MOHAN GANESH

AI Engineer | Generative AI & Full-Stack Developer | B.Tech CSE '27

Passionate AI Engineer and Computer Science Honors student (CGPA **9.5**/10) skilled in Generative AI, Machine Learning, and Full-Stack Development. Experienced in building multi-agent systems, RAG-powered chatbots, and production-ready AI solutions with proven industry internship experience.

mohanganesh165577@gmail.com | +91-6302375782 | github.com/mohanganesh3

EDUCATION

B.Tech in Computer Science
Indian Institute of Information Technology, Sri City, India

08/2023 - Present *CGPA: 9.5 / 10*

WORK EXPERIENCE

AI Intern
Kasolco

05/2025 - 07/2025

Key Contributions

- Engineered an enterprise-grade AI retail shelf vision analytics system for **Vi-John cosmetics company**, transforming a failed **67%** accuracy project into a production-ready solution, achieving **95% accuracy** (100% in optimal conditions). Implemented breakthrough multi-modal architecture: custom **YOLOv10** trained on 11GB SKU-110K dataset → hybrid **ResNet-34** + **OCR embeddings** (0.7:0.3 ratio) → **FAISS** similarity search → real-time product classification with automated gap detection, enabling **automated inventory management** and eliminating manual shelf auditing.
- Engineered a multi-agent AI chatbot for a fitness application using **Google Gemini LLM** with 5 specialised agents (Chief Strategist, Health Knowledge Council, Plan Generator, User Profile, Mental Wellness) and **ThreadPoolExecutor parallel processing**, achieving **3x response time** improvement. Developed **full-stack solution** with FastAPI backend and React.js frontend, implementing **Pydantic** models for **structured diet plan** (daily meals with calorie/macro breakdowns) and **fitness plans** (weekly exercises with sets/reps/rest periods)tailored to specific user profiles, goals, and preferences.
- Contributed to multiple healthcare AI applications during internship: developed HealthAdvisorAI-POC using **quantised LLaMA-3.2-1B** with FastAPI and prompt engineering for clinical recommendations; built ClarityScan medical imaging platform with **Google Genkit** for X-ray/MRI/ultrasound diagnostics featuring confidence scoring and heatmap visualisations; created Health & Wellbeing AI Companion using **LangChain multi-agent framework** and Google Gemini for personalised BMI calculations, stress assessment, fitness planning, and mental wellness support. Implemented RESTful APIs, conversation memory management, and real-time processing capabilities across all three healthcare AI solutions.

SKILLS

Python | Machine Learning | Deep Learning | React.js | CV | NLP | LangChain | SQL | C++ (DSA) | MERN Stack | LangGraph | RAG | Vector Databases | Fine-tuning | CrewAI | FastAPI | MCP | Docker | Git | YOLO | MongoDB | PyTorch | TensorFlow | Kubernetes

PERSONAL PROJECTS

BrewBuddy – Multi-Agent Coffee Shop AI Assistant

- Architected a **multi-agent pipeline** (Guard, Classification, Details, Order-Taking, Recommendation) using **Google Gemini Pro** and **Pinecone vector database** for semantic search and intelligent routing
- Implemented **market basket analysis recommendation engine** with **Apriori algorithm** for personalized product suggestions based on purchase pattern analysis
- **Deployed** serverless architecture on **RunPod** with **Next.js** frontend, enabling real-time conversational ordering and. **dynamic cart management.**

Maritime AI: LoRA Fine-tuned LLM for Ship Gear

- Implemented **LoRA fine-tuning on Mistral-7B** model using Unsloth framework with **4-bit quantization**, PyMuPDF text extraction, and **custom dataset curation** from 50+ maritime PDFs, achieving **parameter-efficient domain adaptation** for ship equipment guidance and technical documentation
- Engineered **production-ready inference pipeline** with PyTorch, gradient checkpointing, memory optimization, and optimized generation parameters (temperature=0.7, top_p=0.9, repetition_penalty=1.5), **enabling offline maritime Q&A system** with **6.2GB VRAM efficiency**

RAG-Powered Web Content Q&A Chatbot

- Built **Retrieval Augmented Generation pipeline** using **LangChain**, **LLaMA-3-70B**, **FAISS** vector database, and **GROQ API** with **BeautifulSoup** web scraping, text chunking, vector embeddings, and similarity search for context-aware question answering over dynamic web content
- Developed **interactive Streamlit web application** with real-time content processing, chat history management, sample question generation, custom CSS styling, and session state handling for scalable document retrieval and conversational AI interface