YUVAN RAJ R

Phone: +91-9344321132 | Email: yuvanraj1132@gmail.com

PROFILE SUMMARY

- Results-driven AI/ML Engineer & **Agentic AI Architect** with 4 years of experience in designing, deploying, and scaling advanced machine learning and LLM systems across banking, insurance, telecom and enterprise domains.
- Specialized in building autonomous, multi-step conversational agents using **LangGraph**,Multi-Agent Orchestration Systems stateful workflows,parallel asynchronous workflow orchestration and tool-calling architectures.
- Hands-on expertise in **MCP (Model Context Protocol)** for secure agent-tool integration and **LangSmith** for LLMOps, evaluation, tracing, and production monitoring.
- Implemented Elasticsearch-based long-term memory + **RAG pipelines**, enabling context-aware, persistent interactions for enterprise-grade AI assistants.
- Proficient in fine-tuning LLMs, optimizing prompt engineering, and deploying scalable inference systems with multi-GPU acceleration.
- Strong background in data annotation, dataset creation for YOLO models, and computer vision pipelines using OpenCV, PyTorch, TensorFlow, and Detectron2.
- Proven ability to collaborate with clients, understand domain-specific requirements, and deliver impactful AI solutions that improve business efficiency and decision-making.
- Skilled in ChatGPT, Claude, Gemini, Ollama, LLaMA models, LangChain, LangGraph, LangSmith, SQL/MySQL, PostgreSQL, Pinecone, ChromaDB, and Elasticsearch vector databases.

TECHNICAL SKILLS

- **Programming & Frameworks:** Python, OpenCV, Flask, FastAPI , WebSockets, scikit-learn, Keras, PyTorch, TensorFlow, Gradio, Streamlit, LangChain, LangGraph, vLLM, Hugging Face, Pandas, NumPy, BERT, NLTK, YOLO, LlamaIndex, Vector Databases, MCP (Model Context Protocol), LangSmith (LLMOps)
- **Database:** MySQL, SQL, PostgreSQL,VectorDB (Pinecone, ChromaDB), Elasticsearch
- **data annotation Tools:**VoTT, LabelImg
- **Tools:**Git,Docker,Docker Compose, Jenkins, Jira, Multi-GPU Inference (Accelerate),VSCode, Copilot, Cursor
- **Awarness**:AWS(EC2,S3Sagemaker),Spring Boot,Java

WORK EXPERIENCE

Senoir AL/ML Architect | Prodapt | June 2025 - Present

- Led a team of AI/ML engineers, managing end-to-end development cycles, sprint planning, code reviews, and technical decision-making for production-grade AI systems.
- Architected and developed an **Agentic Conversational AI System using LangGraph**, enabling autonomous, AI Safety Guardrails, multi-step workflows with tool-calling and state management.
- Integrated **MCP (Model Context Protocol)** servers to securely connect LLM agents with enterprise tools, APIs, and business systems, ensuring standardized and scalable interactions.
- Implemented **Elasticsearch-based long-term memory + RAG pipelines**, enabling context retention, document retrieval, and persistent multi-session conversation capabilities.
- Designed **LLMOps pipelines using LangSmith**, including tracing, evaluation, debugging, and performance monitoring for iterative model and workflow improvement.
- Conducted code reviews, provided mentorship, and ensured adherence to best practices in agent design, workflow orchestration, and scalable LLM system development.
- Improved system reliability by building deterministic, testable state flows and reducing failure rates in multi-step autonomous agent tasks

Software Engineer | NCS Softsolution Pvt Ltd | Mar 2022 – June 2025

- Collaborated with R&D teams to analyze requirements, deliver innovative solutions, and create **Proof-of-Concept (PoC) prototypes.**
- Fine-tuned Large Language Models (LLMs) like **LLaMA** for specific applications, optimizing hyperparameters to improve model performance and adapt to domain-specific needs using **LoRa**.
- Implemented multi-GPU for LLM inference using the **Accelerate** package to enhance performance and reduce computational costs, resulting in efficient and scalable solutions.
- Trained computer vision models by collecting and annotating image data for tasks like object detection, classification, and segmentation.
- Worked on tabular data, image-text, and multi-language **ASR systems**, developing models for text recognition and transcription in images and audio

PROJECTS UNDERTAKEN

Project 1: **Enterprise Agentic Conversational AI Platform**

Technologies: LangGraph, MCP, LangChain, Python, Elasticsearch, LangSmith, FastAPI, VectorDB, React Agent

Overview: Built an enterprise Agentic AI platform using LangGraph, MCP, and Elasticsearch to enable autonomous multi-step workflows, secure tool integrations, and persistent, context-aware conversations.

- Architected a production-grade Agentic Conversational AI system using **LangGraph**, enabling autonomous multi-step reasoning and dynamic tool execution.
- Integrated **MCP (Model Context Protocol)** servers to provide secure, standardized connectivity between LLM agents and enterprise APIs, databases, and internal business tools.
- Developed a long-term memory + RAG pipeline with Elasticsearch to support persistent conversations, historical recall, and context continuity.
- Implemented **LLMOps workflows using LangSmith**, including evaluation, tracing, debugging, and latency monitoring.
- Built and optimized stateful agent workflows, improving reliability, reducing task failures, and enabling consistent end-to-end task completion.
- Led a team of AI/ML engineers by conducting code reviews, managing merge requests, mentoring team members, and driving architectural decisions for scalable deployment

Project 2: **Generative AI - Rag pipeline for Insurance Document Automation**

Technologies: Fast API, Opencv, Langchain, Pytorch, CUDA, VectorDB, Transformer model

Overview: Built a system to convert insurance documents into structured data using LLaMA 3.2 LLMs, RAG with LangChain, and ChromaDB.

- Utilized LLaMA 2 LLM models to effectively transform complex insurance documents into organized data, enabling accurate information extraction and structured representation.
- Developed a Retrieval-Augmented Generation (RAG) system with LangChain for efficient data retrieval and generation.
- Implemented ChromaDB as a vector database, enabling fast and accurate document similarity searches using optimized embeddings.
- Implemented optimized chunking strategies, metadata filtering, **Hybrid Search**, and **Re-Rank** fine-tuning parameters such as chunk size, overlap, retrieval top-k, and similarity thresholds to enhance retrieval relevance and ensure efficient handling of complex insurance documents.
- Successfully integrated extracted data into SQL databases, enabling organized storage and easy retrieval for further analysis.

Project 3: **vKYC - Automation**

Technologies: Fast API, Opencv, Transformer model, Pytorch, CUDA, TensorFlow, Hugging Face, Docker, Yolov8

Overview: Led the integration of speech recognition and image detection technologies into the vKYC process, revolutionizing customer verification methods for improved compliance and efficiency.

- Utilized YOLO to detect PAN cards, Aadhaar cards, signatures, and count individuals in real-time from video stream, streamlining the verification process.
- Implemented FastAPI async for handling multiple requests and processing multi-file uploads, optimizing the vKYC workflow and ensuring smooth parallel processing of large documents.
- Developed CNN models for signature similarity assessments, ensuring robust verification accuracy.
- Employed a fine-tuned Whisper model to deliver precise Hindi and Tamil speech-to-text conversion.
- Deployed Docker using Docker Compose to manage multiple containers, enabling seamless orchestration with multi-GPU support for enhanced computational efficiency.

AWARDS & CERTIFICATIONS

- Successfully developing vKYC module using AI technology for Kotak Bank
- AI / ML Certified From Guvi Institute

PERSONAL DETAILS

- **Nationality:** Indian
- **Languages Known:** English, Tamil