

Hill Patel

+91 97232 88714 | hillaniljppatel@gmail.com | LinkedIn | GitHub
Pune, Maharashtra | hillpatel.tech

SUMMARY

AI Engineer with 1.5 years of experience building agentic AI systems in production, specializing in multi-agent orchestration, deterministic tool-use boundaries, and human-in-the-loop governance for enterprise workflows. Currently architecting orchestration infrastructure at CloudRedux, alongside independent open-source work in long-horizon agent recovery (Cairn) and persistent developer memory systems (memex).

TECHNICAL SKILLS

LLMs & Retrieval: RAG, Hybrid/Vector Search, Re-ranking, LangChain/LangGraph, Multi-Agent Orchestration, Google ADK, Claude Code, Memory & Storage Management, Weaviate.

Model Optimization: Quantization (1–8 bit), Pruning, Distillation, ONNX Runtime, TensorRT, CUDA tuning.

ML Engineering: CNNs, Transformers, Multimodal Models, Evaluation (AUROC, F1, hit@K, latency, cost).

System Design: Distributed Systems, Microservices Architecture, Event-Driven Design, Async Processing, API Design, RBAC & Row-Level Security, Scalability & High Availability.

Backend & MLOps: FastAPI, Docker, APScheduler, CI/CD (GitHub Actions), Redis, PostgreSQL, Neo4j, Observability (OTel, Prometheus), GCP, AWS, SQL.

Languages & Tools: Python, PyTorch, TensorFlow, NumPy, Pandas, OpenCV, Gradio, Git, Linux.

PROJECTS

Cairn Jun 2026
Recoverable Long-Horizon AI Agents Framework *GitHub*
Framework-agnostic recovery harness and live benchmark testing "Checkpoints Are Compactions" via Re-grounding Recovery.
v1.0 held pending a powered live-LLM study to validate recovery claims.

PRISM — Madhuban Present
E-Invoicing & Payment Reconciliation Platform *prism.madhuban.com*
Built cron-job pipelines for automated e-invoicing and payment reconciliation at CloudRedux.
Automated multi-source ledger matching, exception flagging, and scheduled invoice dispatch at scale.

memex (PyPI, npm) May 2026
Persistent Memory for AI Coding Agents via MCP *GitHub*
Bitemporal knowledge graph of the codebase served via MCP to Claude Code, Cursor, and Gemini CLI.
12 MCP tools with hierarchical clusters and two-regime confidence decay, published on PyPI and npm.

EXPERIENCE

AI Engineer May 2026 – Present
CloudRedux *Pune, India*
Architecting a meta-level harness pattern, generalizing execution, validation, and recovery logic across agentic workflows.
Driving loop engineering for multi-agent systems, structuring orchestrator-critique-executor cycles with human-in-the-loop control.
Designed PRISM's read-only agent boundary using a 42-entry default-DENY tool allowlist, gating all write actions behind deterministic Python.
Automated 12 scheduled compliance and billing jobs (GST/TDS, escalation, reconciliation) across 3 legal entities and 10+ properties.
Contributing to FosterX PaaS orchestration engine (v2) - workload scheduling, service lifecycle management, and multi-tenant deployment abstractions.

AI Engineer Intern Feb 2025 – Apr 2026
CloudRedux *Pune, India*
Designed and implemented stateful agentic workflows using Vertex AI and Gemini models for enterprise use cases.
Built memory-aware AI systems with structured reasoning, vendor evaluation logic, and human-in-the-loop control.

EDUCATION

NIMS University Rajasthan
B.Tech in Artificial Intelligence & Machine Learning (CGPA: 8.422)

Sept 2022 – May 2026

CERTIFICATIONS & ACHIEVEMENTS

- Published a peer-reviewed research paper: *Transforming Urban Solutions for Smart Cities through Crowdsourced Feedback* (Mar 2025).
- Completed the **RAG and Agentic AI Professional Certificate** - Coursera.
- Open-source impact: **imgshape** (8.84K+ downloads) and **memex-mcp** (5K+ downloads) on PyPI.