Kalaiyarasi G

AI/ML Developer

+919751718523 - kalaiyarasi04ai@gmail.com - Chennai - LinkedIn

SUMMARY

AI/ML Developer specializing in agentic Al systems, LLM integrations, RAG architectures, multimodal pipelines, and PostgreSQL-backed automation workflows. Experienced with building full-stack intelligent applications using LangGraph, Google Gemini APIs, FAISS, Sentence Transformers, and PEFT fine-tuning techniques. Delivered measurable results across multiple real-world deployments, including reducing response times from minutes to seconds and achieving 90% memory optimization in LLM fine-tuning. Seeking roles in Al Engineering, Machine Learning, Intelligent Systems, or Agentic Al Development.

EXPERIENCE

Least Action Company Sep '25 — Present
Al/ML Developer Vellore

- Developed multiple production-grade Al 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.

Graditwin Feb '25 — May '25
Machine Learning Intern chennai

. Built an Al 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.

Strydo Technologies Jun '23 — Jul '23
Al Developer Intern 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.

PROJECTS

Multimodal RAG System Jul '25 — Aug '25

- Developed a full-stack, multi-modal Al 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) Jun '25 — Jul '25

- Developed a real-time Al 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 and integrated cloud-based noise cancellation (BVC) to enhance audio quality.
- Delivered a responsive, human-like Al 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 May '25 — Jun '25

- The project aimed to develop an Al-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 Al Agent framework with MCP server (mcp fetch server and firecrawl) integration to manage Al communication asynchronously, ensuring efficient and scalable Al 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 Apr '25 — May '25

- Built a Customer Churn Prediction System for telecom data to identify at-risk customers, addressing revenue loss from customer churn by predicting likelihood of churn 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, Tableau
Development & Databases Python, SQL, React, TypeScript, Streamlit, FastAPI, PostgreSQL, MySQL, Psycopg2
Languages English

EDUCATION

B.Tech in Artificial Intelligence &Data Science, Annai Mira College of Engineering and Technology - 8.51 Ranipet - 2025