

Kumar Sanu

AI Backend Engineer
Bengaluru, Karnataka

7091512373

sanuatmasai@gmail.com
LinkedIn · GitHub · Portfolio




Summary

Innovative **AI Engineer** specializing in **LLMs, RAG pipelines, multi-agent systems, and vector databases**. Skilled in **LangChain, LangGraph, fine-tuning (PEFT, QLoRA)**, with proven experience cutting inference costs and latency. Strong in backend (**Java, Python, FastAPI, Spring Boot**) and scalable cloud systems, delivering high-performance **AI-powered backend architectures and services**.

Skills

- **AI/ML & LLMs:** MCP, RAG, LangChain, LangGraph, Multi-Agent, Fine-tuning, HuggingFace, Ollama
- **Cloud/DevOps & Architecture:** Docker, CI/CD, REST APIs, Load Balancers, Microservices, Event-Driven Systems
- **Backend & Databases:** PostgreSQL, MySQL, MongoDB, Redis, Vector DBs, Java 8+, Python, Spring Boot, MVC, Security, JPA, FastAPI, RESTful APIs

Projects

- **JumpApp - AI Hinglish Learning Assistant**  
 - Built AI tutor that transforms Hinglish queries into structured lessons (scripts, audio, quizzes), achieving **85% engagement**.
 - Integrated Gemini Live Kit for natural voice-first conversations and multimodal delivery.
 - Used RAG + web search with Redis vectors to cut redundant LLM calls **40%**, reducing API costs and improving latency **25%**.
 - **Stack:** Python, FastAPI, Next.js, MongoDB, LangChain, Gemini 2.0, ElevenLabs, Sarvam, Redis
- **Automated Job Agent - Resume-to-Job Matcher** 
 - Agent on **n8n** delivering curated job alerts with tailored cover letters on **Telegram**.
 - Integrated job boards via RSS and APIs to automate role discovery and filtering, increasing relevant job matches by **85%**, cutting search time by **95%** and saving **15+ hrs/week**.
 - Reduced manual effort by **90%** with an end-to-end pipeline (fetch → parse → match → notify).
 - **Stack:** n8n, RSS/REST APIs, PostgreSQL, JavaScript, Telegram Bot API, Gemini 2.0

Experience

- **Software Engineer - Quicktouch Technologies Ltd.** Mar 2023 - Jun 2025
 - Built an **AI-powered RAG chatbot** using **LangChain, LangGraph, and LangSmith** with embeddings + vector DBs, enabling school-specific knowledge base queries in natural language.
 - Scaled multi-tenant ERP and websites for **3,500+ schools & 100K+ users** with **99.9% uptime**.
 - Optimized APIs with **Redis caching, SQL tuning, and rate limiting**, cutting latency by **60%**.
 - Designed and optimized **database schemas** for high-throughput, multi-tenant workloads.
 - Integrated 5+ payment gateways, enabling **50K+ monthly transactions**.

Education

- **MisogiAI by Masai - AI Engineering** Jun 2025 - Present
Coursework: LLMs, RAG, LangChain, LangGraph, Multi-Agent Workflows, Fine-tuning
- **Crio.Do** Apr 2023 - Apr 2024
Fellowship Program in Software Development (Backend Specialisation) Bengaluru, India

Certifications

- **Crio Certified - Distributed Cloud Components** Mar 2024