

BALKRISHNA MEHTA

AI Engineer · Surat, India · Open to remote
krishmehta208@gmail.com | balkrishnamehta.in

AI engineer with a full-stack foundation, building production agentic systems across LangGraph workflows, hybrid RAG, evaluation, observability, and self-hosted inference—from React interfaces to Python services and GPU deployment.

Experience

Propelius Technologies

Sep 2024 – Present

Software Engineer

Surat, Gujarat · Full-time

- Own a production AI stylist and support agent serving 2.3K customers and 4.4K monthly messages across Android, iOS, and web, peaking at 294 messages in a day.
- Shipped semantic text and image search handling 70K monthly requests across 11K products and 60K variants.
- Run chat, vision, embeddings, and two fine-tuned adapters on self-hosted GPUs for about \$500/month, with a costed path to \$300 on reserved capacity; shipped six adapter versions through a champion-versus-challenger replay gate with 95% confidence intervals.
- Raised language consistency from 0.754 to 0.965 and tool-choice accuracy from 0.833 to 0.900 on held-out conversations.
- Delivered features across 10+ client projects, operated AWS infrastructure (EC2, S3, nginx), and worked directly with clients through requirements and demos.

Blue Data Consulting

May 2023 – Mar 2024

Intern

Surat, Gujarat · Internship

- Built React interfaces and containerized and shipped applications with Docker.

Projects

Harrir — agentic AI platform for Arabic-language commerce *LangGraph · FastAPI · Qwen3-VL · vLLM · Qdrant · MongoDB · Redis · AWS*

- Architected two LangGraph specialist agents with strict-schema handoffs, conditional routing, post-generation validation, and checkpointed Slack human-in-the-loop interrupts.
- Implemented corrective RAG with script-routed dense and bilingual BM25 retrieval, Reciprocal Rank Fusion, cross-encoder reranking, answerability gating, and citation validation.
- Wrote a fully async, schema-compatible MongoDB checkpointer to keep the FastAPI request path free of synchronous thread-pool detours.
- Designed stable Qdrant chunk identities, section-aware splitting, and content status flags for downranking or blocking without re-embedding.
- Served Qwen3-VL with two hot-swappable LoRA adapters plus embeddings on one 16 GB GPU using vLLM multi-LoRA, with Prometheus and Grafana monitoring.
- Built the training-data flywheel: an LLM judge grades production turns, human corrections generate validated synthetic conversations, and a replay gate compares challenger against champion before release.
- Integrated Instagram and Facebook messaging behind HMAC-verified webhooks with comment triage, OTP verification, and background processing; currently in Meta app review.

Riddhi Siddhi — AI-assisted commerce apps

Flutter · On-device/cloud LLMs · Tool calling · Supabase

- Built customer and business Flutter apps with local/cloud AI tool calling for product lookup, cart actions, disambiguation, and handwritten-list scanning.
- Solved catalog cold start by accepting requests for unlisted products and surfacing them to staff for incremental catalog creation.
- Delivered catalog, staff, order, barcode, and sales workflows; cleared Google Play review and entered internal testing.

GemMate — mental wellbeing companion

- Designed the Supabase/PostgreSQL schema and row-level security protecting each user's private family-and-friends circle.
- Built the Twilio/SendGrid escalation engine and Next.js landing and admin experience with monitoring and blog CRUD.

Thinkhome — catalog extraction and publishing platform

- Built parallel page- and product-level PDF catalog extraction with editable review workflows that gate static publication.
- Delivered an optimized PDF viewer with separate page and product views so administrators can inspect and correct extracted content before publishing.
- Updated and cleaned up the existing Next.js and Strapi architecture for new requirements rather than layering patches onto the old structure.

ChristmasPulse — seasonal mobile application

- Built React Native screens including image-overlay composition and set up the Supabase/PostgreSQL backend.
- Added OpenAI image generation and captioning, managed the client relationship, and owned the Google Play release through publication.

Skills

AI: LangGraph, LangChain, tool calling, multi-agent systems, human-in-the-loop, MCP, evaluation, guardrails, RAG, hybrid retrieval, reranking, Qdrant, LoRA, vLLM, Hugging Face, on-device inference

Languages and backend: Python, TypeScript, JavaScript, Dart, SQL, FastAPI, Node.js, webhooks, WebSockets, MongoDB, PostgreSQL, Supabase, Redis, Celery, Twilio, SendGrid

Frontend and mobile: React, Next.js, React Native, Flutter, Tailwind CSS | **Cloud and ops:** AWS, nginx, Docker, Vercel, Prometheus, Grafana

Education

B.Tech, Computer / Software Engineering

Uka Tarsadia University

Sep 2021 – Aug 2025

Surat, Gujarat