SANJAY RAM A

Dindigul, Tamil Nadu | sanjayram26102005@gmail.com | linkedin.com/in/sanjayram-ai | github.com/sanjayram-a
huggingface.co/sanjayram-a | sanjayrama.vercel.app

About

Engineering student and AWS Certified AI Practitioner specializing in on-device inference, LLM fine-tuning, and edge AI. Experienced in the full product lifecycle—from optimizing complex models for budget hardware to engineering and shipping production-ready, cross-platform AI applications to the Google Play Store. Seeking an internship to bring hands-on product development experience and scalable problem-solving to an applied AI team.

Education

SSM Institute of Engineering and Technology, Dindigul — **Aug 2023 – Expected May 2027**
B.Tech in Artificial Intelligence & Data Science — **CGPA: 7.50**
Core coursework: Deep Learning, NLP, Computer Vision, Optimization, Statistics, DSA

MSP Solai Nadar Memorial Higher Secondary School, Dindigul — **2022 – 2023**
Class XII — Tamil Nadu State Board — **Percentage: 80%**

Experience

Freelance Software Developer | *Independent Client Work* — **Oct 2025 – Present**

- **E-Commerce Backend:** Engineered a Python/FastAPI e-commerce backend with geospatial database queries to dynamically route users to the nearest physical stores.
- **Gold Valuation & Invoicing Software:** Developed a desktop application that streamlines point-of-sale operations by instantly calculating gold purity and automating physical receipt printing.

Selected Projects

AnyLLM | play.google.com/store/apps/details?id=com.dotwellabs.anyllm — *LLM, Flutter, MCP, Agentic Loops*

- Architected a unified chat app connecting GPT-4o, Claude, Gemini, Llama, and Groq within a single workspace.
- Acquired 100+ downloads and secured 1 paid customer within the first month of launch on the Google Play Store.
- Integrated MCP server support, web search, AI debates, and OpenAI-compatible BYOK (bring-your-own-key) providers.

FixGemma — On-Device Repair Assistant | github.com/sanjayram-a/fixgemma — *Gemma 4, PyTorch, Cactus, Flutter*

- Engineered a multimodal appliance-repair assistant running fine-tuned Gemma-4(E2B/E4B) fully on-device via Cactus inference.
- Achieved an inference speed of 10 tokens/sec and 800ms Time To First Token (TTFT) on sub-$200 budget smartphones.
- Eliminated cloud dependency, ensuring zero ongoing inference costs and complete user data privacy.

Humana TTS — Offline Mobile TTS | github.com/sanjayram-a/humanatts — *ONNX, React Native, Edge-AI*

- Built an offline text-to-speech mobile application utilizing on-device ONNX inference pipelines.
- Enabled low-latency voice synthesis on standard Android hardware without requiring an active network connection.

Technical Skills

AI / LLMs:	LLMs, PyTorch, TensorFlow, Scikit-learn, Transformers, Diffusion, Vision Transformers, Unsloth, OpenCV, Fine-tuning (SFT/PEFT), RAG, AI Agents, MCP
NLP & Retrieval:	LangChain, LlamaIndex, Vector Search, Embeddings, Tokenization, Prompt Engineering
Edge AI:	Edge Inference, ONNX, CoreML, MediaPipe , Quantization, On-Device Model Deployment
Languages:	Python, Java, Rust, C++, SQL, HTML, CSS, JavaScript
Backend:	Flask, FastAPI, REST APIs, Microservices
Mobile / Frontend:	Flutter, React Native, Android
Tools:	Git, Docker, MLflow, Streamlit, Gradio, Hugging Face, Linux

Certification

AWS Certified AI Practitioner | *Amazon Web Services* — **Feb 2026 – Feb 2029**
credly.com/badges/29282ecf-4871-48f6-be8e-69e94d818eb6

Interests & Languages

Technical Interests: LLM Systems, LLM Training & Fine-Tuning, Edge AI & On-Device Inference, Model Evaluation, AI Agents, Data Products, APP Development
Languages: English (Fluent), Tamil (Native), Saurashtra (Spoken)