# SATYABHAN UMED BHADORIYA

Senior AI Engineer | Pune, India | +91 79727 85226 | satyabhan.bhadoriya@gmail.com
satyabhan007.github.io | GitHub | LinkedIn | Open To Work • Hybrid/Pune

## PROFESSIONAL SUMMARY

**Senior AI Engineer | Production GenAI, LLMs & Agents | 8+ YOE (Enterprise & AI)**

Creator of CLINIRISE, a production-ready healthcare WhatsApp AI CRM System powered by Google Gemini, Qdrant, Redis, and GCP Cloud Run. Specialist in Multi-Agent System design, RAG architecture, and AI/GenAI model optimization. Experienced with LangChain, Vertex AI, Groq, AI SDLC tools, and enterprise infrastructure monitoring, as well as authoring over 400 chapters of CI-validated open-source engineering curricula. Adept at applying TypeSafe Jev typed decision models for guardrail enforcement in agentic workflows.

## TECHNICAL SKILLS

**AI & Systems Engineering:** LLM Deployment, GenAI, RAG Architecture, Tool/Function Calling, Prompt Engineering, and Multi-Agent Systems

**Frameworks & Libraries:** PyTorch, Hugging Face Transformers, FastAPI, LangChain, LangGraph, and TypeSafe Jev

**AI Model Optimization:** LoRA/DPO Fine-Tuning, INT8 Quantization, vLLM Serving, and KV-Cache Optimization

**AI Platforms & Inference:** GCP Vertex AI (Gemini), Groq (Fast Inference), Ollama, Azure STT, and ElevenLabs

**Vector, Data & Cloud:** Qdrant (Vector DB), Redis, PostgreSQL, Firebase/Firestore, GCP Cloud Run

**DevOps & Infra:** Linux/RHEL 7, Docker, Terraform, Ansible, Prometheus, Grafana, GitHub Actions, VMware, OpenSCAP, STIG Hardening

**Languages:** Python (Primary), TypeScript, Node.js

## FEATURED PROJECT

**CLINIRISE** | ***Production Multi-Agent Healthcare CRM & RAG System***

Solo-built, end-to-end multi-clinic CRM currently in pilot testing with 2 live clients. (Live App: gen-lang-client-0155046437.web.app)

- **Multi-Agent System Design:** Architected a supervisor-worker multi-agent topology using LangGraph-style orchestration (intent-router, scheduling, FAQ RAG, escalation agents). Implemented typed tool calls, shared agent state, and confidence-gated handoffs.
- **RAG Architecture & Implementation:** Engineered hybrid retrieval via Qdrant vector search and two-tier Redis caching to achieve sub-2 second latency. Built semantic intent routing to separate deterministic actions from generative queries, reducing LLM token consumption by ~40%.
- **AI Model Optimization:** Developed multi-tier model resiliency (Gemini Flash → Flash-Lite → template fallback) integrating INT8 quantization, KV-cache optimization, and continuous batching.
- **AgentOps & Responsible AI:** Implemented clinical guardrails including PII redaction (DPDP Act compliance), consent management, and human handoff escalation. Applied TypeSafe Jev (Choice/Score/Boolean) for fast, low-cost verification in place of full LLM calls.
- **Cloud & DevOps:** Built with React 18, TypeScript, Vite, Node.js/Express, GCP Cloud Run. Validated with 860+ automated tests (Vitest, Cypress E2E, k6).

## PROFESSIONAL EXPERIENCE

**Independent AI Engineer (Self-Directed)** — May 2026 – Present | Pune, India

- Lead active development and pilot management of CLINIRISE for 2 live clinic clients, continuously iterating on core RAG quality and prompt reliability.
- Architect and deploy comprehensive open-source engineering curricula, building full-scale labs for eBPF observability, vLLM model serving, and AI threat defense.

**Technical Support Engineer** | ***Aman Security India Pvt. Ltd.*** — March 2024 – May 2026 | Pune, India

- **AI Product Development:** Architected and independently built CLINIRISE (outside of core support duties) applying GenAI, Python, RAG, and Vertex AI.
- **AI SDLC & Voice Prototyping:** Utilized Claude Code and Antigravity IDE for an AI-assisted development workflow, setting up CI/CD pipelines via GitHub Actions.
- **Infrastructure & Monitoring:** Managed virtualized VMware environments, monitoring infrastructure health and service uptime using Prometheus and Grafana dashboards.
- **Technical Operations:** Provided Tier-level technical support and issue resolution using ServiceNow, Jira, and remote access diagnostic tools.

**Linux Administrator** | ***Oceanic EOPS LLP*** — 2023 – 2024 | Pune, India

- Managed RHEL 7 production servers, ensuring high availability and secure infrastructure.
- Applied DISA STIG hardening, executed OpenSCAP compliance auditing, and configured strict SELinux policies.

**Application Engineer** | *Precision Automation* 2018 – 2023 | Pune, India

- Delivered end-to-end automation engineering across ~15 major industrial projects, including assembly lines, packaging, and material handling systems.
- Programmed advanced HMI/PLC logic (GX Works / Mitsubishi), consistently maintaining 99.8% operational uptime across deployed systems.

## OPEN-SOURCE ENGINEERING & BENCHMARKING

**Author & Maintainer of CI-Validated Engineering Curricula (satyabhan007.github.io)**

- **AI-ML & AI-Lab:** Authored a 9-course curriculum covering scalar autograd, BPE tokenizers, attention mechanisms, RAG + LangGraph/ReAct agent loops, LoRA/DPO fine-tuning, INT8 quantization, KV-cache, and vLLM serving. Maintains 69/69 passing CI checks.
- **AI Threat Defense:** Developed production-grade training materials for AI Threat Detection, Adversarial Robustness, and Go Gateway Engineering.
- **DevOps-Infra:** Built extensive SRE playbooks and infrastructure code utilizing Terraform IaC and Docker multi-stage hardened builds, covering SLI/SLOs, golden signals, and incident runbooks secured by a 4-gate CI pipeline.
- **Linux-HPC-Security:** Created comprehensive systems engineering modules detailing Linux internals, hardening, DISA STIGs, Ansible, Slurm HPC, and featuring an advanced eBPF Observability Capstone.

## EDUCATION

**Bachelor of Engineering (B.E.), Mechanical Engineering — First Class** 2014

Savitribai Phule Pune University — Dhole Patil College of Engineering (DPCOE)