

Anupam Kumar

Software Development Engineer | Backend & Generative AI

+91 78763 02026 · anupamkumar.nith@gmail.com · Gurugram, India · Open to Relocation

[linkedin.com/in/anupam8nith](https://www.linkedin.com/in/anupam8nith) · github.com/anupam8nith · medium.com/@anupam8nith · leetcode.com/u/anupam8nith

SUMMARY

Software Development Engineer with 2 years 7 months building production-grade serverless backends and Generative AI applications on AWS and Microsoft Azure. Shipped 4 production GenAI products including multi-agent RAG chatbots and a 14-stage multi-tenant pipeline on Azure Kubernetes Service. Strong in Python, FastAPI, AWS Lambda, DynamoDB, LangChain, AWS Bedrock, Azure OpenAI, Qdrant, and Azure AI Search. Mentored 3 junior engineers, led architecture reviews, communicated with cross-functional stakeholders, and delivered production releases under Agile and Scrum.

SKILLS

Languages: Python, SQL, JavaScript, TypeScript, C++, C, Bash, YAML

Backend & APIs: FastAPI, Django, Flask, REST APIs, WebSocket, GraphQL, Server-Sent Events, async/await, Pydantic v2, OpenAPI, Swagger

Generative AI / Machine Learning: LangChain, RAG (Retrieval-Augmented Generation), Multi-Agent Systems, AWS Bedrock, Azure OpenAI, Large Language Models (LLM), Prompt Engineering, Intent Classification, Embeddings, Vector Search, Document Chunking, Semantic Search

Amazon Web Services (AWS): AWS Lambda, API Gateway, AppSync, DynamoDB, S3, SQS, SNS, EventBridge, Cognito, CloudFormation, CloudWatch, IAM, AWS Bedrock, Aurora, RDS, Serverless Framework, Boto3

Microsoft Azure: Azure Kubernetes Service (AKS), KEDA, Azure AI Search, Azure OpenAI, Azure Document Intelligence, Azure AI Foundry, Service Bus, Event Grid, Azure Blob Storage, Managed Identity, Kustomize

Databases: PostgreSQL, DynamoDB, Qdrant (vector database), Azure AI Search (vector database), Redis

DevOps & Tools: Docker, Kubernetes, Git, GitHub, Bitbucket, CI/CD, GitHub Actions, Bitbucket Pipelines, SonarQube, pytest, mypy, flake8, Bruno

Architecture & Methodologies: Serverless, Microservices, Event-Driven, Multi-Tenant SaaS, Distributed Systems, RAG Pipelines, Multi-Agent Orchestration, System Design, Agile, Scrum, Test-Driven Development

PROFESSIONAL EXPERIENCE

Software Development Engineer, Q3 Technologies (Q3 Infotech Pvt Ltd)

October 2023 - Present

Gurugram, India · Clients: Samsung, Michelin, Panasonic, a UK rail operator (delay-repay), Mistral Data, Synthrics

- Built and deployed 3 production microservices on Azure Kubernetes Service (AKS) with KEDA autoscaling for a GenAI bid-writing platform; led delivery of 19+ tickets across a 14-stage multi-tenant pipeline with 410+ tests passing.
- Created a Docker-based AWS Lambda debugger with selective invoker, reducing cloud iteration cycles by 50% and saving Q3 approximately \$9,000 per year in AWS spend.
- Designed a CI/CD pipeline auto-deploying 124+ AWS Lambda functions via reusable CloudFormation modules and boto3 scripts, cutting manual provisioning effort by 75%.
- Led backend implementation across 4 production Generative AI products covering RAG pipelines, multi-agent systems, vector search, and document indexing on AWS Bedrock and Azure OpenAI.
- Authored setuptemplate, an internal pip-installable Python project scaffolder for FastAPI, Flask, and Django integrating mypy, pytest, and flake8, cutting project setup time from hours to under 2 minutes.
- Mentored 3 junior engineers across AWS (Lambda, API Gateway, DynamoDB, CloudFormation) and Generative AI (Qdrant, local LLMs, RAG architecture); collaborated cross-functionally with product, frontend, and operations stakeholders.

Research Intern, Solid State Physics Laboratory (SSPL), DRDO

June 2022 - August 2022

New Delhi, India

- Investigated GaN HEMT device performance for telecom applications using Silvaco simulations and TonyPlot analysis; co-authored a comparative study against silicon benchmarks that informed future device design direction at SSPL.

KEY PROJECTS

AI Bid-Writing Platform - RFP Document Indexing, Extraction & Summarisation Pipeline (Synthrics)

February 2026 - Present

- Developed and deployed a 14-stage multi-tenant document pipeline with indexer, extraction, and summarisation modules for a GenAI bid-writing platform, shipping 3 production microservices on Azure AKS with KEDA autoscaling and 410+ tests passing within 2 months.
- Implemented token-aware embedding batching (100 texts / 15k tokens on Azure OpenAI) with Retry-After backoff, eliminating 429 rate-limit errors during peak load; introduced stale-job reaper, graceful SIGTERM handling, and multi-tenant cache using Pydantic v2 and asyncpg.
- Reduced Docker image size by 23% (775 MB to 600 MB) via build-stage trimming and rolled out a DOCX-to-PDF fallback path using headless LibreOffice, expanding processable document coverage where direct Azure Document Intelligence extraction failed.

Q3 Ask Contracts AI - RAG-Based Contracts Q&A Backend

January 2026 - February 2026

- Owned end-to-end backend from initial scaffolding to UAT and production deployment in 5 weeks using FastAPI, PostgreSQL, and Qdrant; built a JWT-secured RAG pipeline with Server-Sent Events streaming, multi-page citation, async attachment indexing, and checksum-based upload deduplication.
- Shipped to production at Q3 with active usage by senior-level internal teams, receiving positive user feedback on accuracy, citation quality, and response latency; hardened security with CORS, rate limiting, security headers, and intent-based query classification.

Keypath Compass Chatbot - Multi-Agent Canvas LMS Assistant

December 2025 - January 2026

- Designed a 4-agent multi-agent system (intent classifier, internet-fallback agent, document agent, identity capability) for Canvas LMS Q&A using Python, FastAPI, Azure AI Foundry, and LangChain; refactored intentAgent reducing cognitive complexity and clearing all SonarQube issues across 7+ merged pull requests.
- Improved response caching, Canvas front-page detection, and model-independent agent design, reducing LLM coupling across model versions; initial release delivered to client and accepted, with client returning for a follow-on proposal that expanded engagement scope.

BerthMaps - Real-Time Train Berth Tracking Platform (Mistral Data)

July 2025 - November 2025

- Reduced DynamoDB network overhead by 25% and Lambda runtime by 20% by refactoring batch operations on the Python AWS Lambda backend; redesigned the SVG asset upload flow with S3 presigned URLs to eliminate the 10 MB upload cap.
- Engineered a WebSocket live-tracking layer using AppSync subscriptions over DynamoDB streams enabling real-time berth updates across regional rail networks; resolved 3 production incidents and authored knowledge-transfer documentation adopted by the team.

Claims Management Portals - UK Rail Delay-Repay (Customer + Agent Workflow)

December 2023 - mid-2025

- Built AWS Lambda-based claims processing portals (customer submission + agent workflow) for a UK rail operator's delay-repay platform, serving 1,000+ daily users with 99.5% uptime via CloudFormation and Serverless Framework.
- Integrated Loqate API for UK address and IBAN validation; deployed 38+ Lambda functions via automated CI/CD pipelines; cut claim processing time by 40% with conditional CSV/PDF report generation from Aurora PostgreSQL.

EDUCATION

Bachelor of Technology in Computer Science & Engineering, National Institute of Technology, Hamirpur 2019 - 2023

CGPA: 8.41 / 10.0 · Hamirpur, Himachal Pradesh, India

ACHIEVEMENTS & RECOGNITION

- Codeforces Global Rank 280 (Round 993, Div. 4); CodeChef Global Rank 121 (September Long Challenge 2022, Div. 4); 900+ Data Structures and Algorithms problems solved across LeetCode, GeeksForGeeks, HackerRank, and Codeforces.
- JEE Mains 99.08 percentile; National Science Olympiad Rank 3 (Std 7).
- Saved Q3 Technologies approximately \$9,000 per year in AWS cloud spend through the Docker-based Lambda debugger initiative.
- Mentored 3 junior engineers across AWS and Generative AI domains; led architecture reviews; authored knowledge-transfer materials adopted by the team.
- Building [HealAll India](#), a personal social-impact open-source platform that helps connect people seeking medical assistance; independent passion project as a way of giving back, not associated with any company or organization.
- Published a Medium article on Google ADK, Multi-Agent Systems, and Model Context Protocol (MCP), cross-posted on LinkedIn.