

Jaimin Panchal

AI Engineer — LLM | RAG | Agents | Python/FastAPI

📍 Ahmedabad, India | 📞 +91 99984 57004 | ✉️ jaiminmpanchal02@gmail.com

🌐 linkedin.com/in/jaiminpanchal2002 | 🐙 github.com/jaiminpanchal2002 | 🌐 jaiminpanchal.com

2.5+ yrs Software Engineering • LLM/RAG Production Systems • 10K+ Orders/Month at Scale • 6 Languages Spoken

PROFESSIONAL SUMMARY

AI engineer building production **LLM systems** — RAG pipelines, tool-calling agents, and evaluation harnesses — on a foundation of 2.5+ years shipping backend software at scale. Hands-on with **Python/FastAPI, LangChain/LangGraph, vector search (pgvector, ChromaDB), embeddings, and RAG evaluation (RAGAS, DeepEval)** against OpenAI and Google Gemini. Backs LLM work with real production engineering: a Shopify order backend sustaining **10,000+ orders/month**, event-driven queues, REST APIs, Docker, and CI/CD. Focused on making LLM applications reliable — retrieval quality, deterministic guardrails, and measurable evaluation — rather than demos. Based in Ahmedabad, available immediately.

TECHNICAL SKILLS

AI / LLM: RAG (hybrid search, RRF, reranking), LangChain, LangGraph, Agents & Tool Calling, Vector Databases (pgvector, ChromaDB), Embeddings, Prompt Engineering, Structured Outputs, RAG Evaluation (RAGAS, DeepEval), OpenAI, Google Gemini, Ollama

Languages: Python, PHP, JavaScript, TypeScript, SQL

AI Serving & Backend: FastAPI, Streamlit, Laravel, Node.js, REST APIs, Event-Driven Architecture, Queues (Redis, BullMQ), WebSockets, JWT/OAuth 2.0

Databases & Retrieval: PostgreSQL + pgvector, MySQL, ChromaDB, Redis, MongoDB, Query Optimization, Indexing

Cloud & DevOps: Docker, AWS (EC2, S3), GitHub Actions, CI/CD, Git, Linux, Nginx

Testing & Evaluation: pytest, RAGAS, DeepEval, PHPUnit/Pest, Jest

Concepts: OOP, SOLID, System Design, Microservices, Scalability, Agile/Scrum

PROFESSIONAL EXPERIENCE

Freelance AI & Software Engineer

Sep 2025 – Present

AI / LLM & Backend Engineer

Remote

- Ship independent **AI/LLM** and full-stack engagements using **Python/FastAPI, LangChain/LangGraph, pgvector/ChromaDB, and OpenAI/Gemini** — production RAG pipelines, tool-calling agents, and evaluation harnesses (detailed under Projects).
- Built the **Laravel + MySQL** backend and REST APIs for **Spot It Now**, a cross-platform wildlife-sighting product — geospatial sighting feeds, photo uploads, secure auth (sign-up / login / OTP), and push notifications consumed by the Android and iOS clients — plus an admin panel for moderation and AdMob.

Sapphire Software Solution

Jul 2023 – Aug 2025

Software Engineer (Backend Focus)

Ahmedabad, India

- Delivered five client platforms end-to-end (Laravel, MySQL) across SaaS, e-commerce, fintech, and social products; all five remain live in production.
- Optimized MySQL performance through indexing and query restructuring and moved order fulfilment into queued background workers for a Shopify order backend sustaining **10,000+ orders/month**.
- Engineered event-driven Laravel backends (queues, events, secure webhooks) powering live bid broadcasting, a transactional wallet ledger, and automated revenue-split payouts — removing manual payout calculation.
- Owned end-to-end delivery — requirements, architecture, implementation, code review, and post-launch support — earning **Employee of the Quarter (2025)** and the **Emerging Star Award (2024)**.

AI / LLM PROJECTS

DocIntel — Multimodal Document Intelligence API

 GitHub

Python, FastAPI, Vision LLM (Gemini / GPT-4o), Pydantic, Docker

- Built a multimodal document-intelligence API (FastAPI) that extracts structured fields from PDFs and images using a **vision LLM**, returning schema-validated **Pydantic** outputs with per-field review flags and fallback handling for low-confidence extractions.

- Added a rules-based validation layer (schema, required-field, and total-consistency checks), pytest coverage, Docker packaging, and a demo UI where users upload a document and receive validated JSON.

production-rag-eval — RAG Chatbot with Evaluation Harness



Python, FastAPI, LangChain, ChromaDB, RAGAS / DeepEval, Gemini + Ollama, Docker

- Built a RAG chatbot (FastAPI + Streamlit) with bge-base embeddings, ChromaDB vector search, and MMR retrieval, served by Google Gemini and local Ollama models for cost-controlled inference.
- Automated quality evaluation with RAGAS and DeepEval (faithfulness, answer relevancy, context precision) in a GitHub Actions pipeline, reaching 92% test coverage and sub-1.2s average response latency.

ResolveAI — LLM Support Agent with Hybrid Retrieval



Python, LangGraph, FastAPI, PostgreSQL + pgvector, Redis

- Built a LangGraph agent that resolves support cases via tool calling and hybrid retrieval — pgvector (HNSW cosine) fused with Postgres full-text via Reciprocal Rank Fusion, then cross-encoder reranked — benchmarked against a dense-only baseline.
- Enforced support policies as deterministic checks in application code (not prompts) so high-risk cases always escalate to a human, eliminating unsafe auto-resolutions on the evaluation set.

ShopSense — E-Commerce API with AI Product Search



Node.js, NestJS, TypeScript, PostgreSQL + pgvector, Prisma, Redis, Stripe

- Added natural-language product search and an LLM shopping assistant to a NestJS/TypeScript e-commerce backend using **pgvector semantic search and function calling**, documented with OpenAPI and tested with Jest.
- Backed it with JWT auth, Prisma/PostgreSQL, Redis/BullMQ queues, Stripe checkout, and webhook-driven order reconciliation with Socket.io real-time updates.

FeedbackIQ — Laravel SaaS API with LLM Analysis



Laravel 11, PHP 8.2, MySQL, Redis, Sanctum, Docker

- Built a multi-tenant SaaS REST API that ingests user feedback and classifies **sentiment, themes, and urgency via an LLM** dispatched to queued Redis workers, keeping the ingest endpoint fast.
- Added OpenAPI documentation, Pest tests, and a Dockerized GitHub Actions CI pipeline, with a deployed live demo.

EDUCATION

- **B.Tech in Computer Science & Engineering** — Indus University, Ahmedabad | 2019–2023 | CGPA 8.89/10

CREDENTIALS

- **Generative AI Mastermind** — Outskill (2026): applied LLMs, prompt engineering, and AI application building.
- **IBM Project Manager** — Professional Certificate, Coursera (2026): project lifecycle, risk and stakeholder management, Agile/Scrum; CAPM exam preparation. [Credential](#)

LANGUAGES

- **Gujarati** — Native | **Hindi** — Native | **English** — Fluent (Professional Working Proficiency)
- **Marathi** — Conversational | **German** — Basic | **French** — Basic