Rohith K Bobby

rohithkbobby04@gmail.com | +91 8590642088 | Mavelikara, Kerala, India
GitHub | LinkedIn

Summary

ML and MLOps engineer focused on production voice AI, backend services, GPU-backed inference, and Kubernetes deployments.

Professional Experience

ML & MLOps Engineer Dec 2025 – Present
GenerativeStudio.dev, CHVT OÜ *Remote*

- Built low-latency voice AI pipelines with LiveKit, self-hosted ASR, LLM, and TTS services for production call handling.
- Served Parakeet ASR and Kokoro TTS on bare-metal L40S infrastructure with Kubernetes, KubeRay, Ray Serve, and Docker.
- Validated inference services through concurrency sweeps, reaching 155 RPS for ASR, 56 RPS for TTS, 630x real-time audio throughput, p50 latency below 200ms, and full transcription match across 2,400 test requests.
- Containerized multi-model inference stacks, wrote KubeRay manifests, and managed registry workflows for production Kubernetes deployments.
- Reduced cold-start latency and timeout failures in async Ray Serve deployments by tuning worker concurrency, batching, and deployment settings.

Machine Learning Intern Mar 2023 – Dec 2023
IHRD / SBCID Kerala *Kerala, India*

- Built an OCR pipeline to digitize and summarize local newspaper articles, reducing manual review time for analysts.
- Created keyword extraction and article tagging systems for police case identification, improving classification accuracy by 20%.
- Developed Indic-language translation and summarization workflows used by 20+ senior officials.

Education

B.Tech in Computer Science, Cyber Security Nov 2022 – Apr 2026
College of Engineering Kallooppara *Kerala, India*

- Completed coursework in April 2026; final results awaited.
- Dissertation: Fast Fully Homomorphic Encryption via Module-LWE with parallel NTT algorithms, GPU Roofline modeling, and IND-CPA security analysis.

Projects

TinyServe
FastAPI, PyTorch, Transformers, CUDA, OpenAI-compatible API

- Building a minimal local LLM inference server for developers who need simple model serving without the overhead of larger serving stacks.
- Added OpenAI-compatible chat completions, streaming responses, request queuing, token usage logs, per-model config, and GPU memory display.
- Designed the system for one-machine deployment, readable internals, simple concurrency limits, request cancellation, and model warmup.

Offline Voice Assistant for Field Technicians
Parakeet RNNT, Kokoro TTS, ColPali, Qwen3-VL

- Led ML architecture for a fully offline edge AI voice assistant at the Armada Hackathon, built for field technicians without internet access.
- Combined local ASR, noise suppression, Vision-RAG document retrieval, vision-language reasoning, and TTS into one offline assistant.

goKyber
Go, Post-Quantum Cryptography

- Implemented the Kyber key encapsulation mechanism in Go with clean structure, practical benchmarking, and educational clarity.
- Compared lattice-based key exchange against RSA and ECC baselines to study performance tradeoffs.

Skills

Languages	Python, Go, Rust, C
Backend & APIs	FastAPI, REST APIs, OpenAI-compatible APIs, async services, service integration
ML & Inference	PyTorch, Transformers, ONNX, CUDA, Ray Serve, TensorFlow, LLM serving
Voice AI	Parakeet RNNT, Conformer-CTC, Whisper, Kokoro TTS, DeepFilterNet, LiveKit
MLOps & Infra	Docker, Kubernetes, KubeRay, Linux, CI/CD, GPU-backed deployments
Research Areas	Post-Quantum Cryptography, Kyber, Module-LWE, FHE, NTT, GPU Performance Modeling