

HARSH NANAVATI

Applied AI Engineer | LLMs · RAG · Agentic Systems · Evaluation

nanavatiharsh00@gmail.com · [linkedin.com/in/harshnanavati](https://www.linkedin.com/in/harshnanavati)

Ahmedabad, India (IST, UTC+5:30) · Remote · 4+ hrs overlap with EU & US-East · Open to full-time, contract or EOR

SUMMARY

Applied AI engineer with 2+ years shipping production LLM systems end to end — enterprise RAG, agentic pipelines with automated verification, real-time self-hosted LLM services and GPU computer vision — for clients in the US and India. Systems serve **600+ users** and cut manual work by up to **~70%**.

SKILLS

GenAI: LLMs, RAG, agentic workflows, LangGraph, LangChain, OpenAI API, Ollama (self-hosted LLMs), structured outputs, prompt engineering, guardrails, verification harnesses, human-in-the-loop

Retrieval & Evaluation: FAISS, vector search, OpenAI embeddings, Docling, chunking strategies, reranking, citation grounding, RAG evaluation (RAGAS)

ML / CV: PyTorch, scikit-learn, FaceNet/MTCNN, YOLOv8, ONNX, Elo & probabilistic models, backtesting

Deployment & MLOps: Docker, Docker Compose, Nginx, Prometheus & Grafana, CI/CD (GitHub Actions)

Engineering: Python, FastAPI, asyncio, REST APIs, SQL Server, PostgreSQL, MongoDB, Redis, Spark/Scala, Git, Angular

EXPERIENCE

SDE – Artificial Intelligence (Applied AI Engineer) | Masterly Solutions Pvt. Ltd.

Oct 2024 – Present

Ahmedabad, India · Building in-house AI products and client systems for the US and India

- **DocuPilot – Enterprise RAG Platform:** cut manual SOP queries by **~40%** for **600+ staff** across 14 document categories — Docling ingestion (OCR + tables), custom structure-aware chunking with page-level provenance, and text-embedding-3-small + FAISS retrieval over **12,000+ chunks** at **sub-100 ms** median latency.
- **DocuPilot – Answer Grounding & Guardrails:** GPT-4o-mini answers constrained by answer-grounded citation selection and a two-stage citation gate, with every claim linked to its source page region; role-based document access, JWT auth, SHA-256 dedup and per-query observability (latency, citations, feedback) on FastAPI + SQL Server.
- **CrickPal Intelligence Layer – AI Win-Prediction Platform:** LLM agent turns analysts' plain-English specs into feature code, gated by an automated verification harness (golden examples, symmetry, replay-based leakage, determinism) with self-repair and human approval; Elo baseline at **log-loss 0.643 / Brier 0.227** over 4.1k T20Is; **~800 automated tests**.
- **CrickPal AI Summary – Real-Time LLM Commentary Engine:** live ball-by-ball data → narrative and headline summaries on self-hosted Ollama; incremental summarization keeps prompts flat as a match grows; exactly-once batching on a single instance, a circuit breaker and a no-redeploy tuning console.
- **AI Face Matching System:** cut manual photo tagging by **~70%** with an unattended MTCNN + FaceNet pipeline — self-healing job queue, retry budgets, stale-lease recovery and embedding cache across three SQL Server databases; extended to video with YOLOv8 detection and ONNX/DirectML export.
- **QuantumCopilot – AI Loan & Financial Assistant:** maps natural-language queries to typed parameters across 6 loan products; parallel scenario fan-out (**~3x faster scenario stage**), multi-turn memory, constant-time API-key checks, and 8 Uvicorn workers on IIS.
- **RAG Support & Concierge Chatbots:** support bot with a human-in-the-loop correction loop (**~60% fewer escalations**); concierge bot redesigned from RAG to cached long-context with single-call intent routing.

Intern – Data Science & Big Data | Deloitte Haskins & Sells LLP, remote

Jun 2023 – May 2024

- Used Spark, Scala and SQL for entity-resolution and anomaly-detection analyses on large-scale financial datasets.
- Built Power BI dashboards for enterprise audit reporting workflows.

PROJECT & RESEARCH

ViperMind – Multi-Agent Python Tutor (personal project) — LangGraph router over four GPT-4 agents; code grades answers against an LLM-written key; Redis cache, PostgreSQL, 7-service Docker Compose stack with monitoring.

Blockchain-Driven Federated Learning for Parkinson's Detection (M.Tech research), with explainable AI.

EDUCATION

M.Tech, Data Science | Nirma University, Ahmedabad | CGPA 8.79/10

2022 – 2024

B.E., Computer Engineering | LDRP-ITR, Gandhinagar | CGPA 8.75/10

2017 – 2021