VIJAYAKUMAR S

Applied AI Engineer | LLM Systems, RAG, Agents & Real-Time Speech

+91 94889 55846 | vijay.aieng@gmail.com | linkedin.com/in/vijay-aieng | github.com/VijayAIeng | vijayai.site | Chennai, India

Objective

Applied AI Engineer with 3.5+ years of experience shipping LLM, RAG, agent and real-time speech systems into production. I design for latency, scale and cost from day one and measure every change.

Technical Skills

LLM & Agents: Llama 3, Qwen3, GPT-4o, RAG, Agentic RAG, LangChain, LangGraph, CrewAI, Function Calling, Long-Term Memory, LoRA, QLoRA, vLLM, Ollama

Retrieval: Weaviate, Qdrant, ChromaDB, FAISS, Neo4j, BGE-M3, Hybrid Search, Knowledge Graphs

Speech & Audio: Whisper, wav2vec 2.0, HuBERT, XTTS-v2, ASR, TTS, Voice Cloning, Audio Streaming

Deep Learning: PyTorch, TensorFlow, Hugging Face, ONNX Runtime, Quantization, GPU Batching

Backend & MLOps: FastAPI, WebSockets, REST APIs, Docker, Kubernetes, CI/CD, MLflow, A/B Testing

Cloud & Data: AWS (SageMaker, EC2), PostgreSQL, Redis, MongoDB, Streaming Data, Large-Scale Data Processing

Languages: Python, SQL, TypeScript, JavaScript, NumPy, Pandas, Scikit-learn, React, Next.js

Work Experience

AI / ML Engineer **Jul 2024 to Present**

Tapaya Technologies LLP, Kerala, India (Remote)

- Built a real-time voice assistant (Whisper, Llama 3, XTTS-v2) with WebSocket audio streaming on Kubernetes. Streaming each stage instead of waiting for full outputs brought response time under **500ms**.
- Cut infrastructure cost by **30%** after profiling GPU usage; added request batching and model quantization with no drop in output quality.
- Built a hybrid RAG platform (BGE-M3, Weaviate, Neo4j, Qwen3-8B on vLLM) for **1M+ enterprise documents**. Added graph context to vector search, plus access control and audit logging; retrieval precision improved **40%**.
- Developed a multilingual TTS engine for **12 languages** with first-token latency under **200ms**, and a Vision RAG index of **2.1M+ vectors**; user satisfaction rose **25%**.
- Designed multi-agent pipelines in LangGraph and CrewAI with tool calling, long-term memory and multi-step reasoning, automating manual steps and cutting workflow time by **60%**.
- Co-developed Shabdh and shipped it into the production TTS pipeline; represented Tapaya at LIMMITS'24 and LIMMITS'25 (Track 2).

Machine Learning Engineer, Intern **Oct 2023 to Jun 2024**

Tapaya Technologies LLP, Kerala, India (Remote)

- Fine-tuned ASR and TTS models on **50+ hours** of multilingual data with LoRA/PEFT, cutting Word Error Rate by **18%**; set up CI/CD and MLflow dashboards to watch models in production.

Data Scientist, Intern **Nov 2022 to Sep 2023**

Simplilearn, India (Hybrid)

- Built and evaluated **10+** supervised and deep learning models (classification, regression, NLP) in Python, TensorFlow and Scikit-learn.

Projects & Open Source

Brahm Studio: AI Visual App Builder (multi-model, agent orchestration, memory) **2026**

- Built an AI app builder with a Figma-style drag-and-drop canvas, component library, properties panel, Monaco IDE and file system, powered by FastAPI services for chat, RAG, agents, memory, database and LLM calls.
- Orchestrated multiple LLMs and specialized agents behind one chat pipeline, with long-term memory and RAG so each response uses project context and past conversations.
- Evaluated AST-based canvas-to-code sync and redesigned it as a dedicated in-canvas code editor for stability.

Shabdh: Multilingual Zero-Shot Voice Cloning (GitHub · Demo) **2025 to 2026**

- Disentangled speaker embeddings for zero-shot cloning: **97.3%** speaker verification accuracy across **12+ languages**. Submitted to LIMMITS'25; **18+ GitHub stars**.

EMOD: Zero-Shot Emotion Transfer for Expressive TTS (GitHub · Demo) **2025 to 2026**

- Disentangled emotion embeddings give controllable prosody without per-emotion fine-tuning; **18+ GitHub stars**.

Education

Sasurie Academy of Engineering, Coimbatore **2016 to 2020**

B.E. Computer Science and Engineering