Jeevarathinam V

AI/ML Engineer

Chennai, India | 8946038533 | jeevav62@gmail.com
linkedin.com/in/jeevarathinamv | github.com/Jeevav62 | huggingface.co/jeevav62 | Portfolio

Professional Summary

AI/ML Engineer specializing in production Voice AI, LLM fine-tuning, retrieval-augmented systems, and open-source AI development. Combines deep learning research with full-stack engineering to ship end-to-end AI products from prototype to deployment. Passionate about bridging cutting-edge ML research with practical engineering to drive real business impact.

Education

Anand Institute of Higher Technology **2022 – 2026**
B.Tech in Artificial Intelligence and Data Science | CGPA: 8.00 *Chennai, Tamil Nadu*

Experience

F22 Labs **Chennai, India**
AI Engineer Intern *Dec 2025 – Present*

- Authored **95+** technical POC research documents and evaluated **20+** TTS/STT/LLM/OCR models; findings contributed to internal TTS Leaderboard used by engineering teams.
- Fine-tuned 3 production TTS models (Kokoro-82M, XTTS-v2, VoxCPM) reducing WER from **60% to 22%** and improving NISQA MOS by **18%**.
- Fine-tuned LFM 2.5 1.2B instruct LLM, served via **vLLM** on a multi-GPU server; deployed full STT + LLM + TTS pipeline into LiveKit as a production voice AI agent.
- Designed production **Hybrid RAG** architecture with dense retrieval (Qdrant) + LLM reranking (Groq), achieving ˜500ms avg / ˜800ms p90 latency; implemented **GraphRAG** pipeline over Neo4j; applied **prompt engineering, prompt caching, and hallucination prevention** across production LLM pipelines.
- Benchmarked **Zvec, Qdrant, and Milvus** on RAG retrieval accuracy and latency, identified Zvec as the fastest with highest recall accuracy; published findings as a technical blog on the F22 Labs engineering blog. [Blog]
- Researched **Task Arithmetic** for TTS model merging, combined 2 fine-tuned Kokoro models in shared weight space without retraining, achieving **55% listener preference**.

Shiash Pvt Ltd **Chennai, India**
Data Science Intern *Jul 2025 – Nov 2025*

- Engineered data pipelines using Pandas/NumPy to preprocess **50+GB** datasets, improving training efficiency by **20%**.
- Performed systematic hyperparameter tuning using GridSearchCV and RandomizedSearchCV across multiple ML algorithms, boosting classification accuracy by **25%** on test data.
- Integrated trained models into Flask APIs for real-time inference, reducing latency by **30%**.

UptoSkills **Remote**
Data Analytics Intern *Jan 2025 – Apr 2025*

- Built Power BI dashboards for 500+ colleges, enabling regional insights and accreditation analysis.
- Automated data preparation with Power Query and Excel, reducing manual reporting effort by 60%.

Arul Technologies Pvt Ltd **Chennai, India**
AI/ML Intern *Nov 2024 – Dec 2024*

- Developed regression model for real estate pricing achieving R^2 of 0.85+ using NumPy, Pandas, and scikit-learn.
- Executed full ML pipeline from data ingestion through feature engineering, model tuning, and evaluation.

Projects

TTS Fine-Tuning & Task Arithmetic Research **2026**
Kokoro-82M, XTTS-v2, VoxCPM, PyTorch, Hugging Face

- **Pioneered Task Arithmetic for TTS**, combined fine-tuned female voice + Indian accent Kokoro models in shared weight space at α=0.6, β=1.0 **without any retraining**; achieved MOS 4.4 and **55% listener preference** vs 27% baseline. [Blog]
- Fine-tuned XTTS-v2 GPT component (DDP across 2 GPUs, 30 epochs, 500 synthetic clips) reduced WER by **58.4%** (18.54% → 7.71%), semantic similarity +12.1%, NISQA MOS +5.5%.
- Fine-tuned Kokoro-82M on 4,358 Indian-English audio clips (2-stage StyleTTS2) with custom Indian G2P phoneme lexicon; improved Indian proper noun pronunciation from **3.4/10 to 8.8/10**.

– Fine-tuned VoxCPM 1.5 with LoRA (r=16, α=32), NISQA MOS improved **18%** (3.19 → 3.77), WER reduced 9.1%; fixed phoneme errors for Indian names and acronyms. [Results]

Real-Time Multilingual Translation System **2026**

Node.js, WebSocket, Real-time Audio Streaming, STT/TTS APIs

– Architected browser-native live speech-to-speech translation system across **5+ Indian languages** (Hindi, Tamil, Telugu, Kannada, Malayalam, Bengali) at **˜380ms E2E latency** supporting **25+ concurrent listeners** per room.
– Reduced cross-lingual TTS latency by **83%** (650ms → 75ms) through systematic 5-provider benchmarking; engineered 3-stage pipeline: Deepgram Nova-3 STT (˜150ms) → Sarvam Translate (˜45ms) → ElevenLabs Flash v2.5 TTS (˜75ms).
– Built multi-room WebSocket architecture (host/speaker/listener roles) with API key pool rotation, real-time cost tracking, and persistent latency logging. [Demo]

LLM SEO — AI-Native Content Engine for Citation Optimization **2026**

Python, LLM APIs, Web Scraping, JWT Authentication, Cryptography, Async Processing

– Architected multi-stage AI content engine that researches, verifies, and generates citation-optimized articles directly cited by ChatGPT, Claude, Gemini, and Perplexity; engineered 5-stage pipeline: Question Discovery → Source Authority Mapping → Fact Verification → Hub & Spoke Knowledge Map → Article Generation.
– Implemented hallucination-prevention layers: quote-traceable fact extraction, cross-article contradiction detection, citation-to-content validation, and resume-from-checkpoint for cost-safety. [Demo]

Offline LLM on Android — Edge AI Inference **2026**

Python, llama.cpp, CMake, Android, LFM 2.5 1.2B Instruct

– Engineered on-device LLM inference system deploying LFM 2.5 1.2B on Android (Poco X3) via llama.cpp + CMake — **fully offline inference with zero internet dependency**, running entirely on consumer mobile hardware.
– Proved edge AI viability: quantized LLM runs on-device with no cloud backend; tested and demoed to validate LFM 2.5 support on resource-constrained edge devices. [Demo] [Blog]

AI Hoax Buster — Chrome Extension **2025**

Python, Django, Hugging Face Transformers, scikit-learn, JS/CSS, Chrome Extension APIs

– Built browser-integrated NLP Chrome extension for real-time bias and hoax detection with **sub-800ms latency** on news articles and web content.
– Engineered deterministic inference pipelines with chunked processing, label normalization, reproducible scoring, and manifest-compliant Chrome extension logic.

Technical Skills

Languages: Python, SQL
Frameworks: PyTorch, LangChain, LlamaIndex, LangGraph, CrewAI
Libraries: scikit-learn, Hugging Face Transformers, PEFT, LoRA, DDP, Pandas, NumPy, Gradio
Generative AI & LLMs: LLM Fine-tuning, Prompt Engineering, Context Engineering, Prompt Caching, Hallucination Prevention, Hybrid RAG, GraphRAG, Task Arithmetic, vLLM, llama.cpp, Hugging Face Hub
Voice AI: TTS/STT Fine-tuning, Phoneme Engineering (G2P, IPA), LiveKit
Databases & Vector Stores: MySQL, Supabase, Redis, Qdrant, Neo4j
Backend & Security: REST APIs, FastAPI, Flask, WebSocket, JWT Authentication, Cryptography
Analytics Tools: Power BI, Microsoft Excel
Cloud & DevOps: AWS, Docker, Git, RunPod, Multi-GPU Training,EasyPanel

Technical Articles & Research Blogs

I Merged Two AI Voice Models With Math And It Actually Worked | Task Arithmetic for TTS Research
Zvec vs Qdrant vs Milvus: Vector Database Comparison for RAG | F22 Labs
What Is TOON and How Does It Reduce AI Token Costs? | F22 Labs
How to Run Local LLM on an Android Phone? | F22 Labs
Reflection Prompting Explained: Why One Prompt Is Not Enough | F22 Labs

Certifications

IBM Data Science (Coursera) • Google Python Crash Course (Coursera) • AI Primer & Generative AI (Infosys) • Data Analytics Job Simulation (Deloitte) • Automation Developer (UiPath) • Introduction to Networks (Cisco)