

NEERUDI SAI VIKAS

AI Systems Engineer | Forward-Deployed Engineer (FDE)

Multi-Tenant AI Backends · Real-Time Telephony RAG · Financial Engines · SLM Compression · IIT Indore CSE
Hyderabad, India (Open to US Remote & Relocation) | neerudisaivikas@gmail.com | saivikas.work | github.com/kee1bo | linkedin.com/in/neerudi-sai-vikas

Summary

IIT Indore CSE alumnus building high-throughput multi-tenant AI backends, sub-second telephony RAG agents, bitemporal institutional financial reconciliation engines, reinforcement-learning decision systems, and compressed small language models from scratch. 4+ years of 0-to-1 production engineering across 15+ enterprise deployments.

Work Experience

TechRAQ

Backend & AI Systems Engineer

Apr 2026 – Present

Hyderabad, India

- **Backend and AI systems architect** for an enterprise multi-tenant EdTech SaaS platform, translating executive and client requirements into asynchronous **FastAPI** microservices, logical tenant isolation (**tenant_id** across PostgreSQL, Redis, Qdrant), and bounded LLM validation graphs.
- **Forward-Deployed AI Engineer** for US-based logistics and enterprise clients, shipping real-time cascaded telephony CX agents (Twilio/Vapi/Deepgram + Qdrant with **<350ms TTFT** and out-of-prompt Zero-Trust AuthZ), computer-vision automated billing MVPs, and document forensics pipelines.

Independent AI Systems Engineering & R&D

AI Systems Engineer & Applied ML Researcher

Jul 2022 – Apr 2026

Hyderabad / Remote

- **Delivered 15+ end-to-end production AI systems** across enterprise hybrid RAG (**Qdrant HNSW + BM25** boosting Recall@10 from 0.62 to 0.89), institutional evergreen-fund workspaces, unsupervised document layout clustering, and multi-provider LLM gateways (Claude, OpenAI, Mistral) with **99.9% uptime**.
- **Built first-principles PyTorch Transformer checkpoints** (21M BPE baseline with Shannon conditional-entropy patching & CSA2 KV-cache compression R&D) and hierarchical reinforcement learning (**HMM + PPO**) quantitative engines over TimescaleDB.

Flagship Production Systems & R&D

Multi-Tenant Autonomous Course Generation & Eval Engine | *FastAPI, PostgreSQL, Redis, Qdrant, Claude/OpenAI/go/course-en*

- Architected an async multi-stage curriculum DAG (Outline → Slides → Quizzes → TTS → SCORM) with strict **tenant_id** isolation and bounded Pydantic + Ragas LLM-as-a-Judge grader loops, slashing manual authoring time by **70%**.

Low-Latency Cascaded Telephony CX Agent | *FastAPI (SSE/WebSockets), Twilio, Vapi, Deepgram, Qdrant, Redis/go/voice-cx*

- Engineered a sub-second telephony voice agent with **<350ms LLM TTFT**, Redis sliding-window context pruning (**-40% token spend**), and out-of-prompt Qdrant payload AuthZ preventing cross-customer data leaks.

Reservoir & Altium: Institutional Evergreen-Fund Engines | *Python, TypeScript, FastAPI, PostgreSQL, Qdrant, Anthropic/go/reserv*

- Decoupled LLM source locator interpretation from a deterministic **28-digit Decimal** financial engine with bitemporal **as_known_at** cutoffs, append-only evidence gaps, 2,000-trial Monte Carlo tender simulations, and maker-checker governance.

First-Principles 21M PyTorch Transformer & SLM Compression (CSA2) | *PyTorch, BPE, RoPE, KV-Cache, FastAPI/go/slm-csa*

- Trained a 21M decoder-only Transformer from scratch ($O(1)$ tuple-hashed BPE, MHA, RoPE); evaluating Shannon conditional-entropy byte patching $H(X_t | X_{<t})$ and cross-layer KV-cache reuse (CSA2 Full/Reindex/Reuse modes).

Hierarchical RL Quantitative Engine (HMM + PPO) | *Python, PyTorch, Scikit-Learn, TimescaleDB, Docker /go/quant-rl*

- Decoupled unsupervised Gaussian HMM market regime filtration from a downstream PPO policy network and tactical execution layer backed by TimescaleDB hypertables.

Technical Skills Matrix

AI, LLMs & Post-Training: PyTorch, From-Scratch Transformers (MHA, RoPE, BPE, KV-Cache, CSA2/MLA Compression), Teacher-Student Distillation, LoRA/QLoRA, Multi-Provider LLM Gateways (Claude, OpenAI, Mistral, DeepSeek), MCP, LangGraph.

Retrieval, Voice & Multimodal: Hybrid RAG (Qdrant Dense HNSW + BM25 Sparse, RRF), ChromaDB, pgvector, Cascaded Telephony Voice Agents (Twilio, Vapi, Deepgram STT/TTS), PageIndex Trees, OpenCV, Unsupervised Layout Clustering.

Evals, RL & Reliability: Ragas, LLM-as-a-Judge, Deterministic Schema Validation (Pydantic, JSON Schema), 28-Digit Decimal Financial Contracts, RL Environments (PPO, Hidden Markov Models), Zero-Trust Retrieval AuthZ / RBAC.

Backend, Data & Infrastructure: Python (asyncio), TypeScript, FastAPI (SSE/WebSockets), Flask, Multi-Tenant Microservices, PostgreSQL, TimescaleDB, Redis (Active Context Caching & Queues), Docker, Linux, CI/CD Release Gates.

Education

Indian Institute of Technology (IIT) Indore

B.Tech in Computer Science & Engineering — Algorithms, Distributed Systems, Deep Learning, Reinforcement Learning

Graduated Jul 2022

Indore, India