

Anupam Kumar

Generative AI Engineer | LLMs · RAG · Multi-Agent Systems · LangChain · Python

Gurugram, India · +91 78763 02026 · anupamkumar.nith@gmail.com · linkedin.com/in/anupam8nith · github.com/anupam8nith · anupam.info

SUMMARY

Generative AI Engineer with 3 years of software engineering experience, building and operating production LLM systems - RAG pipelines, multi-agent orchestration, prompt engineering, and retrieval/embeddings - deployed on AWS Bedrock and Azure OpenAI. Backed by a strong backend engineering foundation (Python, FastAPI, Kubernetes, serverless) that carries GenAI systems from architecture through production and keeps them reliable once real users depend on them. Currently own the GenAI architecture for an enterprise document-automation platform and a production RAG system for contract Q&A.

CORE SKILLS

LLMs & Frameworks	LangChain, Retrieval-Augmented Generation (RAG), Multi-Agent Systems, AI Agents, Tool / Function Calling, Prompt Engineering, Structured Outputs, Model Context Protocol (MCP)
LLM Providers	Azure OpenAI (GPT-4-class), AWS Bedrock (Claude, Llama), Azure AI Foundry, Ollama (local models)
Retrieval & Embeddings	Vector Search, Semantic Search, Embeddings, Qdrant, Azure AI Search
Languages	Python, SQL, TypeScript, JavaScript, Bash, C++
Engineering Foundation	FastAPI, Docker, Kubernetes (AKS), KEDA, AWS (Lambda, API Gateway, DynamoDB, S3), Azure (Document Intelligence, Service Bus), PostgreSQL, CI/CD, System Design

PROFESSIONAL EXPERIENCE

Software Development Engineer - Q3 Technologies (Q3 Infotech Pvt Ltd) Oct 2023 – Present · Gurugram, India

- Own the architecture and backend across **two production Generative AI products** and a multi-agent client build, spanning RAG, vector search, and document AI. Chose between **AWS Bedrock** and **Azure OpenAI** based on cost, latency, and ops fit; built on LangChain, Qdrant, and Azure AI Search.
- Shipped the GenAI backbone of an enterprise bid-writing platform: **3 production microservices on Azure Kubernetes Service** with KEDA autoscaling, running a **14-stage multi-tenant document pipeline**, live in two months. Eliminated Azure OpenAI **429s under peak load** with token-aware embedding batching and Retry-After backoff.
- Took a JWT-secured **RAG pipeline** for internal contract Q&A from an empty repo to production in **five weeks** - SSE streaming, multi-page citations, intent-based query routing. Live and used daily internally.
- Designed a **4-agent system** (intent classifier, internet-fallback, document, identity) for an LMS Q&A assistant on Azure AI Foundry, keeping agents model-independent so swapping LLM versions doesn't break behavior. Delivered the first release; the client returned with a larger follow-on scope.
- Also own the broader backend/infra: serverless AWS systems serving **~1,000 daily users at 99.5% uptime**, a CI/CD pipeline auto-deploying **124+ Lambda functions**, and internal tooling (a Lambda debugger saving **~\$9K/yr**, a project scaffolder cutting new-service setup from hours to minutes). Fixed 3 critical production incidents; mentor 3 junior engineers on the GenAI and AWS stack.

Research Intern - Solid State Physics Laboratory (SSPL), DRDO Jun – Aug 2022 · New Delhi

Studied GaN HEMT performance for telecom use via Silvaco simulations, co-authoring a silicon-benchmark comparison that fed into device-design decisions.

FEATURED GENERATIVE AI PROJECTS

AI Bid-Writing Platform Python · FastAPI · Azure AKS · KEDA · Azure OpenAI · Azure AI Search · Document Intelligence · PostgreSQL 14-stage multi-tenant document indexing/extraction/summarisation pipeline across 3 AKS microservices, live in two months. Token-aware embedding batching and Retry-After backoff eliminated Azure OpenAI 429s under peak load; DOCX to PDF to Document Intelligence fallback via headless LibreOffice for files Azure couldn't read directly.	Feb 2026 – Present
Internal Contracts AI - Production RAG for Contracts Q&A FastAPI · PostgreSQL · Qdrant · LangChain · AWS Bedrock · Ollama JWT-secured RAG pipeline from empty repo to production in five weeks: SSE streaming, multi-page citations, async attachment indexing, checksum dedup. Security-hardened (patched an SSO bypass, added rate limiting and intent-based query routing). Live and used daily internally.	Jan – Feb 2026
Multi-Agent LMS Assistant Python · FastAPI · Azure AI Foundry · LangChain 4-agent system (intent classifier, internet-fallback, document, identity) for LMS Q&A. Model-independent agent design; refactored to clear every SonarQube issue across 7+ PRs. Client returned with a bigger follow-on scope after the first release.	Dec 2025 – Jan 2026

ADDITIONAL ENGINEERING EXPERIENCE

Built and operated two production serverless platforms for a UK rail operator on AWS Lambda: a real-time train-tracking system (AppSync subscriptions over DynamoDB streams; cut network overhead 25% and Lambda runtime 20%) and delay-repay claims portals now serving **~1,000 daily users at 99.5% uptime** (Loqate address/IBAN validation, 38+ Lambda functions via CI/CD, 40% faster claim processing).

EDUCATION

B.Tech, Computer Science & Engineering - National Institute of Technology, Hamirpur 2019 – 2023 · CGPA 8.41/10.0

ACHIEVEMENTS

- Competitive programming: Codeforces best rank 280, CodeChef best rank 121, 900+ DSA problems solved across LeetCode, GeeksforGeeks, HackerRank, and Codeforces.
- JEE Mains 99.08 percentile (2019).
- Published technical writing on multi-agent systems and the Model Context Protocol (MCP); write a public field-notes series on LLM reliability, RAG, and production AI engineering at anupam.info/articles.
- Building HealAll India, an open-source platform helping people find medical assistance across India - personal project, not tied to any employer.