Raif Karani

Bhatkal, India | +91 7624916594 | karaniraif@gmail.com | linkedin.com/in/raif-karani | github.com/RAIF-KARANI | raifkarani.in

Relocating to Dubai, UAE | **Visa:** Requires UAE employment visa | **Availability:** Immediate | **Languages:** English, Hindi, Urdu

SUMMARY

Computer science graduate building LLM systems end to end: retrieval, structured generation, evaluation with statistical gates, and shipping them as reliable products, e.g. CodeLens, a local-LLM code reviewer released as a Windows app.

EDUCATION

Visvesvaraya Technological University (VTU), Karnataka 2022 - 2026

Bachelor of Engineering (B.E.) - Computer Science and Engineering CGPA: 7.9 / 10

Relevant Coursework: Machine Learning, Artificial Intelligence, Data Structures & Algorithms, DBMS, Operating Systems, Networks

PROJECTS

CodeLens - AI Code Reviewer for GitHub Pull Requests **Live** | **Code**

Python, FastAPI, Celery, Redis, PostgreSQL (pgvector), tree-sitter, Ollama (qwen3:8b), Next.js 16, Electron, Docker, Playwright, Vercel

- Built a GitHub App that reviews pull requests with a local 8B LLM over a tree-sitter code graph and pgvector retrieval, posting inline comments about 40 seconds after a push; a validation layer drops findings not on a changed line or not citing code the model saw.
- Built an 83-case evaluation harness (planted bugs, safe counterfactual twins, clean refactors) gated by a paired-bootstrap 95% CI: raised F1 from 0.65 to 0.77 and recall from 0.71 to 0.94, rejected 4 ideas that failed the gate, caught a data leak early.
- Shipped CodeLens Desktop, a Windows installer (Electron + Docker Compose) whose setup wizard installs Docker and Ollama and creates the user's own GitHub App; released as v0.1.1 after a fresh-machine rehearsal.
- Shipped a live demo on Vercel + Supabase (fixed a connection-limit outage; DB latency 1.2 s to 9 ms); 900+ tests in CI, incl. Playwright end-to-end and WCAG 2.2 AA checks.

my-rag - Personal RAG Assistant over GitHub Repos and Websites

Python, FastAPI, SQLite (FTS5, sqlite-vec), Ollama (qwen3:8b, qwen3-embedding), Playwright, Server-Sent Events, Web Components

- Built a local RAG assistant over my GitHub repos and websites with hybrid search (keyword + vector + project cards, RRF fusion) that cites the exact file and lines for every claim, streamed via a FastAPI SSE API to an embeddable chat widget.
- Evaluated on a 38-question golden set: correct source in the top 10 for 30/30 answerable questions (~100 ms per search), 8/8 unanswerable questions refused by an evidence gate, 0 hallucinated answers, ~3 s per answer.

RAG Document Assistant with Automated Evaluation **Code**

Python, LangChain, Chroma, Hugging Face sentence-transformers, Ollama (Llama 3.2), Ragas, Streamlit, pytest

- Built a fully local RAG assistant (chunking, Chroma vector search, grounded answers with cited sources, document scoping) with a Streamlit chat UI and evaluation dashboard; no API keys or cloud cost.
- Built a Ragas evaluation harness over 32 hand-labeled questions (incl. unanswerable ones) scoring faithfulness 0.77 and context recall 0.78, with a pytest regression gate and 10 isolated experiments on chunk size, top-k and embedding models.

QueryCraft - AI-Powered SQL/NoSQL Query Generator **Code**

Next.js 15, React 19, TypeScript, Node.js, Express, MongoDB, PostgreSQL, MySQL, Neo4j, Gemini / OpenRouter / Ollama

- Built a natural-language query engine that generates and executes queries on 5 database engines (PostgreSQL, MySQL, SQLite, MongoDB, Neo4j), routing prompts across 3 LLM providers with automatic retries.
- Exposed 12 REST endpoints with JWT and bcrypt auth, rate limiting, and Helmet security headers, and built a multi-turn Next.js chat UI that summarizes history to keep context across query refinements.

LinkedOut - Reverse-Hiring Career Platform *(Collaborative Project)* **Live** | **Code**

Next.js 15, React 19, NestJS, TypeScript, PostgreSQL (Supabase), Drizzle ORM, Docker, Turborepo, GitHub Actions

- Co-built an employee-first hiring platform (Turborepo monorepo: Next.js + NestJS, 300+ Jest tests) with a 9-stage hiring pipeline, custom JWT auth with rotating refresh tokens, per-IP rate limiting and Row Level Security, deployed with GitHub Actions.

EXPERIENCE

Tap Academy Jan 2026 - Jul 2026

Software Engineering Intern

- Completed industry-oriented Java full-stack training (core Java, OOP, JDBC, Servlets/JSP, SQL, front-end), applying data structures, algorithms and database design in hands-on builds with MVC practices.

Golden Bird Education Feb 2025 - May 2025

Android & Web Development Intern

- Built 3 production-ready mini-applications with Firebase and the MERN stack (REST APIs, secure auth) in a 5-member Agile team, and reduced Android network latency by 25% via caching and optimized data fetching.

TECHNICAL SKILLS

AI / ML: Generative AI, Large Language Models (LLMs), RAG, hybrid search (full-text + vector, RRF), LangChain, vector databases (pgvector, Chroma, sqlite-vec), embeddings, Hugging Face, LLM evaluation (Ragas, bootstrap confidence intervals), prompt engineering, structured outputs, Ollama, Gemini, OpenRouter, Machine Learning, NLP, PyTorch, TensorFlow, scikit-learn, Pandas

Languages: Python, TypeScript, JavaScript, Java, SQL, C/C++

Backend & Data: FastAPI, Celery, Node.js, Express.js, NestJS, Flask, Streamlit, PostgreSQL, Redis, SQLite, MySQL, MongoDB, Neo4j, Supabase, SQLAlchemy, Drizzle ORM

Frontend & Desktop: React, Next.js, Tailwind CSS, Electron, Socket.io

Tools & Platforms: Git, Docker, pytest, Playwright, GitHub Actions, CI/CD, Vercel, Railway, Cloudflare, AWS, Azure, Postman

CERTIFICATIONS

Build Your Own Generative AI Model, NxtWave (Oct 2024); Full Stack Web Development Specialization, Coursera (Jul 2024)