

HRUTHIK M

+91-7483229386 mhrithik450@gmail.com mhrithik.com

Summary

Software & AI Engineer specializing in **Large Language Models (LLMs)**, **RAG**, **Agentic AI**, and **LLM Fine-Tuning (LoRA, QLoRA, PEFT, SFT, Quantization)**. Experienced in building production-grade AI applications, data pipelines, and full-stack solutions using **React**, **Next.js**, **Node.js**, **Django**, **PostgreSQL** and **AWS/GCP**.

Skills

Programming Languages: Python, TypeScript, JavaScript, SQL, C++

AI & ML: LLMs, 1-Bit LLMs, RAG, GraphRAG, Agentic AI, NLP, Prompt Engineering, Embeddings, Vector Databases (ChromaDB, Pinecone), LangGraph, LangChain, LangSmith, OpenAI API

LLM Fine-Tuning: LoRA, QLoRA (Quantized LoRA), PEFT (Parameter-Efficient Fine-Tuning), Supervised Fine-Tuning (SFT), Quantization

Full-Stack & Cloud: React.js, Next.js, Node.js, Django, REST APIs, PostgreSQL, Docker, GitHub Actions, CI/CD, AWS, GCP

Experience

AI Engineer

June 2026 – Present

TechThink Innovations

Python, PostgreSQL, ChromaDB, LangGraph, Langchain, Streamlit

- Built an **AI-powered email intelligence platform** over **12,000+ emails** using **RAG** and **vector search**, achieving **85% retrieval accuracy** on **100 benchmark queries** validated via **human evaluation** and **LLM-as-a-Judge scoring**.
- Engineered **RAG retrieval pipelines** with a **cross-encoder re-ranker** that filtered **100+ candidate chunks down to the top 20 relevant documents**, improving **LLM response quality** and reducing hallucinations; validated via manual review of 50 representative queries.
- Developed a **conversational AI interface** using **LangGraph** agentic workflows and **LangChain**, enabling non-technical users to perform **natural language querying** with responses delivered within **3–5 seconds**.

Software Development Engineer

Apr 2025 – May 2026

Codedale Private Ltd

Next.js, PostgreSQL, Chroma DB, Langgraph, Langchain

- Developed a **speech recognition and alignment system** using **Python**, **Longest Common Subsequence (LCS)**, and **Levenshtein Distance** to solve real-time transcript accuracy issues in an AI speech-driven application, achieving sub-second latency.
- Built **voice-enabled conversational AI interfaces** integrating **OpenAI TTS** and **LLM-powered chat**, with real-time voice waveform visualization using **React Media APIs**, enabling natural, low-latency human–AI voice conversations.
- Designed **PostgreSQL schemas**, transactional workflows, and caching layers for multiple AI-powered applications, reducing average API response latency by up to **80%** across **100 representative requests**.

Projects

Neylon AI | Lead Engineer — Django, Next.js | Live

October 2025 – Present

- Built **Neylon AI**, an **AI knowledge platform** that curates domain-specific corpora from websites and documents; fine-tuned **Phi-3 Mini (3.8B)** using **QLoRA**, **PEFT**, and **Supervised Fine-Tuning (SFT)** with **4-bit Quantization** on the extracted data for accurate domain-specific Q&A.
- Augmented the fine-tuned model with a **hybrid RAG pipeline** (**BM25 + ChromaDB** vector search) achieving **95% retrieval accuracy**; deployed end-to-end on **GCP Cloud Run** via **Docker** and **GitHub Actions CI/CD**.

Education

B.Tech. Computer Science and Engineering

CGPA: 7.61 / 10

University of visveswaraya college of engineering, K.R Circle, Bangalore