

Ayush Kumar Bar

Kharagpur, West Bengal • +91-7021854808 • ayush.kumar.bar.official@gmail.com

• **LinkedIn** • **GitHub** • **ResearchGate** • **Portfolio** • **Google Scholar**

AI Engineer experienced in building and deploying scalable Speech AI, Voice AI, and LLM systems. Skilled in model fine-tuning, distributed training, and real-time inference using LoRA, LangChain, and vLLM. Proven track record in delivering state-of-the-art (SOTA) ASR models and cloud-based AI solutions on AWS and GCP.

EXPERIENCE

Shunya Labs • 04/2026 - Present

Applied AI Scientist, Full-Time

- Fine-tuned and extended large-scale Text-to-Speech (TTS) models including **SparkTTS**, **MiraTTS**, **PiperTTS**, and **QwenTTS** by adding language tokens and updating vocabularies for **22 Indic languages**.
- Optimized **SparkTTS** inference using Triton Inference Server and vLLM, achieving ~400 ms Time-to-First-Token (TTFT) and throughput of 315 tokens/sec. Evaluated encoding/decoding bottlenecks caused by the Qwen2.5-0.5B architecture generating BiCodec tokens before waveform reconstruction through the BiCodec decoder.
- Deployed **PiperTTS** for edge-device inference using ONNX Runtime, enabling real-time speech synthesis on resource-constrained Android/iOS smartphones.
- Fine-tuned ASR models for 55 Indic languages based on Qwen3-ASR (0.6B and 1.7B parameters), optimized for customer support, banking, finance, and conversational speech use cases across India.
- Served ASR models through vLLM APIs, achieving 200 concurrent requests on a single NVIDIA A100 GPU with ~450 tokens/sec throughput while maintaining a **Real-Time Factor (RTFx) of 40**.
- Built production-grade voice agents for automated appointment booking and scheduling using **Twilio**, custom ASR/TTS models, and LLMs from multiple providers, and POCs with **pipecat** and **livekit**.
- Developed Retrieval-Augmented Generation (RAG) pipelines using **Milvus** for company-specific knowledge retrieval within voice agents.
- Implemented **Silero VAD** for silence detection, turn-taking, and barge-in handling to improve conversational responsiveness and user experience.
- Trained an any-to-any Indic machine translation model on top of **NLLB-600M**, enabling multilingual translation across Indian languages and South Asian languages for clients.
- Contributed to the development and deployment of Automatic Speech Recognition (ASR), TTS, and translation systems hosted on organizational GPU infrastructure and exposed through the VAK platform (<https://www.shunyalabs.ai/vak>)
- Built speech-to-intent and text-to-intent models for edge deployments, eliminating the traditional ASR → LLM → TTS pipeline for specific task-oriented applications.
- Delivered intent-based AI solutions for automotive and telecom clients, reducing inference latency, lowering infrastructure costs, and offloading significant traffic from ASR, LLM, and TTS services at scale. Building **Text and Speech to intent models**.

United We Care • 06/2025 – 03/2026

Applied AI Scientist, Full-Time

- Fine-tuned large-scale **Whisper**, **Indic Conformer**, and **TTS** models using **LoRA (PEFT)** and **Seq2Seq Trainer** frameworks on **A100 GPUs**, optimizing training with Accelerate for distributed data parallelism (**DDP**).
- Led, developed, and deployed **real-time speech transcription** service: shunyalabs.ai/realtime on **L40 GPUs**, capable of handling **250+ concurrent requests** with low-latency inference.
- Fine-tuned Whisper ASR model [shunyalabs/pingala-v1-en-verbatim](https://shunyalabs.ai/pingala-v1-en-verbatim) achieving **2.94% WER**, surpassing industry benchmarks for English verbatim transcription.
- Contributed to **state-of-the-art Speech AI and Voice AI research**, including dataset preparation, model optimization, and GPU utilization optimization.

United We Care • 04/2024 – 06/2025

AI Engineer, Full-Time

- Built **"Stella 2.0"**, a fully agentic mental health chatbot ecosystem with 15 autonomous agents (search, clinician, recommendation, appointment-booking, onboarding, SOS, and protocol agents) orchestrated across 60+ nodes; implemented using **LangGraph**, **LangChain**, **Redis**, **Neo4j (knowledge graph)**, **Milvus (Vector Database)**, **Kafka**, and **Redis Streams** for state and event management, integrated **Serper APIs** and custom web scrapers for external search and dynamic knowledge retrieval. Find the app at [United We Care](https://UnitedWeCare.com).
- Responsible for building Clinical Copilot for health care professionals. Experienced in designing prompts (Prompt Engineering) for highly detailed and accurate responses due to sensitivity of the business. Also implemented **Chain-of-Thought (CoT)** to reduce hallucinations and also to solve a known issue of **"Needle in the Haystack"**.
- Cloud Infrastructure & Microservices: Successfully deployed and managed multiple microservices on **Google Cloud Platform (GCP)** and **Amazon Web Services (AWS)**, leveraging **NGINX** as a **reverse proxy** for load balancing and traffic management. Designed and maintained production-grade APIs using **FastAPI**, optimizing them for high scalability, low latency, and reliability in cloud environments.
- Retrieval-Augmented Generation (RAG) & Vector Search: Implemented **RAG** for vector search with **L2 (Euclidean Distance)**, **Inner Product**, and **Cosine Similarity** metrics, developing **real-time vector database APIs** that support dynamic updates under high-throughput conditions. Reduced field generation time **by 83% (30→5 mins)**. Replaced API calls with **Kafka** as a message broker, significantly reducing service load and improving system efficiency.
- Conversational AI & Memory Systems: Developed multiple conversational chatbots with persistent memory stored in **Redis**, enabling seamless continuity in dialogue without breaking conversational flow, enhancing user interaction quality.
- Implemented advanced **LangChain** and **LlamaIndex** features, including Chains, **Pydantic** parsing, Datetime and String parsing, and **Recursive Chunkers**. Optimized workflows with **dynamic prompt routing**, integrated **VectorStores** and **Redis** for persistent context, and enhanced real-time response with streaming. Refined prompt designs using Prompt Templates for scalable and effective text generation.
- Deployed and tested multiple LLMs straight from **huggingface** such as **Meta-Llama-3-8B-Instruct**, **Meta-Llama-3.1-8B-Instruct**, **Deci/DeciLM-7B**, **mistralai/Mistral-7B-v0.1**, and deploying them on **GCP, AWS** instances with **VLLM** for **OpenAI** like inference. Worked with multiple embedding models such as **all-MiniLM-L6-v2**, **paraphrase-MiniLM-L6-v2** for better vector creation and retrieval.

EDUCATION

Bachelor of Technology in Computer Science and Engineering

Techno Engineering College Banipur • Kolkata, West Bengal, India • GPA: 9.21 • 07/2019 - 07/2023

AISSCE

Kendriya Vidyalaya • Mumbai, Maharashtra, India • 74.20% • 04/2018 - 05/2019

SSC

Kendriya Vidyalaya • Mumbai, Maharashtra, India • 8.8 • 04/2016 - 06/2017

SKILLS

LLMs & NLP (LangChain, LlamaIndex, LoRA/PEFT, VLLM, Prompt Engineering, RAG), Speech & Voice AI (Whisper, TTS, ASR fine-tuning), Cloud & Deployment (AWS, GCP, Docker, FastAPI, Microservices, NGINX, Linux), Databases & Vector Stores (Redis, Neo4j, Milvus, Qdrant, FAISS), Programming (Python).

LANGUAGES

RWT: English, Hindi, Bengali