REDDYVARI NITHISH KUMAR REDDY

Generative AI Engineer | LLM Systems, RAG & Inference Optimization

reddynithish15@gmail.com | +91 9989928002 | Hyderabad, Telangana, India
github.com/nithishkumar2022020 | linkedin.com/in/reddyvari-nithish-kumar-reddy

Profile

Generative AI engineer building RAG systems, LLM fine-tuning, and inference optimization, with a cloud-deployed RAG application and a 5.5× speech-inference speedup with ONNX Runtime. Looking for GenAI and LLM application roles.

Technical Skills

GenAI / LLM: RAG, LLM fine-tuning (LoRA/QLoRA), embeddings, prompt engineering, tool calling, MCP, LangChain, Ollama, vLLM
Retrieval & Evaluation: Milvus, Zilliz Cloud, BM25, hybrid search (RRF), semantic chunking, RAG evaluation (hit rate, MRR, NDCG)
Inference & ML: PyTorch, ONNX Runtime, quantization, ASR and WER evaluation, speaker diarization
Cloud & DevOps: AWS (EC2, Cognito, CloudWatch), Docker, Nginx, GitHub Actions, GHCR, Cloudflare R2, Azure Container Apps
Languages & Tools: Python, SQL, Java, C, Dart, Flutter, FastAPI, Git

Work Experience

Cloud Security Web LLC – Associate Software Engineer Mar 2026 – Aug 2026
Remote

- Optimized speech inference with ONNX Runtime on CPU, reducing latency from 2,409 ms to 435 ms (∼5.5×); optimized SymSpell post-processing from 2,330 ms to 282 ms.
- Exported Parakeet TDT 0.6B to ONNX and ran it with ONNX Runtime on CPU, comparing against the PyTorch baseline.
- Built a reusable evaluation framework for speech models covering WER, medical entity accuracy, quantization, latency, and regression analysis.
- Benchmarked Parakeet, MedASR, and Nemotron; Parakeet TDT v2 Q8_0 reached 14.03% medical WER and 1.59% LibriSpeech WER.
- Implemented speaker diarization post-processing and Dockerized async workers on Azure Container Apps and Service Bus.

Viswam – Machine Learning Intern Mar 2025 – Jun 2025

- Built an LLM fine-tuning workflow for Qwen 2.5 0.5B using 4-bit quantization, Unsloth, and LoRA/QLoRA; set up distributed inference with Petals across low-memory hardware.

Digi Quanta – Machine Learning Intern Mar 2025 – Jun 2025

- Designed a system prompt for a Telugu clinical-assistant LLM; researched Telugu ASR and built Streamlit and Flutter demos.

Projects

Synqra – Cloud-Deployed RAG Application
Python • FastAPI • Milvus • Zilliz • Gemini • Flutter • AWS EC2 • Docker • Nginx

- Built an end-to-end RAG application with hybrid retrieval (BM25 + dense vectors, RRF), semantic chunking, and page-level citations, with a Flutter frontend.
- Deployed on AWS EC2 with Docker Compose and Nginx (HTTPS with automated renewal), AWS Cognito JWT authorization, and GitHub Actions CI/CD publishing to GHCR.
- Added CloudWatch logging, infrastructure monitoring, and custom metrics for RAG latency, API performance, and vector database operations.
- Evaluated retrieval on 6 questions: hit rate, MRR, and NDCG of 1.0, answer F1 of 0.67; identified citation over-reporting (precision 0.38) as the main weakness, with a planned fix.

Custom MCP Tools & Reflective Coding Agent
MCP • Claude • Python

- Built MCP tools giving LLMs structured access to local files, and a coding assistant with an iterative generate-evaluate-refine loop.

Education & Certification

B.Tech, Electronics & Communication Engineering, Vignana Bharathi Institute of Technology, Hyderabad Graduated Jul 2025
Certification: Programming Fundamentals Using Python, Infosys Springboard