Kalaiyarasi G

Chennai • kalaiyarasi04ai@gmail.com • +919751718523 • in/kalaiyarasigopal

SUMMARY

AI/ML Developer specializing in agentic AI, RAG architectures, and LLM applications using LangGraph, Google Gemini APIs, FAISS, Sentence Transformers, PEFT, PostgreSQL, React, and Python. Delivered 90% memory optimization, sub-second response times, and production-ready multimodal pipelines across automated workflows.

EXPERIENCE

AI/ML Developer

Least Action Company **September 2025 - Present,**
Vellore

- Developed multiple production-grade AI agents (personal growth, sales automation, social engagement, productivity) solving the problem of unstructured workflows and inconsistent user tracking.
- Built intelligent pipelines using LangGraph, StateGraph, Google Gemini Pro, enabling advanced intent classification, multi-step reasoning, and agent memory handling.
- Designed a scalable PostgreSQL system (users, goals, tasks, interactions) with Psycopg2, ensuring efficient storage, querying, and workflow execution.
- Implemented conditional routing and multi-agent orchestration to automate goal tracking, sales follow-ups, and content workflows, improving task automation accuracy and reducing manual effort.

Machine Learning Intern

Graditwin **February 2025 - May 2025,**
chennai

- Built an AI document assistant solving the problem of slow, manual document review across multiple file formats.
- Created a RAG pipeline using LangChain, Sentence Transformers, FAISS, delivering fast semantic search and accurate document-grounded responses.
- Integrated multi-format text extraction using PyPDF2, python-docx, pandas, openpyxl, python-pptx supporting PDF, DOCX, TXT, CSV, Excel, and PPTX.
- Leveraged the Groq API (Llama3-8B-8192) for high-speed inference, reducing query response time from minutes to seconds.

Al Developer Intern

Strydo Technologies **June 2023 - July 2023,**
Vellore

- Addressed the challenge of domain-specific LLM adaptation under limited GPU resources.
- Fine-tuned models using Unsloth, PEFT, LoRA, and 4-bit quantization, enabling training on restricted hardware like Colab T4.
- Optimized attention layers (Q, K, V) to achieve 30x faster training and 90% memory reduction.
- Delivered multiple domain-tuned LLMs with high performance on targeted use cases.

PROJECT

Corrective-RAG

- Identified critical limitation where basic RAG systems generated hallucinated responses due to poor initial document retrieval without quality detection mechanisms.
- Engineered CRAG pipeline with dual-stage retrieval featuring automated relevance evaluation (0-1 scoring) that triggers hybrid search combining FAISS vector similarity and BM25 keyword matching.
- Developed using Python, LangChain, GROQ LLM, FAISS, BM25, Sentence-Transformers, and PyPDF.
- Achieved 95% relevance accuracy, improved answer quality by 40%, and implemented deduplication logic reducing redundant context by 35%.

Multimodal RAG System

- Developed a full-stack, multi-modal AI chat application that allows users to upload various file types and ask questions to receive clear answers.
- Text files (PDFs, Word, TXT) extracted using standard libraries like PyPDF2 and python-docx, images analyzed with Google Gemini Vision to extract detailed descriptions, audio files are transcribed using the Deepgram API along with PyDub and soundfile libraries.
- Video files are processed by extracting audio (via moviepy) and key frames (via OpenCV), the audio was transcribed with Deepgram, and frames are analyzed using Google Gemini Vision to generate rich multimodal content.
- The system uses FAISS as a vector database for semantic search, Sentence Transformer's "intfloat/e5-large-v2" model for embedding generation and chunking to enable efficient retrieval-augmented generation (RAG).
- The frontend built with React and TypeScript provides an interactive chat interface supporting file uploads, creating an easy-to-use app for querying diverse file content.

Real-Time Voice Assistant (Jarvis)

- Developed a real-time AI assistant named Jarvis to provide natural, clear, and friendly conversational support tailored specifically for personalized user assistance.
- Used LiveKit SDK for real-time session management (AgentSession, Agent, RoomInputOptions), Google Gemini 2.0 Realtime Model for streaming andintegrated cloud-based noise cancellation (BVC) to enhance audio quality.
- Delivered a responsive, human-like AI assistant capable of smooth real-time voice interactions with improved audio clarity and customizable, supportive communication style that enhanced user engagement.

Al-powered website summarizer

- The project aimed to develop an AI-powered website summarizer that automates extracting and summarizing key content from websites.
- Built the system integrating the LLM powered by llama-3.3-70b-versatile for advanced natural language understanding.
- It uses the Pydantic AI Agent framework with MCP server (mcp fetch server and firecrawl) integration to manage AI communication asynchronously, ensuring efficient and scalable AI processing.
- Utilized AsyncIO for concurrent website fetching to improve processing speed, Pandas for data manipulation and analysis, and Plotly for creating interactive,insightful data visualizations.
- This addressed the challenge of manual, time-consuming web research by providing users with concise, actionable insights efficiently.

Customer Churn Prediction

- Built a Customer Churn Prediction System for telecom data to identify at-risk customers, addressing revenue loss from customer churn by predicting likelihood ofchurn based on demographic, billing, and service usage data.
- Used Pandas and NumPy for preprocessing, XGBoost Classifier (churn_model.pkl) for high-performance predictions on tabular data, and SMOTE-ENN to effectively balance churn vs non-churn classes; applied robust label encoding with an "Unknown" fallback for unseen categorical values and used Joblib to serialize/load the trained model and encoder (label_encoders.pkl).
- Integrated with a CSV dataset (customer_churn.csv) to auto-fetch existing customer details, enabling real-time churn prediction for business teams.
- The project has potential to reduce churn rates by 10-15% through early customer engagement strategies.

SKILLS

AI & LLMs: Llama-3.x, Gemma, Qwen, Gemini APIs, Claude, PEFT, LORA,

QLORA ML Tools: Scikit-learn, XGBoost, TensorFlow, PyTorch

Multimodal Vision: Gemini Vision, Deepgram, OpenCV,

LiveKit Data & Visualization: Pandas, NumPy, Plotly, power BI

Development & Databases: Python, SQL, React, TypeScript, Streamlit, FastAPI, PostgreSQL, MySQL, Psycopg2

EDUCATION

B.Tech in Artificial Intelligence & Data Science

Annai Mira College of Engineering and Technology • Ranipet • 2025 • 8.51