Venkat Akhil Lakkapragada

akkii2006 | huggingface.co/akkiisfrommars | lvakhil06@gmail.com | LinkedIn

Summary

AI engineer specializing in training transformer-based language models from scratch on multi-billion token datasets. Deeply interested in speech and audio ML — ASR, TTS, speaker diarization, and expressive speech systems.

Education

Ongoing – 2027	B.Tech in CSE (AI & ML), Manipal University Jaipur
2023	Grade 12 (CBSE), Oakridge International School

Skills

Programming	Python, C/C++, Swift
Machine Learning	PyTorch, Transformers, torchaudio, librosa, Diffusion Models
Speech & Audio	ASR, TTS, Speaker Diarization, Emotion Recognition, pyannote, Whisper, FFmpeg, yt-dlp
LLM Development	Pretraining, Tokenization, Dataset Construction, Model Evaluation
System and Tools	Linux, Bash, Git, Hugging Face, WandB, Kaggle, Claude, Codex, Copilot
Deployment	On-device inference, model optimization

Experience

Mistyoz AI **Founder**

- Built the **CosmicFish** LLM family (19M–300M parameters) from scratch on a 60B-token multi-domain dataset, with efficient transformer architectures (RoPE, SwiGLU, RMSNorm, GQA) and full pretraining pipelines.
- Deployed models on Hugging Face and on-device via CoreML and MLX. Built audio pipelines covering ASR, speaker diarization, and emotion tagging.

Research

- **CosmicFish-HRM: Adaptive Reasoning via Hierarchical Recurrent Mechanisms in Compact Language Models** (arXiv:2605.28919 | GitHub | HuggingFace). Introduced a Hierarchical Reasoning Module (HRM) that dynamically allocates reasoning compute during inference, enabling adaptive reasoning depth in compact language models. Published on arXiv in 2026.

Projects

SpeakScan — Audio Annotation Pipeline GitHub

- Built an end-to-end audio pipeline: YouTube download, Whisper transcription, speaker diarization via pyannote, and emotion tagging via DistilRoBERTa.
- Outputs structured JSON/CSV annotations and training-ready datasets for speech model development.

CosmicFish — Custom LLM Family GitHub | HuggingFace

- Designed and trained transformer-based language models ranging from **19M–300M parameters**, quantized to int4 for efficient inference.
- Built pretraining datasets including TreeCorpus (300+ monthly downloads on HuggingFace) totalling 60B tokens across web, Wikipedia, math, and code.
- Outperforms similarly sized models by up to 15% on major benchmarks with 200+ monthly downloads sustained over a year.

Intervo — AI Voice Interview Agent GitHub | Live

- Built an AI-powered voice interview agent using Sarvam's ASR and TTS stack.
- Upload a resume and job description, conduct a full voice interview, and receive detailed performance feedback.

SpecBench — LLM Eval Suite Generator GitHub | Live

- Generates targeted benchmarks for language models from a natural language description.
- Used in production to evaluate and improve EvoMind, Mistyoz AI's moderation model.

CosmicChat — Offline AI Chat App App Store

- Built an iOS app for offline interaction with CosmicFish models, running locally via CoreML and the Apple Neural Engine.
- 500+ downloads with zero marketing or ads. Integrates EvoMind for real-time harmful prompt filtering.