HAREESH LOTTI

Eluru District Nuzvid, AP ☎ 6374248411 lottiharish@gmail.com GitHub LinkedIn

Profile Summary

Emerging **AI/ML Engineer** with hands-on experience building **RAG systems, vector search pipelines**, and **backend APIs** using FastAPI, LangChain, and Qdrant. Skilled in developing scalable document ingestion workflows, semantic search, and intelligent retrieval for real-world AI applications. Strong problem-solver with a focus on clean engineering, automation, and practical AI deployment.

Education

Bachelor of Technology in Computer Science and Engineering **2022–2026**
Rajiv Gandhi University of Knowledge Technologies (IIIT Nuzvid), Andhra Pradesh *CGPA: 8.85*

Pre-University Course – PUC **2020–2022**
Rajiv Gandhi University of Knowledge Technologies (IIIT Nuzvid), Andhra Pradesh *CGPA: 9.64*

Board of Secondary Education, Andhra Pradesh **2019–2020**
Vignan Vidhya Nikethan *CGPA: 10*

Technical Skills

Programming: Python (advanced), Java, C, SQL

AI/ML Frameworks: PyTorch, TensorFlow, Scikit-learn, Keras, Transformers

Generative AI: LangChain ,LangGraph, LangSmith, RAG Agents, Fine-tuning (PEFT, LoRA/QLoRA), OpenAI/Gemini/Groq APIs

NLP: Hugging Face, NLTK, spaCy, fuzzywuzzy, BeautifulSoup, SentenceTransformers

Data Analysis & Visualization: Pandas, NumPy, SciPy, Matplotlib, Seaborn, Plotly, Geopandas

Web Development & APIs: Streamlit, Flask, FastAPI, Pydantic, REST APIs, Spring Boot, HTML/CSS/JavaScript, **Map Servers & Clients (GIS)**

Audio Processing & Real-time Systems: Livekit, real-time audio mixing, resampy, scipy, soundfile, TTS pipelines

MLOps & Deployment: Docker, Git, CI/CD basics, MLflow, Hugging Face Hub

Databases: MySQL, Vector Databases (Qdrant, FAISS, Chroma, Pinecone)

Projects

Pneumonia Detection from Chest X-Ray Images using Custom CNN Architecture *{Deep Learning, Medical Imaging}*
GitHub | Project Paper

- Developed a 5-layer **Custom CNN architecture** in TensorFlow/Keras for automated pneumonia detection from pediatric chest X-rays.
- Achieved **90.4% test accuracy**, with **93% precision for Pneumonia** and **87% for Normal** cases on the test dataset of 5,863 images.
- Implemented **comprehensive data augmentation** and **Batch Normalization/Dropout** techniques to counter class imbalance and overfitting.
- Utilized **AI-assisted development tools (Cursor IDE)** for intelligent code generation, debugging, and documentation.
- Proposed future enhancements including **ensemble methods, interpretability with Grad-CAM, and clinical validation**.

Ask Chatbot! (Integrated RAG & Tool-Using AI Agent) *{Generative AI, LangGraph, RAG, Agents, LangSmith}*
GitHub

Knowledge Assistant – Document-Based RAG System

- Developed a **Retrieval-Augmented Generation (RAG)** chatbot using **LangChain**, **FAISS**, and **Groq API** for accurate document-based question answering.
- Enabled multi-document ingestion, semantic search, and adaptive context chunking for improved query relevance.
- Optimized retrieval latency and precision through custom embedding management and query ranking logic.

Action Assistant – LangGraph & Gemini Agent

- Built a **LangGraph**-powered intelligent agent integrated with **Google Gemini** for tool-based actions including **DuckDuckGo Search** and **WeatherStack API**.
- Designed a refined **Streamlit UI** with tool execution progress indicators ("Using tool..." → "Tool finished") and clean final-response-only message rendering.
- Integrated **LangSmith** for workflow debugging, LLM trace inspection, and performance analytics of multi-tool execution pipelines.

Kaggle: NYC Taxi Fare Prediction *{Machine Learning, Regression}* **GitHub**

- Developed a regression model to predict NYC taxi fares, securing a **top 30% Kaggle rank** with an RMSE of **$3.0**.
- Cleaned and preprocessed large-scale trip data, removing outliers and computing features like **Haversine distance** and temporal patterns.
- Compared multiple algorithms and selected **XGBoost** as the best performer based on cross-validation results.

Insurance Premium Prediction API *{ML API, FastAPI, Docker}* **GitHub** | **Live API Endpoint**

- Engineered features (BMI, age group, lifestyle risk) and trained a **Random Forest** model to predict insurance categories, achieving **90% accuracy**.
- Built a production-ready REST API using **FastAPI** to serve the trained model and provide real-time inference.
- Implemented server-side data validation and automatic feature computation from raw inputs (e.g., age, height) using **Pydantic** models and `@computed_field` decorators.
- Designed a clear API response schema to return the predicted category, confidence score, and class probabilities.
- Containerized the complete application, model, and dependencies using **Docker** for scalable and isolated deployment.

YouTube Video Analyzer *{RAG}* GitHub

- Built a RAG-based application to "chat" with YouTube videos by asking questions about their transcribed content.
- Engineered a pipeline to download video transcripts, create a searchable **FAISS** vector index, and generate answers using LangChain and Gemini.

Experience

AI Backend & Applied Research Experience

AI Backend Engineer (Contract Project) – Zudu.ai **Nov 2025 – Dec 2025**

Product & SaaS-Based Conversational AI Company

- Designed and deployed an enterprise-grade **RAG (Retrieval-Augmented Generation) backend microservice** using FastAPI, acting as the core knowledge engine for a production voice assistant.
- Implemented a **Hybrid Search pipeline** in Qdrant combining dense embeddings with sparse keyword scoring, significantly improving retrieval precision for technical documents.
- Enhanced query processing with **query rewriting capabilities** and fixed search threshold configuration bugs to improve relevance scoring.
- Engineered a multi-format **document ingestion pipeline** with Adaptive Text Splitting (Recursive, Markdown, Semantic) and **Parent-Child Indexing**, ensuring full-context retrieval for long documents.

- Integrated **semantic chunking** using LangChain Experimental and implemented **LLM-based metadata extraction** (Title, Author, Summary) to enhance document understanding and retrieval quality.
- Built **high-fidelity PDF parsers** using PyMuPDF to handle complex layouts, tables, headings, and implemented **OCR fallback with Tesseract** for image-dominant documents.
- Optimized end-to-end system latency using asynchronous I/O, vector batching, efficient Qdrant query parameters, and connection pooling—supporting real-time voice queries.
- Developed a scalable and production-ready API infrastructure with structured logs, exception handling, health checks, and **Docker-based deployment** for Azure cloud environments.
- Built a **real-time background sound mixing feature** for Livekit voice agents, implementing a **BackgroundSoundMixer service** in Python that integrates with the TTS pipeline to overlay ambient audio (9 sound types: coffee_shop, call_center, office, restaurant, outdoor, etc.) with configurable volume control (-30 to -10 dB).
- Developed the backend API layer in Java (Spring Boot) for background sound configuration, including database migrations, entity lifecycle hooks, validation logic, and CRUD operations supporting agent-level audio customization.
- Implemented comprehensive debug logging and error handling for audio processing workflows, ensuring robust troubleshooting capabilities for production voice assistant deployments.

Summer Internship - Black Bucks – AI & ML Intern **2025**

- Developed and optimized machine learning models for real-world business use cases, contributing to core product features.
- Engineered data preprocessing and feature engineering pipelines to enhance model accuracy and robustness.

Web Team Member – eCrush, IIIT Nuzvid **2022 – Present**

- Developed web automation tools, including a results scraper, to reduce manual work and support student engagement.

Project Contributor – IIIT National Level Fest **2021 – Present**

- Designed a Lottery Ticket & Email Automation System, generating over **₹60,000 revenue** and adopted annually by juniors.

Certifications & Achievements

Certifications

- NPTEL – Deep Learning
- NPTEL – Large Language Models (LLMs)
- View All Certificates

Achievements

- 1st Place – Madcode Coding Competition
- 1st Place – Tech Hunt Competition
- Organized and mentored AI/ML workshops for juniors
- View Achievement Proofs