Neeraj Gupta

Staff Software Engineer | Full Stack Developer | Generative AI Engineer

Bengaluru, India • +91 6280904977 • **neeraj111261cse@gmail.com**

GitHub • **neerajtechtalks.com** • **YouTube**

PROFESSIONAL SUMMARY

Staff Software Engineer with 11+ years of experience building production-grade distributed systems and AI-powered products at scale. Designed and deployed 15+ AI agents across multi-agent orchestration, RAG, and voice AI — shipping 8 AI-first products and 13+ total products end-to-end in payments, fintech, edtech, and government contracting. Currently driving Generative AI adoption at Visa while independently architecting and shipping AI-first products from zero to production. AWS Certified Solutions Architect • Global Rank 6 in AWS CodeJam.

TECHNICAL SKILLS

AI / ML: LangChain, LangGraph, AWS Bedrock Agents, RAG Systems, Vector Databases (OpenSearch, FAISS), Embeddings, Agentic Workflows, Prompt Engineering, Fine-Tuning, Deepgram STT, ElevenLabs TTS, Claude API, Transformer Models

Languages: Java, Python, JavaScript, TypeScript **Frontend:** React, Next.js, React Native (Expo), Angular, Tailwind CSS, Material UI

Backend: Spring Boot, FastAPI, Node.js, Express.js, NestJS, GraphQL, REST APIs, SOAP Web Services

Cloud & DevOps: AWS (Lambda, ECS Fargate, DynamoDB, RDS/Aurora, S3, CloudFront, API Gateway, Cognito, SQS, Bedrock), Terraform, AWS CDK, Docker, GitHub Actions, Jenkins

Data & Architecture: PostgreSQL, MongoDB, Redis, OpenSearch, Apache Kafka, IBM MQ • Microservices, Event-Driven, System Design, WebSocket, Serverless, GraphQL Federation

AI PRODUCTS — DESIGNED, ARCHITECTED & SHIPPED END-TO-END

AI Proposal Automation Platform — Multi-Agent Government RFP System

FastAPI • Next.js • AWS Bedrock • LangChain • DynamoDB • Terraform • S3 • Cognito

- Designed a stateful 6-agent orchestration architecture where each agent (Ingestor, Scorer, Generator, Reviewer, Compliance, Communications) operates as an independent microservice with shared context via DynamoDB — enabling parallel execution and reducing proposal turnaround from weeks to hours
- Engineered a chunked RAG pipeline with hybrid retrieval (vector + keyword) over 10K+ documents, implementing bi-directional traceability between source requirements and generated proposal sections to ensure zero compliance gaps
- Built a probabilistic P-Win scoring engine combining 15+ weighted signals (past performance, team strength, pricing, incumbency) with automated Red/Gold/White Glove quality gates that reject sections below configurable thresholds

AI Recruitment & Video Interview Platform PRODUCTION

Next.js • FastAPI • NestJS • Aurora PostgreSQL • OpenSearch • Redis • ECS Fargate • D-ID • CloudFront

- Designed an event-driven ETL pipeline (Lambda → Step Functions → SQS) that ingests, normalizes, and deduplicates jobs from 5+ aggregators, using content hashing to prevent duplicate listings while maintaining freshness with TTL-based expiry
- Architected a real-time interview system integrating D-ID AI avatars with WebSocket-based bidirectional streaming, maintaining sub-200ms latency through connection pooling and Redis session state management
- Built a search tier on OpenSearch with custom BM25 + vector hybrid ranking across 100K+ listings, backed by Aurora PostgreSQL for ACID transactions and Redis write-through caching for concurrent session handling at scale

Real-Time AI Voice Conversation Coach — Sub-Second Latency Voice AI

React Native (Expo) • FastAPI • WebSocket • Deepgram Nova-3 STT • Claude API • ElevenLabs TTS • ECS • DynamoDB

- Designed a streaming voice pipeline (Deepgram STT → Claude LLM → ElevenLabs TTS) using chunked transfer encoding over persistent WebSocket connections, achieving <1s end-to-end latency by overlapping TTS synthesis with LLM token generation
- Built a persona engine with 6 distinct AI characters (3 interview coaches + 3 daily scenarios), each with isolated system prompts, voice profiles, and adaptive difficulty — deployed cross-platform (iOS, Android, Web) on cost-optimized ECS (~$100/month)

Trinity Transitions — Multi-Agent IEP Planning for Students with Disabilities

TypeScript • Node.js • React • Claude 3.5 Sonnet (Bedrock) • AWS CDK • Lambda • DynamoDB • S3 • WebSocket • PDFKit

- Designed an Orchestrator → Specialist agent fan-out architecture coordinating 5 domain agents (College, Employment, Training, Independent Living, Community) that run concurrently and merge results with conflict resolution and priority ranking
- Built a structured knowledge base indexing 30,000+ colleges and 1,000+ occupations with disability-specific accommodation matching and semantic deduplication (>70% overlap removal) to eliminate redundant recommendations
- Implemented HIPAA-ready architecture (AES-256 at rest, TLS in transit) with FERPA-compliant PII isolation in DynamoDB, 90-day TTL auto-purge, and real-time WebSocket streaming for progressive recommendation delivery

University VIC — Enterprise RAG Chatbot with Multi-Language Voice PRODUCTION

TypeScript • Node.js • Amazon Nova Lite • S3 Vectors • Bedrock KB • Titan Embeddings v2 • Amazon Polly • AWS CDK • GitHub Actions

- Architected a hybrid retrieval system combining S3 Vectors (semantic) + text search (keyword) with automatic citation extraction and source attribution — achieving 90% cost reduction vs OpenSearch/Pinecone while maintaining retrieval accuracy
- Designed a document ingestion pipeline that auto-processes PDF, Word, Excel, and PowerPoint with intelligent chunking (1000 tokens, 20% overlap) and Titan v2 embeddings — new documents are queryable within minutes of upload
- Built multi-language support (English, French, Arabic, Spanish) with language-aware Polly voice synthesis and full CI/CD via GitHub Actions with isolated DEV/PROD environments and API key authentication

Agentic Sales Enablement — AI Lead Scoring for Charter Bus Industry PRODUCTION

React • Python • Kafka • ECS Fargate • Lambda • DynamoDB • API Gateway • Terraform • EventBridge

- Designed a real-time CDC architecture consuming 7 Kafka topics from Confluent Cloud with idempotent consumers, syncing customer profiles and garage locations across a multi-tenant data model in DynamoDB with single-table design
- Built an AI lead scoring engine computing 0-100 scores across 5 weighted dimensions (attendance history, geo-proximity, event type, timing, verification) with configurable thresholds per organization — replacing a fully manual sales prospecting workflow

AI Transition Planning Assistant — RAG-Powered Special Education AI

FastAPI • AWS Bedrock Agent • AWS CDK • Lambda • API Gateway • DynamoDB • S3 • OpenSearch

- Designed a serverless RAG architecture (Bedrock Agent + S3 vector KB + OpenSearch) with session-aware context windowing in DynamoDB, enabling multi-turn domain-expert conversations at <$20/month fully loaded infrastructure cost

PROFESSIONAL EXPERIENCE

Staff Software Engineer

Aug 2025 – Present

Visa | Bengaluru, India

- Designing horizontally-scaled REST APIs and data pipelines powering Visa's payment intelligence platform — handling millions of daily transactions with sub-100ms P99 latency, circuit breakers, and graceful degradation across global regions
- Leading the design of multi-step Generative AI agent pipelines (LangChain + AWS Bedrock) for automated risk assessment — replacing manual review workflows with intelligent decision chains that reduce processing time by orders of magnitude
- Building production RAG systems with embedding pipelines, vector indexing, and context-window optimization — enabling semantic search and knowledge retrieval across millions of enterprise documents with sub-second query latency
- Establishing AI-first engineering standards (prompt versioning, evaluation frameworks, guardrails) adopted across multiple teams — accelerating org-wide AI adoption from experimentation to production deployment

Generative AI Developer & Senior Full Stack Developer

Jan 2020 – Aug 2025

IRIS Software | Pune, India

- Led a monolith decomposition into 12+ microservices using GraphQL federation as the API gateway layer — eliminating single points of failure, enabling independent deployments, and reducing inter-service latency by 25% for millions of concurrent users
- Designed event-driven data pipelines with Kafka (partitioned by tenant) and AWS SQS dead-letter queues for exactly-once processing — achieving 99.99% uptime SLA across EC2, Lambda, and multi-AZ RDS clusters
- Architected an AI sales platform ingesting real-time Kafka CDC streams (7 topics), computing multi-factor lead scores (0-100), and running automated event scouting — transforming a fully manual prospecting process into an intelligent pipeline
- Built a real-time investment dashboard with GraphQL subscriptions for live fund performance; deployed LangFlow-based AI workflow tools that reduced manual review cycles by 40%; mentored 8+ engineers on system design and production readiness

Full Stack Developer

Jan 2018 – Jan 2020

BLOCK8 Technologies | Pune, India

- Architected the entire cloud infrastructure for myStake (blockchain share tokenisation) from scratch — Terraform-managed GCP with auto-scaling, implemented smart contract ICO on Ethereum/Solidity, and built REST APIs with Spring Boot + MongoDB
- Reduced P95 API latency by 50% through GraphQL query optimization, Redis + CDN multi-layer caching, and N+1 query elimination; designed CI/CD pipelines that cut deployment cycles by 20%

Senior Systems Engineer

Jan 2015 – Jan 2018

Infosys Limited | Client: Morgan Stanley

- Owned the end-to-end migration of Morgan Stanley's Stock Plan Call Center Desktop to a firm-standardized platform — a mission-critical system serving corporate officers and directors globally with zero-downtime blue-green cutover across all regions
- Designed SOAP web services for the IVR Stock Trading system (phone-based trade execution); built real-time Splunk dashboards for anomaly detection; resolved production incidents within strict SLA windows across US, EU, and APAC

EDUCATION

B.Tech in Computer Science

Jaypee University of Engineering & Technology, Guna | 2015

CERTIFICATIONS & OPEN SOURCE

AWS Certified Solutions Architect

Contributor to **LangFlow** — enhanced LangChain integration, build processes, real-time orchestration

ACCOMPLISHMENTS

Global Rank 6 — AWS CodeJam • **Ranked 22nd & 34th** — CodeChef

Winner — IRIS Hackathon & CodeChef League

13+ products shipped • payments, fintech, edtech, recruitment, gov-tech

15+ AI agents designed and deployed in production

CONTENT & COMMUNITY

YouTube: **NeerajTechTalks** — AI tutorials, system design, full-stack development