

GAURAV PHADALE

+91-9769275195 | garyphadale@gmail.com | [linkedin.com/in/gaurav-phadale-961388275](https://www.linkedin.com/in/gaurav-phadale-961388275) | github.com/gauravgoodreads | gauravphadale.me

EDUCATION

Mukesh Patel School of Technology Management & Engineering, NMIMS

B.Tech in Computer Engineering; CGPA: 3.59/4.0

Aug. 2023 - May 2027

Mumbai, India

Pace Junior Science College, Mumbai

HSC (Class XII), Science; 70.5%

2021 - 2023

I.E.S Secondary School, Mumbai

SSC (Class X); 92.6%

2021

EXPERIENCE

Full Stack & Agentic AI Development Intern

Mentoria | Project Arya (AI-Powered Job Search Platform)

Jan. 2026 - Present

Mumbai, India

- Grew AI resume scoring and job-matching to **27,000+ users** on **arya.mentoria.com** by shipping semantic search over **PostgreSQL + pgvector** embeddings and **OpenAI APIs**, delivered in two-week sprints alongside product and design
- Held p95 latency at **760ms** for **50 concurrent users** by moving resume parsing, AI inference, and email automation off the request path into **Redis-backed Celery/RQ** pipelines, with **pytest** coverage gating every release
- Built and version-controlled (**Git**, feature-branch workflow, PR review) the **FastAPI REST APIs** behind authentication, application tracking, and Gmail workflows secured with **OAuth2** and **JWT**, plus the **React/TypeScript** dashboard consuming them
- Cut recurring hosting spend to zero by migrating **50+ client sites** to **GitHub Pages + Cloudflare Workers** with automated **GitHub Actions** deploys, and onboarded non-technical clients onto **Sanity Studio** so they could publish edits without filing a dev ticket
- Shipped a **Chrome extension (MV3)** for job-application autofill and integrated **Razorpay** payments with **role-based access control** for enterprise accounts, coordinating rollout directly with client stakeholders

Full Stack Developer Intern

SkyHigh Travels

Jan. 2025 - Jun. 2025

Mumbai, India

- Cut average query latency **35%** by redesigning the customer, itinerary, and invoice schema and eliminating **N+1** access patterns through JOIN restructuring and targeted indexing, all tracked through **Git** with peer code review
- Shifted operations pricing from intuition-based to data-driven by building a **SQL** analytics module aggregating booking and revenue data by month, segment, and route
- Delivered production features across a **250-hour** engagement through weekly stand-ups, translating ambiguous requirements from non-technical stakeholders into scoped, shipped functionality

Student Research Intern | SRIP (0.90 AUC, 1,166-Patient Cohort)

NMIMS MPSTME, under Dr. Mahesh Patil

2024

Mumbai, India

- Built the full ML pipeline (imputation, normalization, outlier handling, **PCA**) over a published clinical cohort of **1,166 tuberculosis patients**
- Beat Logistic Regression and Random Forest baselines with **SVM (0.90 AUC, 87% accuracy)**, isolating alcohol use as the strongest severity predictor across every model; co-authored the resulting paper

PROJECTS

InferGate | Distributed LLM Inference Gateway | FastAPI, Redis, Docker, Nginx, Prometheus | GitHub

2026

- Halved p95 latency (**394ms to 188ms**) and removed **97.9%** of upstream provider calls against a no-cache control, measured over **2,000 requests** at **50 concurrent users** (median of 3 runs)
- Designed a two-tier response cache pairing **SHA-256** exact matching with cosine-similarity semantic matching, calibrating the threshold to **95% recall** with **zero false positives** across **330 labeled prompt pairs**
- Collapsed **50 concurrent identical requests** into one upstream call with Redis-backed **single-flight coalescing**, and kept every request served through a total primary-provider outage via **circuit breakers** with ordered failover

CampusCopilot | Agentic RAG Platform | FastAPI, PostgreSQL (pgvector), Docker, LLM tool-calling | GitHub

2026

- Lifted Hit@5 from **66.4%** to **88.2%** and MRR from **0.50** to **0.68** on a **110-question benchmark** by fusing BM25 with dense retrieval through **reciprocal rank fusion**
- Indexed **4,661 chunks** from **70 research PDFs** behind an **HNSW pgvector** index, rejecting **28%** of candidate benchmark questions through overlap and answer-leak filters so retrieval scores could not be inflated
- Grounded answers in page-level citations via LLM tool-calling, with a calibrated confidence threshold that abstains on **90.5%** of out-of-scope questions; covered by **106 tests** across both storage backends

Conflux | Real-Time Industrial Safety Risk Engine | FastAPI, MQTT, Neo4j, TimescaleDB, React, Docker |

2026

GitHub

- Caught hazard combinations that single-threshold alarms miss by rescoring every plant zone on a **10-second tick**, multiplying co-occurring weak signals from **five specialist agents** instead of alerting on each independently; pinned by **44 tests**
- Streamed MQTT telemetry to a React dashboard over **WebSockets**, modeling permit-to-zone conflicts in **Neo4j** and sensor history in **TimescaleDB**, and mined code-mixed Hinglish shift logs for equipment-bypass language

TECHNICAL SKILLS

Languages & Backend: Python, TypeScript, JavaScript, SQL, Java, C++ · FastAPI, Node.js, Express.js, Flask, REST APIs, OAuth, JWT

AI/ML: RAG, LLM tool-calling, pgvector retrieval, OpenAI API, scikit-learn (RF, SVM, LR), PyTorch (LSTM)

Infra & Practices: Docker, Redis, PostgreSQL, MongoDB, AWS (S3, Lambda, EC2), CI/CD (GitHub Actions), Git, Agile/Scrum, pytest

Frontend & Core CS: React, Tailwind CSS, Zustand, Chrome Extensions (MV3) · DSA, OOP, DBMS, System Design

CERTIFICATIONS & ACHIEVEMENTS

AWS Certified Cloud Practitioner (CLF-C02), Amazon Web Services, 2025

Top 100 of 1,000+ teams, ZS Campus Beats 2026 National Hackathon, ZS Associates

Research Paper Presenter, TCSC National Conference | Analyzing the Impact of Alcohol Use on Tuberculosis Severity, Dec. 2024