

NAMAN MANOCHA

AI ENGINEER • GENAI • AGENTIC SYSTEMS — LLMs, RAG & Applied ML

Kolkata, India | 8530356515 | namanmanocha42248@gmail.com | [LinkedIn](#) | [GitHub](#)

SUMMARY

AI/ML engineer with hands-on experience across classical ML, multi-agent LLM orchestration, and grounded RAG systems, with a track record of shipping AI products end-to-end. Solo-built **FundersAI**, a live research platform for Indian stocks and mutual funds with deterministic financial metrics, hybrid vector/lexical retrieval, and explainable ML clustering across 12 asset management companies — plus a decoupled report-synthesis microservice containerized and deployed on a self-managed Kubernetes (K3s) cluster on AWS. Currently building evidence-first AI products independently to gain hands-on production deployment experience. Comfortable end-to-end: Python/FastAPI, Next.js/React, Supabase/PostgreSQL, and cloud deployment (Vercel, Docker, Kubernetes). Also experienced translating business requirements into deployed systems as a technology lead.

TECHNICAL SKILLS

ML & AI: LLMs, RAG (LangGraph), Multi-Agent Orchestration (LangGraph, Google ADK), Hybrid Search, Prompt Engineering (few-shot, CoT), Hugging Face, Ollama, Vector Embeddings (pgvector), K-Means Clustering, Logistic Regression, AI Observability (Langfuse), MLflow, scikit-learn

Backend & Infra: Python, FastAPI, REST APIs, SQL, PostgreSQL, Supabase (Auth, RLS), Redis (Upstash), Docker, Kubernetes (K3s), GitHub Actions (CI/CD), Google Cloud Run, Vercel, Prefect, Git

Frontend & Data: Next.js, React, TypeScript, ETL Pipelines, Web Scraping, PDF Parsing, Data Validation, Data Visualization

EXPERIENCE

Technology Manager — PayGain Multiservices Pvt. Ltd.

Feb 2025 – Present

Kolkata, India

- Scoping and leading development of a company-wide mobile app as sole engineer, covering client/employee onboarding, document access, product registration, employee training, and payments — mobile counterpart to PayGain's core platform.
- Designed and deployed a lead-capture automation (Make) triggered on Google Form submissions, eliminating manual lead logging and improving sales outreach.
- Drove 30+ client conversions by building and managing the company website and digital discovery flows.
- Simplified complex investment/insurance products for 300+ seminar attendees via 20+ technical presentations and collateral — translating business requirements into deployed, client-facing systems.

INDEPENDENT PROJECTS

FundersAI — AI-Powered Research Platform for Indian Stocks & Mutual Funds

fundersai.co.in

Python, FastAPI, Next.js, Supabase/pgvector, LangGraph, Groq, OpenAI, Cloudflare R2, GitHub Actions

Solo-built, Apr–Aug 2026 — ~132K lines of code across 752 files, 410 commits, 142+ test suites — submitted to OpenAI Build Week | Live: www.fundersai.co.in | Repo: github.com/Naman6019/FundersAI

- **Citation-Grounded RAG:** built a 7-node LangGraph pipeline (retrieve → grade → synthesize/abstain, with a query-rewrite loop back into retrieval) over ~10,400 embedded document chunks (OpenAI embeddings, Supabase pgvector) with lexical reranking and strict LLM abstention, eliminating hallucinated financial claims when evidence is insufficient.
- **Explainable ML:** engineered a fund-similarity engine using cosine similarity and k-means clustering over 13 standardized features, plus a logistic-regression triage model (MLflow) for parser-review prioritization.
- **Multi-Agent Orchestration:** designed a custom multi-agent pipeline (Planner → Quant → Research → Synthesis agents) with shared state and real-time streamed status updates, where a Groq-hosted LLM router agent classifies query intent and directs execution across stock, mutual-fund, comparison, and market-news paths.
- **Data-Quality Debugging:** diagnosed and fixed a scheme-name normalizer bug that had collapsed 14,000+ mutual fund records into the wrong peer groups, raising field-level data coverage from as low as 83% to 100% on affected AMCs.
- **Resilient Data Engineering:** built a 22-workflow GitHub Actions pipeline scraping and normalizing official disclosures from 12 AMCs, reaching 99.66% data-readiness across 871 tracked fund schemes.
- **Cloud Storage & Billing:** built an S3-compatible storage layer (Cloudflare R2 via boto3/AWS SigV4, Python + TypeScript) for raw documents and cold archival, and shipped Razorpay subscription billing (webhook-verified, Free/Pro/Ultra) with Supabase Auth row-level security.

FundersAI Reports — Containerized LLM Report-Synthesis Microservice

synthesis.fundersai.co.in

Python, FastAPI, LangGraph, Docker, Kubernetes (K3s), AWS EC2

- **Decoupled Architecture:** extracted report synthesis into a standalone FastAPI + LangGraph service, independently deployable from the main monolith so LLM-heavy generation scales without impacting the core app.
- **Kubernetes Deployment:** provisioned and administered an AWS EC2 instance running K3s, deploying the service as a 2-replica Kubernetes Deployment behind a streaming API for real-time fund-comparison report generation.
- **Infrastructure & Secrets:** configured EC2 security groups/networking and managed credentials (Supabase, OpenAI) via Kubernetes Secrets rather than embedding them in the container image.

CareFlow Intelligence — Agentic RAG & Fine-Tuning for Clinical Research

Python, FastAPI, Next.js, Supabase/pgvector, Unsloth, Ollama | Solo-built, Aug 2026 — ~7,300 LOC — non-clinical, synthetic-data-only (MITRE Synthea)

- **Agentic Citation-Grounded RAG:** built a bounded-tool-loop document agent (list_documents/search_documents) with query reformulation and mandatory citation validation against Postgres full-text (GIN tsvector) indexes, backed by 4 automated ingestion pipelines (OpenFDA drug monographs, PubMed E-Utilities, ADA/AHA/KDIGO clinical guidelines, media transcripts) that indexed 37 documents into 620 searchable chunks.
- **RAFT Fine-Tuning:** designed and ran a Retrieval-Augmented Fine-Tuning pipeline generating training data with oracle evidence, hard distractors, and 20% abstention cases, then fine-tuned Qwen2.5-7B-Instruct via 4-bit QLoRA (Unsloth, rank 16/alpha 32) — achieving 88.5% citation precision, 92.3% citation recall, 84.6% distractor rejection, and 84.6% abstention accuracy on held-out clinical QA.
- **Local Model Serving:** quantized the fine-tuned model to GGUF and deployed it behind Ollama for zero-cost local inference, with an OpenRouter (Nemotron 3 Ultra) cloud fallback for the base agent.
- **Human-in-the-Loop Data Agent:** built a second agent for synthetic clinical data ingestion (10 patients, 479 encounters, 300 conditions, 5,368 observations, 296 medications across 18 FHIR R4 tables) with mandatory human-approval gates around autonomous data operations.
- **Full-Stack Delivery:** async FastAPI + SQLAlchemy + Alembic backend and Next.js 15 frontend, Dockerized end-to-end for local and cloud deployment.

TalentOS — Autonomous Opportunity Intelligence Platform (All Things Agentic Hackathon)

Python, Google ADK, LangGraph, Next.js, Firestore, Google Cloud Run | Solo-built, Aug 2026 — ~22K LOC across 130 files, 56 commits, 235+ test suite | Live: [all-things-agentic--allthingsage...](#) | Repo: [github.com/Naman6019/all-things-agentic](#)

- **Dual-Pipeline Agent System:** built a Careers pipeline and a parallel Studio freelance-lead pipeline (LangGraph StateGraphs orchestrating Google ADK agents on Gemini via Vertex AI) ingesting ~2,500 job postings/run across 8 ATS/aggregator sources (Greenhouse, Lever, Ashby, SmartRecruiters, Arbeitnow, Remoteive, RemoteOK, Jobicy).
- **Cost-Disciplined Filtering & Evaluation:** applies deterministic pre-filtering to cut 91% of postings before any LLM call, then runs a 3-state (MET/UNMET/NOT STATED) evaluator agent and a drafter agent to auto-generate tailored resumes, cover letters, and freelance pitches, with persistent run state in Firestore and Cloud Scheduler cron triggers.
- **Guardrails & Observability:** shipped human-in-the-loop send gates and anti-automation platform-compliance checks, with Langfuse tracing and a 235+ test hermetic pytest suite (58 test functions across 21 files, ~6s offline run) validating agent behavior without live API calls.
- **Full-Stack Dashboard:** built a Next.js 15/React 19 dashboard with separate jobs and freelance views, printable tailored resumes, and an interactive pitch editor, deployed on Cloud Run.

Question Answering on SQuAD — Model Fine-Tuning & Benchmarking

Python, PyTorch, scikit-learn, BERT, BiDAF, DistilBERT

- Compared BERT, BiDAF, and DistilBERT for extractive QA; fine-tuned each and built a standardized evaluation pipeline (accuracy, precision, recall, F1); published as a research paper.

EDUCATION & ACHIEVEMENTS

B.Tech, Computer Science Engineering (AI & ML Specialization) — VIT Amaravati

June 2021 – Sept 2025

CGPA: 8.2/10

- Published research paper: "Interactive Question Answering Model using Natural Language Processing" — International Journal of Modern Developments in Engineering and Science.
- Certifications: Azure Administrator Associate; AWS Certified Cloud Practitioner (Foundational).