# Nischay Gupta

Solution Designer · AI/Platform Engineer · Solutions Architect

hello@nischaygupta.dev
nischaygupta.dev
linkedin.com/in/itsnischayyy
github.com/itsnischayyy
Surat, Gujarat, India

## SUMMARY

3+ years shipping production systems, now owning solution design across concurrent enterprise client platforms — from problem framing and architecture through prototype, hardening, production rollout and client UAT sign-off. Previously Lead Developer for a full product team.

Builds applied AI and data platforms in Python: production hybrid RAG (vector + full-text retrieval, rank fusion, cross-encoder reranking, grounded citations), speech-to-text with speaker diarization, entity canonicalisation and self-improving taxonomies, and local GPU inference for data that cannot leave the network.

Architects backends in Java/Spring Boot, C#/.NET Core, Node.js/TypeScript and Python — layered, event-driven design with queue-based ingestion, realtime sockets and multi-provider normalisation — deployed across AWS, Azure and GCP.

Security and governance by default: Microsoft Entra SSO (OIDC/PKCE), OAuth2/JWT, RBAC with database row-level security, AES-256-GCM client-side encryption, TOTP 2FA and append-only audit trails.

## EXPERIENCE

**Solution Designer & Senior Engineer · Xtreme Thoughts SoftTech** Nov 2025 – Present · Surat, India

*Also embedded with Connexink as Solutions Architect & Senior Engineer for their outsourced client accounts.*

Architected a field feedback intelligence platform for a global consumer-health client: unstructured text and multi-speaker Hindi/English voice recordings become RBAC-scoped business intelligence for a five-tier sales hierarchy, with all inference on local GPU.

Built a hybrid RAG assistant behind a question-type router — statistical questions hit TimescaleDB rollups at zero LLM cost; qualitative ones run vector + full-text retrieval with RRF, cross-encoder rerank and grounded citations, capped at two LLM calls per answer.

Enforced RBAC in three independent layers: API scope dependencies, Postgres row-level security via per-session GUC, and scope filters injected into the vector query.

Closed a self-improving vocabulary loop (HDBSCAN drift detection, schema-constrained LLM induction, semantic guardrails) that took the review queue from 22 near-duplicate proposals to effectively zero.

Diagnosed a native-library load-order fault silently killing workers and an IPv6 resolution issue adding 2s to every model call — 149x faster vector queries.

Architected telematics ingestion for an EV fleet platform (webhooks → SQS → unified schema, DLQ auto-retry, per-provider health) — shipped with 148/148 backend and 106 client UAT cases passing.

Delivered an enterprise event & budget lifecycle portal for a pharmaceutical client on ASP.NET Core 8 + Dapper + SQL Server with Azure AD SSO.

Built a data intelligence & network visualisation platform turning unstructured social data into embeddings, semantic clusters and an interactive influence graph.

**Full Stack Engineer & Lead Developer · Shamonk Pvt. Ltd.** Feb 2025 – Oct 2025 · Surat, India

- Architected a User Management System from scratch on Spring Boot serving millions of API requests, with fine-grained role & permission authorization across hundreds of endpoints.
- Designed and shipped two production React Native apps (User and Agent) with real-time chat and voice/video over WebRTC on custom signalling, scaling to 10,000+ concurrent users.
- Ran infrastructure on AWS and GCP over PostgreSQL, MongoDB and Redis at a 99.9% uptime target; Jenkins/Docker pipelines and Prometheus monitoring cut manual deployment effort by 60%.

**Software Engineer — AI & Platform (Full Stack) · Xtreme Thoughts SoftTech**

Jan 2024 – Jan 2025 · Surat, India

- Built an AI audio analytics platform in C#/.NET analysing millions of support calls at ~2x previous throughput, with an LLM evaluation layer scoring agents at 97% agreement with human reviewers.
- Delivered a facial recognition engine at 98% accuracy, matching 5,000–10,000 images in under 10 seconds via an optimised OpenCV + CUDA embedding pipeline.
- Shipped an AI form-processing service classifying and routing 100,000+ submissions with automated email workflows (Python + NestJS).
- Automated 200-page report generation in 20 seconds with Puppeteer, cutting report storage by 80%.

## KEY PROJECTS

**Field Feedback Intelligence & RAG Analytics Platform** Architect, Lead Engineer · 2025 – present

Turns code-mixed voice notes from the field into governed, queryable business intelligence — with every model running on local GPU.

0 LLM calls for statistical questions · ≤2 LLM calls per qualitative answer · 22 → 0 near-duplicate taxonomy proposals · 149x faster vector queries after IPv6 fix · 3 independent RBAC layers

Python, FastAPI, Celery, SQLAlchemy 2.0 (async), PostgreSQL, TimescaleDB, Qdrant, Redis, MinIO, faster-whisper, pyannote, BGE-M3, Cross-encoder reranker, vLLM, Ollama (Qwen), HDBSCAN, Microsoft Entra ID, Docker

**EV Fleet Operations Platform & Telematics Ingestion** Architect, Lead Developer · 2025 – present

One unified telemetry stream from many vendors, powering a platform that runs an electric delivery fleet end to end.

148/148 backend tests passing · 106 client UAT cases passed · 0 deploys needed to add a provider

Java Spring Boot, Node.js, Express, MongoDB, AWS SQS, Redis, Socket.IO, BullMQ, Next.js, React Native, Expo, TypeScript, TypeORM, PostgreSQL, S3, Docker, ECR, GitHub Actions

**Enterprise Event & Budget Lifecycle Portal** Solution Designer, Lead Developer · 2025 – present

Medical-education events tracked from money down to evidence of delivery.

ASP.NET Core 8 MVC, C#, Dapper, SQL Server, Microsoft Entra ID, Azure Blob Storage, Quartz.NET, Serilog, jQuery, Bootstrap 5

**Data Intelligence & Network Visualisation Platform** Solution Designer, ML Engineer · 2025

Raw social data → embeddings → communities, explored as an interactive influence graph.

Python, Vector embeddings, Clustering, Graph visualisation, Multi-source data integration

**AI Audio Analytics & Facial Recognition** Software Engineer · 2024

Millions of support calls scored by LLM, and a face matcher that searches 10,000 images in under 10 seconds.

97% agreement with human reviewers · 98% face recognition accuracy · <10s to match 5,000–10,000 images · ~2x call-analysis throughput

C#, .NET Core, Python, OpenAI APIs, OpenCV, CUDA, SQL Server

**User Management System** Architect, Lead Developer · 2025

Centralised identity and access for a marketplace, serving millions of API requests.

Millions API requests served · 99.9% uptime target · -60% manual deployment effort

Java, Spring Boot, Spring Security, Spring Data JPA, JWT, PostgreSQL, Redis, AWS, GCP, Docker, Jenkins, Prometheus

**Real-Time Consultation Apps** Lead Developer · 2025

Two live Play Store apps with chat and voice/video calling over WebRTC, scaled to 10,000+ concurrent users.

10,000+ concurrent users · 2 apps live on Google Play

React Native, WebRTC, Socket.IO, Java Spring Boot, PostgreSQL, MongoDB, Redis, AWS

## SKILLS

| | |
|---|---|
| **Languages** | Python, Java, C#, TypeScript, JavaScript, SQL, C++ |
| **AI & ML** | Hybrid RAG, LLM evaluation & guardrails, LLM-as-judge, Embeddings (BGE-M3), Cross-encoder reranking, ASR + speaker diarization, Reciprocal Rank Fusion, Grounded citations, Prompt engineering, Structured / schema-constrained outputs, Function calling, Tool-calling agents, MCP servers, HDBSCAN clustering, faster-whisper, pyannote, OpenCV, Face recognition |
| **AI Serving** | vLLM, Ollama (Qwen), OpenAI APIs, Local GPU inference, Provider-abstracted model layer, GPU profile tuning, CUDA |
| **Backend** | Spring Boot, FastAPI, .NET Core / ASP.NET Core, Node.js, Microservices, Spring MVC, Spring Security, Spring Data JPA, Hibernate, Maven, Gradle, Celery, SQLAlchemy 2.0 (async), Dapper, Express, NestJS, TypeORM, REST, WebSockets, Socket.IO, WebRTC, Queue-driven ingestion |
| **Architecture** | Event-driven, Microservices, Clean / N-Tier, DDD, MVC, SOLID, API Gateway, Design patterns (DI, Repository, Factory, Strategy, Adapter) |
| **Messaging** | AWS SQS, Dead Letter Queues, Celery, BullMQ, Socket.IO, Webhooks |
| **Data** | PostgreSQL, TimescaleDB, Qdrant, Redis, Row-level security, Full-text search, SQL Server, MongoDB, MySQL, MinIO, Alembic, Schema design, Query optimisation |
| **Cloud & DevOps** | AWS, Azure, Docker, Kubernetes, EC2, S3, ECR, Secrets Manager, Entra ID, Blob Storage, GCP, GitHub Actions, Jenkins, Prometheus, Serilog |
| **Security** | Microsoft Entra SSO (OIDC/PKCE), OAuth2 / JWT, RBAC & row-level security, AES-256-GCM, Spring Security, PBKDF2, TOTP 2FA, Append-only audit logging, PII handling |
| **Frontend & Mobile** | React, Next.js, React Native, Expo, Angular, Redux, Tailwind, Razor, jQuery, deck.gl, Mapbox |
| **Practices** | Agile / Scrum, Code review, CI/CD, TDD |

## EDUCATION

**B.E. in Information Technology · Sarvajanik College of Engineering and Technology (SCET)**

2020 – 2024 · CGPA 8.2