

VIPUL PONUGOTI

AI Engineer Intern · LLM Systems & RAG · Agentic Workflows · Model Evaluation & Safety · Backend & Full-Stack AI
Vadodara, Gujarat, India | vipulponugoti@gmail.com | [LinkedIn](#) | [GitHub](#) | [LeetCode \(500+\)](#) | [GSSoC '26 Rank 15](#)

PROFESSIONAL SUMMARY

Production-focused AI Engineer and final-year B.Tech CSE (AI & ML) student with hands-on experience building RAG platforms, agentic and MCP-based systems, LLM evaluation harnesses, and self-hosted inference APIs — shipping from architecture through Docker/Kubernetes deployment. Currently an SDE Intern at VibeLevel.ai and a GSSoC '26 contributor ranked 15 of 47,951 with 870+ merged pull requests across 40 projects spanning Node.js, Python/FastAPI, Rust, Dart/Flutter, and Solidity codebases. Guinness World Records participant and 2× National Abacus Gold Topper. Deep commitment to responsible AI: every system ships with hallucination detection, constitutional filtering, reproducible evaluation, and quality-gated CI/CD.

TECHNICAL SKILLS

Languages: Python, Java, TypeScript, JavaScript, SQL

AI & LLM Engineering: RAG (BM25 + dense hybrid, pgvector, FAISS, Pinecone, LlamaIndex, reranking), Agentic Workflows (LangGraph, AutoGen), Model Context Protocol (MCP), Prompt Engineering, RAGAS Evaluation, Constitutional AI, RLHF concepts, Hallucination Detection, Red-Teaming

ML Frameworks: PyTorch (primary), Hugging Face PEFT/Transformers, TensorFlow, Scikit-learn, Pandas, LoRA/QLoRA Fine-Tuning, Phi-4, Llama-3, Statistical Modeling, Feature Engineering

Backend & APIs: Java, Spring Boot, Spring AI, FastAPI, Flask, Node.js/Express, REST/WebSocket/SSE APIs, JPA/JDBC, SQLAlchemy, PostgreSQL (Neon, Drizzle ORM), MongoDB, Redis, Apache Kafka

MLOps, Infra & Testing: Docker, Docker Compose, Kubernetes, Fly.io, GitHub Actions CI/CD, AWS (EC2, S3, Bedrock), Azure, Oracle Cloud Infrastructure, vLLM, Ollama, LangSmith, Weights & Biases, MLflow, OpenTelemetry, Prometheus, Grafana, Pytest, Playwright, Vitest, Git/GitHub, Linux

Frontend: React, Next.js (App Router), TypeScript, Streamlit, Vercel

WORK EXPERIENCE

SDE Intern · VibeLevel.ai · Remote Jul 2026 – Present

AI-native skills & talent platform (practice assessments, vibeathons, hiring funnel) — Next.js App Router on Vercel, FastAPI on Fly.io, Neon Postgres/Drizzle, E2B code sandboxes, MCP server; ~24 merged commits in a team codebase.

- Designed and built Aura Scoring v2, an LLM-based human–AI collaboration scoring engine (spec → multi-phase pipeline: attribution analysis, human-contribution corroboration, substance ceiling, hypothesis provenance) that corrected agency misattribution in AI-assisted sessions, where v1 scored a passive delegator and a strategist identically; shipped with ~2,700 lines of pytest coverage across 12+ test files.
- Built knowledge-graph-grounded session feedback: extracted workspace context and grounded LLM scoring insights in a product feature graph via a custom PFG client.
- Shipped a full-stack workspace signals feature aggregating agent-reported frameworks into user profiles with GitHub repo derivation from session tags (FastAPI + Next.js/TypeScript/React).
- Hardened the MCP server's contracts, evidence verification, and client compatibility; added a read-only scoring-service status endpoint and closed floating-point gaps in score-band lookup.
- Virtualized the hiring candidate table with rAF-based row windowing (with tests) and fixed the auth layer to distinguish database outages from unauthenticated state, eliminating spurious logouts.

ML Engineer Trainee Intern · Kshemam Health Solutions · Remote Feb 2026 – Jun 2026

- Owned end-to-end predictive risk classification pipeline on patient-monitoring data: privacy-compliant feature engineering, Scikit-learn model evaluation, and precision/recall tuning for real-time clinical-risk thresholds.
- Built Java/Spring Boot REST endpoints and scheduled batch pipelines for automated AI analytics; contributed to data-ingestion and model-serving workflows with attention to reproducibility, security, and inference latency.
- Established model performance baselines, latency benchmarks, and structured error-analysis reports — introducing systematic tracking where none previously existed.

Machine Learning Intern · Cognifyz Technologies · Remote Dec 2025 – Jan 2026

- Completed applied data-analysis and supervised-learning tasks in Python, building repeatable preprocessing, training, and evaluation workflows with Pandas and scikit-learn.

PROJECTS

AI Safety & Constitutional Evaluation Harness

Python · PyTorch · FastAPI · Constitutional AI · Red-Teaming · Docker

- Built red-teaming framework for adversarial prompt testing across 6 LLM providers; implemented constitutional filtering pipeline classifying outputs across safety, helpfulness, and harmlessness axes with structured per-category scoring.
- Automated detection of prompt injection, jailbreak patterns, and policy-violating outputs; generated audit reports used to harden system prompts — reducing adversarial success rate by 34% across 1,200 red-team probes.

Enterprise Agentic RAG Platform

Spring Boot · Python · LangGraph · pgvector · LlamaIndex · Redis · Docker · React

- Architected multi-source RAG platform (PDFs, web, structured data) with BM25 + dense hybrid retrieval and cross-encoder reranking over pgvector; achieved 91% RAGAS faithfulness and 87% context precision@5 on internal benchmark.
- Built Spring Boot microservices (JWT auth, RBAC, document management, chat sessions) and Python AI services (adaptive chunking, embedding pipeline, context-aware generation); full-stack ownership from API contract to AI response layer.

- Designed LangGraph agent workflows for multi-hop document comparison, iterative query refinement, and citation-level source verification; implemented responsible-AI guardrails with confidence scoring on every generated response.

MCP-Powered AI Project & Community Assistant

Python · FastAPI · LangGraph · MCP · FAISS · Ollama · Docker

- Architected a local-first multi-agent developer assistant for GitHub analysis, document question answering, web research, and issue workflows using explicit tool boundaries and deterministic routing.
- Implemented REST and streaming (SSE) APIs, FAISS retrieval, Ollama inference, Docker Compose deployment, safe GitHub write controls, configuration validation, and automated tests.

Temporal UI Evidence Graphs — LLM Failure-Diagnosis Benchmark

TypeScript · Node.js · Playwright · React

- Engineered a browser-instrumentation and benchmark pipeline converting console, network, route, DOM, layout, and accessibility events into typed temporal evidence graphs for UI failure diagnosis.
- Ran 3 experiments across 50 controlled scenarios, 3 models, and 10 evidence conditions; produced 2,700 stored diagnoses, verified 10,643 immutable artifacts, and passed 9/9 offline reproducibility checks — the temporal-graph method ranked first in 9 of 10 bug categories.

LLM Evaluation & Observability Platform

Python · FastAPI · Spring Boot · OpenAI · Azure OpenAI · PostgreSQL · LangSmith · Docker

- Automated eval harness across 20+ dataset × model × retriever × chunking configurations; tracked RAGAS faithfulness, context recall, citation accuracy, hallucination rate, and p95 latency per run with LangSmith + W&B dashboards.
- Implemented CI quality gates blocking releases below configurable score thresholds; caught 9 regressions across 4 weeks of sprint iterations before any reached production.

Self-Hosted LLM Inference & Deployment System

Llama-3 · Phi-4 · PyTorch · vLLM · LoRA/QLoRA · Kubernetes · Prometheus · Grafana

- Deployed vLLM-backed inference API on Kubernetes with streaming, batching, per-tenant rate limiting, and token metering; sustained 98.7% uptime under 50 concurrent users in a 72-hour load test.
- QLoRA-fine-tuned Llama-3 8B and Phi-4 on domain datasets; self-hosted inference matched managed API quality at 61% lower cost and 38% lower p95 latency on benchmarked workloads.

Distributed Event-Driven Task & Workflow Engine

Spring Boot · Kafka · Redis · PostgreSQL · Docker · OpenTelemetry

- Building a Kafka-backed distributed task scheduler with idempotency keys and transactional PostgreSQL job claims, coordinating multiple Spring Boot workers without duplicate execution during retries or node failure.
- Implementing dead-letter queues, exponential-backoff retries, circuit breakers, Redis token-bucket rate limiting, and OpenTelemetry/Prometheus telemetry for task latency, throughput, retries, and failures.

OPEN SOURCE

- GSSoC '26 Contributor (GirlScript Foundation, India · May 2026 – Present): Ranked 15 / 47,951 (top 0.03%) with 90,524 pts across 876 merged PRs in 40 projects; AI/Agents & Open Source tracks. Peak throughput of 200 merged PRs in a single week; contributed across Node.js, Python/FastAPI, Rust, Dart/Flutter, and Solidity codebases.
- ai-code-reviewer (multi-agent LLM code review platform): 204 merged PRs across backend, VS Code extension, GitHub Action, and AI engine — atomic circuit-breaker HALF_OPEN admission, symlink TOCTOU fix via O_NOFOLLOW, blocking Redis KEYS → SCAN, safe YAML schema parsing, secret-scrubber and redirect-cap hardening, non-root containers, removal of committed dev credentials, and CI test-artifact repairs.
- Truxify (freight logistics platform): 351 merged PRs — server-side freight pricing, JWT/WebSocket authorization, HMAC webhook verification on raw body, path-traversal protection on ML model endpoints, WASM env-var secret filtering, PromQL input sanitization, hardcoded-credential removal, transaction commit/rollback correctness, TOCTOU race fix in request deduplication, and a Vitest unit-test framework rated critical difficulty / exceptional quality by maintainers.
- MLOps Red Wine Quality Prediction: Enforced quality gate on model promotion, canonical stable-model rollback, surfaced real MLflow failures instead of mock runs, corrected feature-drift ratio, and hardened input validation and error handling across prediction endpoints.
- sahidawa-india, The-Lighthouse, peer-learning, nyay-setu-working, story-spark-ai, commitpulse: Retry-and-throw on stale alert deactivation, pg_cron monitor initialization fix, duplicate-review prevention, timing-safe JWT signature comparison (CVE-class), strict CSP headers, X-Forwarded-For rate-limiter spoofing fix, and input validation + rate limiting on public API endpoints.

EDUCATION

B.Tech Computer Science & Engineering (AI & ML) · Parul Institute of Engineering and Technology, Vadodara 2023 – 2027

CGPA: 7.53/10 · Dean's List 2024–25 · Coursework: Machine Learning, Deep Learning, NLP, Probability & Statistics, Data Structures & Algorithms, Database Systems, Distributed Systems, Operating Systems, Computer Networks

CERTIFICATIONS & ACHIEVEMENTS

- Oracle Cloud Infrastructure 2025: Generative AI Professional · Data Science Professional · AI Foundations Associate
- Advanced Coursework: Machine Learning Specialization (DeepLearning.AI & Stanford Online) · MIT 6.S191 Introduction to Deep Learning
- WISER Quantum Program: Selected among the top 3,000 of 65,000 participants
- Guinness World Records: Official participant — Largest Mental Arithmetic Lesson (3,189 participants), Chennai, Oct 2018
- Gold Topper (2×): 34th & 37th National Abacus Championships, Brainobrainfest 2017–2018