

Anshika Goel

+91-9027060432 | anshika1307goyal@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#) | Ghaziabad, Uttar Pradesh, India

SUMMARY

AI/Backend engineer building GenAI systems that run in production: multi-agent orchestration, RAG pipelines, and the LLMops around them, on Python microservices throughout. Comfortable owning an agentic system end to end and leading a small team through it, from architecture and guardrails down to the evaluation harness and latency SLAs.

TECHNICAL SKILLS

Languages & Backend: Python (async, FastAPI), Node.js, SQL, REST APIs, microservices architecture, API gateway design, distributed systems

GenAI & LLMs: OpenAI, Azure OpenAI, Anthropic, Cohere; prompt engineering (CoT, ReAct, system prompts); RAG (hybrid, agentic); function calling and API orchestration; LLMops; NLP

Agents & Orchestration: LangGraph StateGraph, LangChain Agents, multi-agent system design, supervisor and planner/executor routing, agent-to-agent handoff, tool calling, contextual agent memory (session, user, long-term), workflow orchestration

Vector Search & Data: Qdrant, FAISS, Pinecone, BM25 sparse indexing, hybrid dense/sparse fusion, reranking, vector similarity search, Redis (caching, pub/sub, streams)

Cloud & DevOps: AWS (EC2, S3, RDS), Azure OpenAI Service, GCP Vertex AI, Docker, Kubernetes, GitHub Actions CI/CD, MLflow, event-driven architecture (Kafka, async patterns)

Responsible AI & Observability: guardrail and fallback design, human-in-the-loop escalation, confidence scoring, PII redaction, tenant isolation and RBAC, RAGAS evaluation, drift detection, telemetry (LangSmith tracing, Prometheus/Grafana/Sentry), KPI monitoring (success rate, latency, failure rate)

WORK EXPERIENCE

QuadB Technologies

Feb 2024 – Present

Software Development Engineer II, AI/ML Engineer

Remote, India

- Architected a **multi-agent orchestration platform** spanning 9 Python microservices, with dynamic task routing and supervisor-to-worker handoff. Integration time for a new service dropped **80%**.
- Designed the **plan-reason-act-reflect loop** and function-calling layer that lets agents reach enterprise systems and third-party APIs without hardcoded glue per integration.
- Built a **multi-provider LLM gateway** with automatic failover across OpenAI, Anthropic, and Cohere. AI features stayed **99%+** available through provider outages.
- Cut LLM API spend **~70%** and tripled response speed by resolving requests through rule engines and cached heuristics first, reserving inference for what genuinely needs it.
- Added **agent memory** as stateful context propagation scoped to request, session, and user, removing duplicate downstream calls and taking **~35%** off cross-service latency.
- Containerized the multi-service stack on Docker, shipped via GitHub Actions CI/CD, with logging and request tracing feeding dashboards for success rate, p95 latency, failure counts.
- Own the full SDLC on two-week Scrum cycles: coding, testing, debugging, design reviews, test case reviews, defect analysis across QA and UAT, plus production support readiness reviews before release.
- Mentor 3 junior engineers. Wrote the architecture standards, documentation, and reusable agent scaffolding the team starts from, recording design trade-offs and risk mitigations so the system stayed scalable and maintainable. Onboarding halved.

PROJECTS

Healthcare.AI (ClinicalRAG) [Live GitHub](#)

Personal

- Built a **clinical RAG system** over FDA drug labels and ADA/JNC guidelines: **hybrid retrieval** (BM25 plus dense embeddings, then reranking) on LangGraph, benchmarked across **9 chunking configurations** (3 strategies x 3 token sizes), every answer traced to a cited source.
- Kept low-faithfulness answers out of production with a **RAGAS evaluation gate in GitHub Actions CI**, plus confidence scoring, a refusal path when evidence is thin, and drift detection on Redis. Held **latency SLAs** of p50 <800ms, p95 <2.5s, p99 <5s.

GeniusCFO [Live](#)

Tech Lead · Ongoing

- Tech lead on a production AI accounting platform: **6 microservices** (FastAPI plus Node/Express) over PostgreSQL, Redis, and Qdrant. Posts journal entries and statutory reports across **5,000+ documents a month** and auto-matches **85%** of bank reconciliation lines.
- Built the **supervisor orchestrator** that plans a workflow across **9 specialized agents** and dispatches them over a typed **agent-to-agent protocol**, with LangGraph state machines and Redis checkpointing keeping long runs resumable.
- Owned **document ingestion** for 15+ document types: a four-engine **OCR chain** (PaddleOCR, Surya, AWS Textract fallback), open-set classification, then per-type schema extraction via forced tool use. **91.4%** field accuracy on the golden pack.
- Layered **guardrails** between model and ledger: deterministic checks fire up to 2 targeted repair rounds, then demote confidence and **escalate to human review** rather than post a wrong entry. PII is tokenized before the LLM call.
- Shipped the **conversational AI** layer on hybrid retrieval (dense plus BM25 over Qdrant), with **long-term agent memory** that distills idle sessions into typed facts and preferences and injects them into later chats.
- Instrumented it with Prometheus, Grafana, Sentry, and **LangSmith tracing**, plus a nightly loop turning production corrections into versioned prompt rules, promoted only on a non-regressing eval score.

EDUCATION

B.Tech, Computer Science and Engineering – Raj Kumar Goel Institute of Technology, Ghaziabad

2024

ACHIEVEMENTS

- 3rd Prize, ETHOnline 2025** (Best Use of ASI Alliance / Fetch.ai): built StableGuard.AI, a real-time multi-agent stablecoin monitor.
- National Finalist, Kavach 2023** (Government of India cybersecurity hackathon): led a six-member team through national elimination.
- Winner, BinaryHacks Intracollege Hackathon**: led the team and shipped a full-stack web app within the hackathon window.