

Sivanesan B

Coimbatore | apsiva69@gmail.com | +91-8438692752 | github.com/Sivanesanbalu |
linkedin.com/in/sivanesan-b-871ba7264 | sivanesanbalu.netlify.app

Profile Summary

AI Engineer specializing in Generative AI, Large Language Models, prompt engineering, and AI-powered applications. IEEE-published researcher with work in carbon-efficient LLM inference, achieving 58% lower CO₂ emissions and 56% faster inference. Currently working as a GenAI Executive & Prompt Engineer, developing LLM-based solutions, prompt workflows, AI agents, and RAG applications for practical business use cases. Hands-on experience with Python, PyTorch, Hugging Face, LangChain, LangGraph, FAISS, FastAPI, and Docker, with additional experience in computer vision, voice AI, and MLOps.

Experience

GenAI Executive & Prompt Engineer | *Nunes Instruments* Aug 2026 – Present

- Developing LLM-based applications and Generative AI workflows to automate business processes and work with unstructured business data.
- Designing and evaluating prompts for information extraction, classification, summarization, and structured LLM outputs, with a focus on improving response accuracy and consistency.
- Building AI automation and agentic workflows that connect LLMs with business data and APIs to support practical operational use cases.

AI Coder (Graduate Apprentice) | *Madhu Jayanti International Pvt. Ltd.* Jul 2026 – Aug 2026

- Built workflow automations using openpyxl to process large volumes of operational data and generate structured reports for business workflows.
- Automated the collection and processing of IoT machine data generated during production shifts, converting machine activity into shift-wise operational reports.
- Automated shift-end notifications by sending generated machine reports through email and WhatsApp to the assigned workers, reducing manual reporting and ensuring timely updates.

Innovation Engineer Trainee | *Forge Innovation & Ventures, Coimbatore* Jan 2025 – Jul 2025

- Built an offline voice-driven AI inventory assistant using Whisper STT and a quantized on-device LLM to help warehouse teams search and retrieve stock information through voice commands.
- Integrated FAISS-based semantic search with the local LLM pipeline to retrieve relevant inventory information without relying on an internet connection.
- Reduced manual stock-lookup time by approximately 45% across 3 warehouse pilots while maintaining zero internet dependency.

Intern – Data Networking & Infrastructure | *Bharat Sanchar Nigam Limited (BSNL)* Jun 2022 – Jul 2022

- Observed real-time signal routing and data transmission across optical fiber network nodes in a live telecommunications environment.
- Studied how network traffic moved across connected nodes, with a focus on transmission paths, latency, and network performance.
- Gained practical exposure to large-scale data transmission infrastructure and the operational characteristics of telecom network systems.

Achievements

Best Final Year Project Award – Eco-Transformers Aug 2025

Received the Best Final Year Project Award from the Department of Artificial Intelligence & Data Science, Kumaraguru College of Technology, for research on carbon-efficient LLM inference.

Publications

Eco-Transformers: Carbon-Efficient and Accelerated Inference through Self-Pruning and Activation Freezing in LLMs Aug 2025

3rd International Conference on Next Generation Computing Systems (ICNGCS-2025), PSG iTech, Coimbatore

DOI: [10.1109/ICNGCS64900.2025.11183217](https://doi.org/10.1109/ICNGCS64900.2025.11183217)

Projects

WhatsApp Product & Price Intelligence [GitHub](#)

- Built a local AI system that analyzes WhatsApp conversations and extracts product names, prices, quotations, and other business details from customer messages and shared files.
- Turned unstructured chat history into structured product-price records and automated reports showing which products were discussed, the prices mentioned, and related quotation details.
- Built a searchable SQLite-based archive for messages, attachments, and chat metadata, making historical product and pricing information easier to find and review.

Technologies Used: Python, Generative AI, LLMs, Prompt Engineering, SQLite, Flask, WhatsApp Export, Document Processing, Automation.

- Built a Flutter mobile app that recognizes book covers using a fine-tuned ConvNeXt model, reaching 92% detection accuracy across 1,000+ test images.
- Deployed the model through a FastAPI inference service designed to support 100+ concurrent users, with FAISS-based search for related book information.
- Added VLM and LLM-based processing to enrich recognized books with relevant information and provide a more useful search and discovery experience.

Technologies Used: Flutter, Firebase, FastAPI, PyTorch, ConvNeXt, FAISS, VLM, LLM, Transfer Learning.

- Built an end-to-end voice interview platform using Vapi and OpenRouter LLMs to conduct real-time candidate interviews, generate questions dynamically, and evaluate responses automatically.
- Tested the platform across 50+ mock interviews, reducing manual HR screening time by 70%; added BlazeFace-based computer vision proctoring running at 30 FPS with under 50ms latency.

Technologies Used: Next.js, JavaScript, OpenRouter API, Supabase, BlazeFace, Vapi, Clerk, Computer Vision.

Technical Skills

AI & ML: Machine Learning, Deep Learning, NLP, Computer Vision, Generative AI, LLMs, VLMs, RAG, AI Agents, Prompt Engineering, Fine-tuning, Embeddings, Voice AI
Programming: Python, SQL, TypeScript, C++, R
Frameworks: PyTorch, TensorFlow, Transformers, LangChain, LangGraph, FastAPI, Flask, OpenCV, LoRA, QLoRA
Vector DB & Search: FAISS, ChromaDB, Semantic Search
Data & AI Systems: SQLite, OpenPyXL, IoT Data Processing, Document Processing, WhatsApp Export Processing, Speech-to-Text
Cloud & DevOps: Docker, Google Cloud, Firebase, Supabase, Hugging Face Hub, Git, GitHub
AI Tools & APIs: n8n, Vapi, OpenRouter API, REST APIs, API Integration, AI Automation, Model Deployment, QA Automation
Professional Certifications: IBM AI Engineering Professional Certificate, AWS AI & ML Scholars (Udacity), Tata GenAI Job Simulation (Forage)

Education

B.Tech in Artificial Intelligence and Data Science <i>(Lateral Entry)</i>	2023 – 2026
<i>Kumaraguru College of Technology, Coimbatore</i> CGPA: 8.02/10	
Diploma in Computer Engineering	2020 – 2023
<i>Tamilnadu Government Polytechnic College, Madurai</i> CGPA: 8.0/10	