

Abdur Rahim

Senior AI Engineer | Voice Agents & LLM Infrastructure

erahim3@gmail.com | +880 1978 280 796 | github.com/arahim3 | linkedin.com/in/arahim3

Technical Skills

AI/ML: LLM fine-tuning (LoRA, SFT, DPO, GRPO), vLLM, RAG, HuggingFace, PyTorch, JAX, MLX
Voice AI: Pipecat, LiveKit, Whisper, Parakeet, Ultravox, Kokoro, Deepgram
Agentic: Multi-agent orchestration, LangGraph, ReAct, Function calling, Tool use, MCP
Infrastructure: Python, SQL, Docker, Linux, AWS, CUDA, Multi-GPU serving, Quantization, Vector DBs

Open Source

mlx-tune: LLM/Multimodal Fine-Tuning Framework for Apple Silicon

[GitHub – 1.3k+ Stars](#)

- Built and maintain an open-source MLX-native framework with an Unsloth-compatible API for LLM/VLM, no CUDA required
- Supports LLMs, vision-language, speech (TTS/STT), embeddings, and OCR models via LoRA/QLoRA across 7 trainer algorithms (SFT, DPO, ORPO, GRPO, KTO, SimPO, CPT)
- Authored pure-MLX CTC/RNN-T/TDT transducer losses – first streaming-ASR training path on Apple Silicon – with LoRA support for 39+ MoE architectures and 50+ end-to-end examples
- Distributed via PyPI with HuggingFace/GGUF export and 4-bit quantization via unified memory; documentation organized by user workflow

omniASR Streaming Server

[GitHub Repo](#)

- OpenAI-compatible ASR API with real-time WebSocket streaming, <50ms latency, 100+ concurrent connections
- Multi-device inference (CUDA/MPS/CPU) with VAD-based chunking and Docker/Kubernetes deployment

Professional Experience

Senior AI Engineer

ESAP, Saudi Arabia (Remote)

May 2025 – Present

- Shipped production voice agents on Pipecat with multi-agent orchestration and ReAct-style reasoning for enterprise ERP workflows
- Designed multi-model A100 inference stack (vLLM + Whisper + TTS) achieving <500ms TTFB with custom streaming services over WebRTC
- Built LLM service layer integrating Qwen/Llama with RAG retrieval and dynamic tool selection for ERP-domain agents
- Fine-tuned custom Ultravox speech-LLM (Qwen3-4B base) for bilingual Arabic/English ASR, evaluated on the SADA22 benchmark
- Delivered Playwright-based agentic web automation with function calling for enterprise operations workflows

Competitions

Kaggle Competitions Expert

[kaggle.com/rahim3](#)

Dec 2019 – Present

- Google Landmark Recognition **Silver Medal** | Google Landmark Retrieval **Silver Medal**
- SIIM-ISIC Melanoma Classification, Top **12%** | OpenVaccine mRNA Degradation, Top **12%**

Education

University of Chittagong

Chittagong, Bangladesh

Jan 2017 – May 2023

Bachelor of Science in Electrical and Electronic Engineering