: Guardrails for AI Agents | github.com/rohitlakhani207/agentgate Oct 2026

Python, LangGraph, FastAPI, WebSockets, AWS Strands Decider 2B, Cloudflare Clef-flash, Groq, scikit-learn, SQLite

- Built a two-step gate for a LangGraph agent with two decision models released in Oct 2026: Decider 2B checks every tool call, Clef-flash re-checks unsure ones, and the rest go to a phone web app for approval.
- Benchmarked six gating approaches, including keyword rules, a trained classifier and an LLM judge, on 150 labelled tool calls with 20 hidden prompt injections; the two-step gate missed zero dangerous actions.
- Showed Decider 2B was too overconfident to filter first (69% of calls still reached a human) but routed 93 of 100 requests correctly as a model router; Clef-flash flagged 17 of 20 injections with zero false alarms.

PROFESSIONAL EXPERIENCE

Mobile Application Developer (AI/ML) | WebBrains Technologies, Vadodara Jul 2026 – Present

- Build RAG workflows with embeddings and semantic search that return context-aware answers, served through REST APIs to the company's Flutter apps.
- Prototype Transformer, RNN and LSTM models in TensorFlow and Keras for new AI features.

Technical Solution Architect (Mobile) | Vandana Technologies, RemoteSep 2025 – Jun 2026

- Led **Pranaam Guruji**, a real-time consultation platform with chat, voice and video, from architecture to production release; live on [Google Play](#) (1,000+ installs) and the [App Store](#).
- Architected the backend from scratch: ASP.NET Core REST APIs, SignalR hubs, authentication, custom middleware and a 14-table PostgreSQL schema.
- Built offline-first sync (ObjectBox) with conflict resolution, sync queues and failover, cutting network dependency by about 30% through caching and API design.
- Built the WebRTC and SignalR real-time layer: connection lifecycle, peer negotiation, network recovery.
- Deployed Docker containers on AWS (ECS, EC2, RDS, S3); managed releases in Git, Jira and Confluence.

Flutter Developer | IndaPoint Technologies, VadodaraSep 2024 – Sep 2025

- Provisioned and managed AWS (EC2, RDS, S3) for live backend services, from instance setup to database connectivity and storage.
- Built production WebSocket and SignalR real-time modules with health monitoring and graceful degradation.
- Improved API latency via query optimisation and connection pooling; shipped in Agile sprints with Git and Jira.

Android & Cross-Platform Developer | Rethinkingweb, VadodaraJan 2024 – Aug 2024

- Built Flutter and native Android (Kotlin, Java) apps with Retrofit REST integration, MVVM and Clean Architecture; optimised API response handling and error recovery.

CLIENT PROJECTS

- **LexEdge (AI legal practice management)**: Integrated an AI chatbot into API-driven legal workflows with automation and scheduling; live on [Google Play](#) and the [App Store](#).
- **Poolr Parents (school carpooling and safety)**: Built real-time GPS ride tracking with secure communication; optimised API and database calls for live location updates.

EDUCATION

M.Sc. IT (Cyber Security & Forensics) | Parul University, Vadodara | CGPA 8.352022 – 2024

M.Sc. Chemistry | Pandit Deendayal Energy University, Gandhinagar | CGPA 9.532020 – 2022

CERTIFICATIONS

Microsoft Certified: Azure Fundamentals (AZ-900) | MicrosoftJun 2023

Microsoft Certified: Power Platform Fundamentals (PL-900) | MicrosoftJun 2023