SUJEET KUMAR

AI Engineer (LLM & GenAI) • Python • FastAPI • AWS Bedrock • Node.js

Noida, UP | **+91 7678188698** | **ksujeetkumar7678@gmail.com** | **linkedin.com/in/sujeetkumar-tech** | **github.com/Sujeet76**

SUMMARY

Backend engineer building LLM products from zero to production: ZCoded's multi-model agent pipeline (80–90% client-verified accuracy), persona chat and EUCLID at Acutus AI; sole engineer behind Empathy (20K+ downloads, later sold).

EXPERIENCE

Acutus AI — Noida, Uttar Pradesh, India
Software Engineer — Apr 2025 – Present
Decision-intelligence startup; its products include ZCoded (pre-launch behavioural simulation), QMAPI and EUCLID.

- Built ZCoded's **FastAPI** agent pipeline that simulates real users across 72 metrics with **80–90% client-verified accuracy**
- Designed **multi-LLM orchestration** across Claude, Gemini, GPT and Mistral on **AWS Bedrock** that cut inference cost
- Designed persona generation and synthesis prompts with four merge strategies: Amplify, Elect, Arbitrate and Reconstruct
- Created the **persona chat** end to end, a React UI and backend that let users question each simulated persona
- Defined a structured output schema that turns simulation results into landing-page fixes, such as copy rewrites
- Added **RabbitMQ** for long-running LLM jobs and async notifications, plus credit-based billing and an enterprise API
- Set up Bedrock invocation logging with token and cost tracking per IAM principal, and compared models on cost per run
- Built **EUCLID**, a synthetic-data SaaS whose pipeline generates privacy-preserving datasets that mirror real-world data
- Developed QMAPI's prescreening engine in Node.js and MySQL, with Redis caching, matching respondents on 10+ criteria

Studyspot India Pvt. Ltd. — Hybrid
Software Engineer — Jan 2024 – Feb 2025
*Ed-tech company whose platform, 360tutoring, connects **15,000+ students** with tutors across India.*

- Built the 360tutoring backend in **Express.js and TypeScript** on MongoDB, with MFA, KYC and role-specific onboarding
- Implemented location-based tutor search on the Google Maps API with MongoDB geospatial indexes
- Set up CI/CD with GitHub Actions and AWS ECR, cutting deployments **from hours to about 8 minutes**

Business Curative — Remote
Software Engineer Intern — Apr 2024 – May 2024

- Migrated the WordPress site to React and TypeScript on the WordPress REST API and GraphQL, with zero downtime

PROJECTS

Empathy | *Elysia (Bun), NestJS, MongoDB, Redis, AWS EC2, React Native (Expo)* — App Store | Google Play
Emotional support app connecting users with trained listeners over chat and calls; built as the only engineer on a two-person team; later sold.

- Grew to **20K+ downloads**, **2,000 MAU** and a 4.5-star rating, with **INR 1.5 lakh monthly revenue** (peak INR 10K/day)
- Built the Elysia and NestJS backend on AWS EC2 with a **self-hosted observability stack** (OpenTelemetry, Prometheus, Grafana, Loki, Tempo) and SLO alerting

OPEN SOURCE

GetStream stream-video-js | *React Native SDK, found while building Empathy*

- Fixed a callingx bug where calls with a scheme-less Android Telecom address crashed system_server and rebooted ColorOS, Realme and Oppo phones; maintainers cherry-picked my fix (#2452) into #2453 with authorship preserved
- Root-caused an Expo plugin bug (#2313) that reset ios.forceStaticLinking and broke Skia builds; filed repro and fix

TECHNICAL SKILLS

LLM & Generative AI: Bedrock, Claude, GPT, Gemini, Mistral, OpenAI, multi-LLM orchestration, agents, prompt engineering
LLMOps: Invocation logging, token and cost tracking, structured outputs, async inference queues, Claude Code, Cursor
Languages: Python, TypeScript, JavaScript | **Databases:** MySQL, PostgreSQL, MongoDB, Redis
Backend: FastAPI, Node.js (Express, Elysia, NestJS), REST, GraphQL, WebSockets, RabbitMQ, BullMQ, Jest, PyTest
Cloud & DevOps: AWS (EC2, ECR, S3, Bedrock), Docker, GitHub Actions, OpenTelemetry, Prometheus, Grafana

EDUCATION

Dyal Singh College, University of Delhi — Delhi
Bachelor of Science in Computer Science — Aug 2021 – Jun 2024