Sanjay Kumar

Head of Engineering | AI & Platform Engineering

Delhi, India | sanjox.tech@gmail.com
linkedin.com/in/sanjoxtech | github.com/sanjoxtech | sanjox.tech

Executive Profile

Engineering leader with 10+ years across marketplaces, fintech and AI-enabled products. Led 12–15 engineers at Drive Iah, owning hiring, performance management, roadmap decisions and infrastructure budget. Combines production platform and AI-workflow delivery with hands-on engineering in model serving, adaptation and agent infrastructure.

Professional Experience

Drive Iah / Drive mate — November 2020 – July 2026

Marketplace operations across Singapore and Australia.

Head of Engineering — August 2024 – July 2026

- **People leadership:** Led 12–15 engineers supporting Singapore and Australia; owned hiring, performance reviews, promotions and team management.
- **Strategy and investment:** Owned product roadmap decisions and the infrastructure budget, alongside architecture and delivery responsibilities.
- **Cost and delivery:** Reduced infrastructure costs by 35% and deployment time from approximately 60 minutes to 5 minutes.
- **Platform operations:** Led engineering for infrastructure handling 1M+ daily requests with 99.99% availability; strengthened payment processing, reconciliation and failure recovery.
- **Production AI workflows:** Led delivery of AI-assisted support, sales, identity verification and collections across both markets, with model/tool integration, conversation state and human escalation.
- **Platform transformation:** Directed the Trip Management System redesign into domain-aligned services for reservations, trips, pricing and payments. Established gRPC contracts, Kafka events and transactional outbox patterns.
- **Delivery foundations:** Built reusable Terraform/AWS infrastructure and CI/CD for ECS Fargate and Aurora PostgreSQL. Advanced TMS through development and UAT, with production migration steps documented.
- **Engineering automation:** Connected Jira, AWS Lambda, GitHub Actions and coding agents to produce reviewable pull requests; introduced validation gates and operational runbooks.

Technical Lead — 2020 – 2024

- Led backend delivery, design reviews and engineer mentoring as the marketplace expanded. Evolved APIs, integrations and cloud infrastructure through service decomposition, caching, monitoring and automated deployments.

Selected AI Systems Projects

Independent Engineering **2026 – Present**

Building three connected systems for model adaptation, inference and agent execution, with versioned artifacts, repeatable evaluation and explicit correctness gates.

- **FactoryOS – reliable agent execution.** Built a SQL-backed control plane with dependency-aware work graphs, worker leases, separate planning/execution/verification and Docker/Kubernetes sandbox backends. Implemented repeatable evaluation campaigns, PostgreSQL recovery checks and local Kubernetes execution.
- **InferStack – model serving and qualification.** Developed a Qwen-oriented serving runtime with continuous batching, prefix caching and Triton kernel experiments. Added A100 kernel-equivalence and native-liveness checks, with separate correctness and performance acceptance gates.
- **Model Foundry – adaptation workflow management.** Implemented model inspection, immutable registry/lineage, dataset and benchmark controls, and training-recipe compilation for adaptation and post-training backends. Supported by practical study of LoRA/QLoRA, SFT, preference optimization and distillation.

Earlier Experience

Josudo | Lead Developer **June 2019 – November 2020**

Led backend development for a gaming learning platform: coaching sessions, replay feedback, leagues and rewards. Built APIs and automated delivery workflows.

Attentive AI | Senior Software Engineer **May 2018 – June 2019**

Built geospatial imagery analytics backends and integrated ML models with APIs; improved data pipelines and deployment for mapping and change-detection products.

RenewBuy | Software Developer **September 2016 – April 2018**

Developed insurance aggregation and partner APIs, payment wallet integrations and sales/support tooling across motor, health and travel insurance.

Core Capabilities

Leadership: Hiring, performance management, roadmap prioritization, infrastructure budget ownership, architecture reviews and mentoring.

Platforms: AWS, Terraform, Docker, CI/CD, distributed systems, microservices, gRPC, Kafka, PostgreSQL, MySQL, Redis and ClickHouse.

AI and engineering: Agent workflows, MCP, retrieval/evaluation, model adaptation, inference experimentation, Python, TypeScript, Node.js and React.

Education

BTech in Computer Science | Amity University, Noida | 2016