Bangalore, India · +91-814-790-1810
lijo.jose@gmail.com
linkedin.com/in/lijojose
lijojose.com
github.com/lijo-jose

Lijo Jose

Principal Engineer — AI Platforms & Real-Time Systems

Principal-level engineer with **20+ years** building production systems at the intersection of **real-time media infrastructure** and **applied AI**. Architected and shipped voice AI, LLM evaluation, and RAG platforms that now analyze 100% of customer conversations across enterprise accounts (10K–50K per month); previously built broadcast-grade streaming systems (GStreamer, C++) shipped in major media OEM products. Hands-on across the stack — from C++ media pipelines to LLM orchestration — while setting technical direction for a 14-engineer organization.

EXPERIENCE

Senior Director of AI — VXI Global Solutions

Oct 2023 – Present

Set technical direction and architecture for a 14-engineer AI platform organization built from scratch; personally designed every major system below.

- Architected a **voice AI training platform** that simulates live customer calls end-to-end: a real-time speech loop (streaming STT → LLM → TTS over a WebSocket voice backend) with adaptive scenario difficulty, rubric-driven evaluation, and structured coaching workflows — converting one-on-one trainer-led role-play into on-demand simulation for every trainee.
- Built the **LLM evaluation pipeline** behind interaction analytics: rubric-based structured scoring with regression suites over versioned prompts, moving QA review from sampled spot-checks to **100% of customer conversations** across voice and chat channels (10K–50K per month).
- Designed the **retrieval architecture** for RAG-backed analytics and assistance; chose Qdrant over pgvector for payload-filtered vector search with per-tenant collection isolation, and tracked retrieval quality through the same evaluation-regression harness as the scoring pipeline.
- Built **real-time agent assistance over WebRTC**: live call audio analyzed mid-conversation through streaming transcription and LLM analysis, surfacing intent, compliance flags, and next-step guidance to operations teams while the call is still in progress.
- Designed the platform runtime on **GCP** — Cloud Run services behind an Apigee gateway, Cloud SQL, Cloud Build CI/CD — with multi-tenant isolation across client accounts.

Stack: Python · TypeScript · FastAPI · Next.js · OpenAI · Qdrant · GCP

Open Source & Recent Builds

- **Git Browse** (Open Source · MIT) — local Git repository browser: three-panel UI, command palette (⌘K), interactive rebase, branch/folder/file diff. Next.js, TypeScript, simple-git. github.com/lijo-jose/git-browse
- **Parallel Mind** — personal AI reviewer with a 5-tier privacy router keeping sensitive processing on local models (Ollama, pgvector, FastAPI); further agentic prototypes at lijojose.com/work

AI Systems Architect — Giraffe CCTV *(computer-vision security platform)*

Jan 2022 – Oct 2023

- Designed the **end-to-end computer-vision workflow** — footage collection, labeling, model training, edge deployment — running object detection and tracking on Raspberry Pi with Google Coral acceleration.
- Built a GStreamer **multi-stream RTSP receiver** with real-time alert snapshots (bounding-box overlays), Redis IPC, and a Flask control API.
- Developed a versioned **ML evaluation and regression framework** — the same evaluation-first approach later applied to LLM scoring at VXI.

AI Systems Architect — Vuelogix *(video analytics platform)*

May 2019 – Jan 2022

- Designed a **media workflow engine** on Redis-backed task queues (Python-RQ, Celery) for high-throughput parallel inference.
- Built a GStreamer network media receiver with **per-stream process isolation** for fault containment, coordinated by a FastAPI master controller.
- Architected a **high-throughput data acquisition system** in C++ (libsoup) writing to ClickHouse for batch analytics.

Lead Engineer — Ether Labs *(acquired by Atlassian)* May 2018 – May 2019

- Built a **text-extraction pipeline over HLS video segments** on AWS — frame capture, Tesseract OCR, Rekognition — making conference screen-shares searchable.
- Designed the **media streaming server** for live and post-call conference playback: multi-rate HLS, S3/CloudFront, DRM, voice-to-text.

Lead Engineer — Amagi *(broadcast SaaS, now unicorn)* Sep 2015 – Apr 2018

- Architected the broadcast **ad-localization system** in GStreamer: **Amagi-patented watermark detection (SMPTE-35)** driving frame-accurate, AV-synced ad replacement.
- Built a **digital splicing engine** (C/C++) with shared-memory buffer transfer for low-latency ad insertion on constrained devices without re-encoding.

Engineer → Senior → Lead Engineer — Ittiam Systems Sep 2004 – Sep 2015

- Architected the **multimedia framework SDK** (C/C++) that became the foundation for Ittiam's broadcast encoder, transcoder, and remux products shipped to OEM customers; built AV encoders (UDP/HLS/RTP), MPEG2-TS transcoding and remux with real-time HLS output, and high-density multithreaded audio transcoding.
- **Cut product release time by 50%** by designing and automating the end-to-end release process.

Project Engineer — Wipro Technologies Oct 2003 – Sep 2004

- Built localization support for a billing application in the logistics domain.

SKILLS

AI/ML: LLM orchestration (OpenAI GPT, Claude, Llama) · RAG · Qdrant · pgvector · evaluation frameworks / LLM-based scoring · voice AI (STT/TTS, turn-taking) · Ollama · OpenCV

Languages: Python · C++ · TypeScript

Infra/Cloud: GCP (Cloud Run, Cloud SQL, Apigee, Cloud Build) · AWS (S3, CloudFront, Rekognition) · Docker · Terraform · GitHub Actions · PostgreSQL · Redis · ClickHouse · FastAPI · Next.js

Media/Real-time: GStreamer · WebRTC · WebSocket · FFmpeg · HLS · MPEG2-TS · RTSP · edge AI (Raspberry Pi, Google Coral)

EDUCATION

B.Tech — Electronics and Communication Engineering · Model Engineering College, Cochin University · 2003