Yuvraj Mahendra Sanghai

Pune, Maharashtra, India
+91-8378833508 | yuvrajms008@gmail.com | yuvrajms.tech | yuvraj008 | yuv008

Education

Pune Institute of Computer Technology, Pune **Nov 2022 – July 2026**
Bachelor of Engineering in Computer Engineering *Honors in Data Science*

Experience

VoiceraCX **June 2026 – Present**
AI R&D Engineer – VEE Legion *Pune*

- Defined product scope and engineered **VEE Legion**, a production platform for automated voice-agent simulation and evaluation using **Twilio/LiveKit**.
- Built **Go**-based concurrency orchestration, validating **200 concurrent calls with zero transport errors** through controlled load testing.
- Developed automated evaluation across **WER/CER, latency, tool calls, PII, scenario checks, and LLM-as-a-judge**, with safeguards against failed judge executions skewing pass rates.

VoiceraCX **June 2025 – June 2026**
Artificial Intelligence Intern *Pune*

- Developed real-time voice pipelines using **LiveKit**, implementing **LLM streaming inference**, connection pooling, and smart caching to reduce end-to-end response latency from **3s to 1.2s** (˜**60%**), redundant API calls by **60%**, and network latency by **45%**.
- Owned and maintained a production **Vector database microservice**, migrating from **Chroma to Qdrant** and implementing **hybrid dense–sparse retrieval** to reduce query latency by **90%**.

DAOStreet **Feb 2025 – June 2025**
Software Development Intern *Remote*

- Developed and maintained web application features using **Svelte**, contributing to frontend architecture, UI/UX improvements, and production development workflows.

Alesa AI Ltd, UK **Nov 2024 – Mar 2025**
AI/ML Intern *Remote*

- Fine-tuned **Llama-3.1-8B** with **PEFT** on a custom web-scraped dataset of **10,000+** dream-related terms, improving interpretation relevance.

Projects

Orpheus TTS – Fine-tuned Neural Voice Synthesis | *LoRA, Unsloth, SNAC, GGUF, FastAPI* **March 2026**

- Fine-tuned **Orpheus-TTS (Llama-3.2-3B)** to generate **SNAC audio tokens** using **LoRA** via Unsloth on **1,200 audio-text pairs**; completed in **299 steps (˜19 min)** on a single GPU.
- Debugged silent audio caused by missing **SNAC codebook offsets** across 7 token positions, identifying a codec-model boundary alignment bug through architecture-level inspection.
- Quantized to **GGUF Q4_K_M** (6.3GB → 2GB); optimized CPU inference with **llama.cpp + AVX2/FMA** and achieved **65 tok/sec** (**30× speedup**) on a T4 GPU with FastAPI.

Smart Pathshala – Offline-First AI School Ecosystem | *FastAPI, Qdrant, CrewAI* **July 2025 – June 2026**

- Built an offline-ready, multi-role school management platform for **Tier-2/3 schools** using modular **FastAPI–React**, PostgreSQL, and **Qdrant** for embedding-based search.
- Architected a **CrewAI multi-agent pipeline** for automated **Academic Calendar** generation and an **AI Exam Paper Generator** producing syllabus-aligned question sets.

Technical Skills

Languages: Python, C++, Go
Machine Learning: PyTorch, TensorFlow, scikit-learn, NumPy, Pandas, LoRA/PEFT, Unsloth
GenAI & NLP: LLMs, RAG, LangChain, CrewAI, LLM Evaluation
AI Systems & Data: FastAPI, Qdrant, PostgreSQL, MySQL, MongoDB, LiveKit, Pipecat
Evaluation & Infrastructure: WER/CER, Latency Benchmarking, Docker, Jaeger , Twilio, llama.cpp, GGUF

Achievements

Winner – Devclash 2025: 1st prize, 24-hour hackathon at DY Patil Pimpri, developing ShetNiyojan.
Winner – L&T NeuroHack: 1st prize, 24-hour hackathon at COEP Mindspark 24, developing WarCast.
Winner – Synapse 2.0: 1st prize in an AI hackathon at MKSSS CCOEW, developing LegiFy.
MHT-CET 2022 – 99.35%ile: Merit Rank **971** among 400,000+ candidates.