

Jai Dhingra

☎ +91 8279217748 ✉ jaidhingra402@gmail.com 🔗 linkedin.com/in/jai-dhingra 🌐 github.com/jaidh01

Education

Jecrc Foundation

B.Tech (Artificial Intelligence & Data Science) – CGPA: 7.21

November 2022 – April 2026

Jaipur, Rajasthan

S.V. Public School, Jaipur

XII Passed C.B.S.E. Board: 82.8%

2022

Experience

Humanli.AI

Machine Learning Engineer

May 2025 – Present

Jaipur, Rajasthan

- Leading production GenAI platform Data on Demand (DOD) for enterprise data intelligence.
- Built RAG pipelines with LangChain, FAISS/Chroma using Mistral & Gemma.
- Designed multi-agent query routing and document reasoning workflows.
- Built PDF/image ingestion using Mistral OCR with chunking and embeddings.
- Fine-tuned Gemma (270M, 1B) using LoRA for domain QA.
- Deployed models on Azure VM + Docker via Azure AI Foundry and Azure DevOps.
- Reduced inference latency by 41% using load balancing.
- Built MCP server integrations; implemented short-term memory using Redis and long-term memory using PostgreSQL.
- Designed system and architecture for scalable AI platform components.
- Set up and maintained CI/CD pipelines via Azure DevOps for automated model deployment.

Semi-Hyphen

Machine Learning Intern

June 2024 – August 2024

Remote

- Built a Yoga Pose Detection system for a mobile application using MediaPipe and MobileNetV2.
- Achieved 88% accuracy across 8 distinct yoga poses.

Projects

Post-Discharge Medical AI Assistant (Multi-Agent RAG) | [Link](#)

Python, LangGraph, FastAPI, FAISS, Streamlit, HuggingFace

Nov 2025

- Built multi-agent RAG system for 25+ discharge reports with tool-based routing.
- Clinical agent used RAG over nephrology PDFs with citations + web fallback.
- Vector search via FAISS + sentence-transformer embeddings; Streamlit UI and FastAPI backend.

AI Call Assistant | Web Speech API, FastAPI, Ollama (Phi-3 Mini), Azure VM, ngrok | [Link](#)

June 2025

- Built real-time voice AI using STT/TTS with Phi-3 Mini via Ollama.
- FastAPI backend + HTML frontend integrated with Web Speech API.
- Deployed on Azure VM (B2ms) with ngrok live demo tunnel.

FastSAM Real-Time Inference | Python, PyTorch, ONNX, CUDA | [Link](#)

October 2024

- Upgraded Meta's Segment-Anything model for live video.
- Packaged as a Flask+Docker service with health-checks and auto-scaling scripts for edge-device deployment.

Skills

Languages: Python, C, C++

GenAI: RAG, LangChain, LangGraph, LoRA, HuggingFace, Mistral, Gemma

ML: PyTorch, TensorFlow, Scikit-Learn

Backend: FastAPI, Flask, Streamlit

Data/Vector: Pandas, NumPy, FAISS, ChromaDB, MongoDB, Redis, PostgreSQL

Cloud/DevOps: Azure VM, Docker, Git, Ollama, CI/CD, Azure DevOps

Achievements

Got 133rd Rank in GSSOC-extd 2024.

Certified in Postman API Fundamentals by Postman.