

Ruchir Kalokhe

Pune, India · +91 85306 62440 · ruchirkalokhe@gmail.com · ruchir.dev · github.com/ruxir-ig · linkedin.com/in/ruchirkalokhe

Engineer working across backend systems and ML infrastructure: multi-provider GPU inference, model serving, retrieval and ranking, and the billing, security, and reliability layers that keep them in production. Comfortable owning a feature from runtime container to API to evaluation harness.

EXPERIENCE

Gen AI Intern, Backend and ML | In2peta Services Pvt. Ltd.

Remote · Dec 2025 – Present

TypeScript, Fastify, PostgreSQL, Prisma, PgBoss, Docker

- Built the GPU platform backend in **TypeScript** and **Fastify**, enabling multi-provider inference across Modal, Koyeb, and RunPod with unified runtime contracts, PgBoss job orchestration on PostgreSQL, provider routing, live endpoint resolution, and failover so long-running jobs keep serving when a provider drops.
- Automated Hugging Face model deployment (Dockerfiles, handlers, build workers, Prisma catalog) so models go from registry to a live endpoint without manual packaging.
- Integrated **LTX-2** and **LTX-2.3 Distilled** video-generation runtimes with custom handlers, image conditioning, and duration controls up to 60 seconds across Modal, Koyeb, and Docker.
- Implemented usage-based billing and credit reservations across token, character, second, and image modalities, with Razorpay payments, invoicing, and idempotent Prisma/PostgreSQL workflows that block overspend before inference runs.
- Hardened Fastify APIs with scoped keys, plan-aware access control, webhook verification, authenticated uploads, and SSRF/private-URL blocking; patched LiteLLM for Fireworks cached-token pricing with regression coverage.
- Owned model lifecycle from catalog registration through container build, provider deploy, live invocation, and usage accounting; diagnosed production failures across API, runtime, and inference.

PROJECTS

Auscult | Privacy-preserving observability for healthcare AI agents

Python, Presidio, spaCy

- Made agent runs auditable without retaining raw PHI: prompts, tool calls, outputs, and errors are sanitized with Presidio/spaCy and deterministic Faker replacements **before** database write.
- Fail-closed capture so sanitizer or queue failures never silently persist unsanitized text; JSON-safe redaction, background writes, replay/export/purge, Alembic migrations, and a PHI eval harness under CI.
- Shipped SDK wrappers and LangChain callbacks so apps get traces without changing their core call path, plus a CLI for inspecting and comparing sanitized runs.

Candis | LLM candidate discovery and ranking over 100K profiles

Python, NumPy, embeddings, evals

- Ranked top candidates from **100K profiles** in **13 s** on CPU (NumPy path) via honeypot filters, structured fit scoring, availability weighting, cached LLM reranking, and evidence-guided expansion.
- Beat keyword stuffing: **15** injected weak profiles entered the top-100 under keyword matching, **0** under Candis; pairwise LLM audit **83.7% win rate**, hand-qrel NDCG@10 **0.93+**, prompt-injection audit with **0** effect on final ranking.

MuseTalk API | GPU inference service for real-time lip synchronization

FastAPI, Docker, CUDA

- Turned a research video model into a production GPU FastAPI service with Docker deploy, health checks, model setup, image-driven inference, and chunked downloads for large outputs.
- Cut GFPGAN enhancement time by **1.8×** (e.g. **25–28 min** → **15 min** on RTX 4060) by removing redundant face detection; fixed OpenMMLab/MMPose and Docker build-isolation failures.

TraceLink | Manufacturing supply-chain traceability platform

FastAPI, React, Docker

- Shipped a live factory-floor platform (Render + CI) for intake → batches → QC → dispatch → complaints, with forward/reverse tracing and blast-radius / financial-exposure analysis in milliseconds.
- Ingests **40k+ rows in seconds** via batch inserts and SHA-256 duplicate detection; operator UX for searchable traces, notification history, and Firebase-authenticated dashboards.

More: SAR Colorization, AssetFlow · github.com/ruxir-ig

TECHNICAL SKILLS

- **Languages** Python, TypeScript, SQL, C
- **Machine Learning** PyTorch, TensorFlow, Hugging Face, Diffusers, CUDA, scikit-learn, Presidio, spaCy
- **Backend** FastAPI, Fastify, Node.js, REST APIs, PgBoss, asynchronous workers, SDK design
- **Infrastructure** Docker, PostgreSQL, Prisma, Modal, RunPod, Koyeb, GitHub Actions, Linux

EDUCATION

B.E., Artificial Intelligence and Data Science | Savitribai Phule Pune University · PES MCOE Pune, India · Aug 2023 – Jun 2027

- **Open source & research** Contributed 24-bit truecolor to *ascii-view* (C); **maintainer** of *nitch*; GAN SAR-to-RGB colorization in PyTorch (top 25 internal SIH); BB84 QKD simulator.
- **Community** Member of Google Developer Group on Campus (AIML) - organised IMACE 2026 hackathon and events, and served on the judging jury.