M. Sanjai

sanjaimurugan7319@gmail.com | +91-9600527319 | LinkedIn | GitHub | Portfolio | LeetCode

SUMMARY

- Full-stack engineer who builds AI products 0→1 — from whiteboard to deployed prototype — with React/Next.js front ends, FastAPI and Node.js back ends, and SQL/NoSQL databases underneath.
- Shipped a production SaaS end-to-end (Next.js, FastAPI, MongoDB, Docker on AWS/Vercel) with JWT auth, rate limiting, real-time SSE streaming, and automatic provider failover.
- Hands-on with LLMs, embeddings, vector search, and RAG — comfortable wiring models into real products. Strong fundamentals in DSA, OOP, DBMS, REST APIs, and system design.

EDUCATION

VIT Chennai Chennai, India

Bachelor of Technology in Computer Science Engineering *Aug 2024 – May 2028*

- **Relevant Coursework:** Data Structures & Algorithms, Object-Oriented Programming, Operating Systems, Computer Networks, DBMS (SQL & NoSQL), Machine Learning, Software Engineering

SKILLS

- **Languages:** Python, JavaScript, TypeScript, C, C++, SQL, Java, Go
- **Backend & APIs:** FastAPI, Node.js, REST APIs, SSE Streaming, JWT Auth, Microservices, Concurrency & Fault Tolerance
- **Frontend:** React, Next.js
- **Databases:** MongoDB, PostgreSQL, SQL (Joins, Indexing, Transactions)
- **Cloud & DevOps:** Docker, AWS (EC2, App Runner), Vercel, Git, CI/CD, Agile/Scrum
- **AI/ML:** PyTorch, Vision Transformers, CLIP, LLM APIs, RAG, Embeddings, LoRA, YOLOv8, Quantization

WORK EXPERIENCE

Founder & Lead Developer — HireFaster Jan 2026 – Present

Full-Stack SaaS Platform | *hirefaster.giize.com*

- **Architected and deployed** a production full-stack web app (Next.js, FastAPI, MongoDB Atlas, Docker) on AWS/Vercel with real-time SSE streaming, RESTful APIs, JWT auth, rate limiting, and input sanitization — handling concurrent user sessions reliably.
- **Engineered** a fault-tolerant microservice orchestration layer with 3 concurrent processing pipelines and automatic provider fallback — achieving graceful degradation and sustained request handling during individual service outages.
- **Implemented** end-to-end security hardening including input validation, authentication middleware, and abuse-prevention controls — prioritizing security from the ground up.

PROJECTS

MoE Safe-Band Quantization — Open-Source LLM Optimization Nov 2025 – Mar 2026

OLMoE, JetMoE, Phi-mini-MoE, INT8 Quantization, PyTorch, HuggingFace

- **Built** a systematic benchmarking framework profiling layer-level quantization tolerance, applying selective INT8 to ~25% of layers — ~12.5% memory reduction at near-zero accuracy loss (max -0.66pp across 6 benchmarks). Open-sourced (MIT) with reproducible pipelines.

GLI-Flow — Open-Source RTL-to-GDS ASIC Design Automation Platform Jul 2026 – Present

React, Vite, FastAPI, PostgreSQL/SQLite (Supabase), Docker, CI/CD, LLM Integration

- **Built** a full-stack platform that takes Verilog to GDSII in one command — a React/Vite dashboard (35+ views) over a FastAPI REST backend with a pluggable SQLite/PostgreSQL (Supabase) data layer, containerized with Docker and an 82-file CI test suite.
- **Engineered** an LLM-assisted investigation engine for automated root-cause analysis of failed runs; primary author (15 of 17 commits). Open-sourced under Apache-2.0.

LLM Fine-Tuning & RAG Systems May 2025 – Aug 2025

Apple MLX, HuggingFace, PyTorch, Vector Search

- **Optimized** TinyLlama 1.1B via LoRA (0.435% trainable params) and built a RAG chatbot with vector search backend — reducing hallucination through document-grounded responses.

DriveCLIP — Multi-Task Vision System for Autonomous Driving Mar 2026 – Present

PyTorch, Vision Transformers, CLIP, FPN, Cross-Attention, HuggingFace

- **Designed** a consolidated multi-task architecture extending CLIP ViT-B/32 with FPN and cross-attention, merging 5 sequential tasks into a single optimized forward pass — 2.3x throughput improvement (20.4 FPS) via architectural optimization.

KEY ACHIEVEMENTS

- Shipped a production SaaS end-to-end (architecture → deployment) with live cloud infrastructure, auth, real-time streaming, and monitoring.
- Designed fault-tolerant distributed service layer with automatic failover, sustaining request handling through individual provider outages.
- Built and open-sourced a full-stack RTL-to-GDS chip-design automation platform (React + FastAPI, 82-file CI suite) with an LLM-assisted failure-investigation engine.