

MOHAMMED NASEEF AK

AI/ML Developer | Generative AI & RAG Engineer

✉ mohammednaseef301@gmail.com 📞 +91 7907451701 [in LinkedIn](#) [GitHub](#) Malappuram, Kerala, India

SUMMARY

MCA graduate (2026) specializing in Machine Learning, Deep Learning, and Generative AI, with hands-on experience building Retrieval-Augmented Generation (RAG) applications using LangChain, FastAPI, and locally-served LLMs. Engineered PdfBot, a hybrid BM25/dense-retrieval RAG pipeline with cross-encoder reranking and multi-turn conversational memory, and a Hybrid CNN-LSTM deep learning model achieving 90% accuracy on speech-based distress detection. Strong Python foundation across the ML lifecycle, from data processing and model training to evaluation and REST API integration, with a builder's focus on shipping practical AI solutions.

SKILLS

Languages: Python, SQL

ML/DL Frameworks: TensorFlow, Keras, PyTorch, Scikit-learn, NumPy, Pandas

Generative AI / LLM / RAG: Large Language Models (LLMs), Generative AI, Retrieval-Augmented Generation (RAG), LangChain, Hugging Face Transformers, Vector Databases (ChromaDB, FAISS), Prompt Engineering, Hybrid Retrieval & Reranking, Semantic Search, Natural Language Processing (NLP)

Tools/Platforms: FastAPI, REST APIs, React, Git, Librosa

Concepts: CNN-LSTM, Multi-Head Attention, Model Optimization, Data Augmentation

PROJECTS

PDFBOT — LOCAL RAG-BASED DOCUMENT Q&A SYSTEM, (PYTHON, LANGCHAIN, FASTAPI, REACT, CHROMADB, LLAMA 3.2, OLLAMA)

- Architected a fully local Generative AI application integrating a Large Language Model (Llama 3.2) with a Retrieval-Augmented Generation (RAG) pipeline, eliminating external API dependency for complete data privacy.
- Built a hybrid BM25 + dense-vector (ChromaDB) retrieval system fused through Reciprocal Rank Fusion, resolving the scale mismatch between lexical and semantic similarity scores.
- Developed and integrated REST APIs using FastAPI to connect a React/Vite front end with the retrieval and generation backend, supporting a 5-stage pipeline including multi-query expansion and parent-child chunking (400/1500-token).
- Optimized retrieval precision with a cross-encoder reranking stage (ms-marco-MiniLM-L-6-v2) and hard-floor logit filtering to surface the most relevant context before generation.
- Engineered stateful, multi-turn conversational memory through a sliding-window context manager, using all-mpnet-base-v2 embeddings.

VOICE DISTRESS DETECTION FROM SPEECH, (PYTHON, TENSORFLOW, KERAS, LIBROSA, FLASK)

- Designed and trained a Hybrid CNN-LSTM model with Multi-Head Attention for binary distress classification and 5-class emotion recognition (neutral, happy, sad, angry, fearful), achieving 90% accuracy on binary distress detection.
- Extracted and engineered audio features (MFCC, Spectral Contrast, Tonnetz, Chroma, Zero-Crossing Rate) using Librosa across three benchmark datasets (CREMA-D, RAVDESS, TESS).
- Improved model robustness through data augmentation (time-stretching, pitch-shifting, Gaussian noise injection) and addressed class imbalance using weighted loss functions and oversampling.
- Deployed a Flask-based web interface for real-time audio upload and emotion/distress inference.

EDUCATION

MASTER OF COMPUTER APPLICATIONS, Indira Gandhi National Tribal University, Amarkantak
2024 – 2026

BSC COMPUTER SCIENCE, EMEA College of Arts and Science, Calicut University 2021 – 2024