

Ashish Soni

+91 8076893497 | ashishsoni243k@gmail.com | ashish-soni-engineer | ashishsoni-ai

Summary

ML/AI engineer with hands-on experience building production-grade agentic systems, RAG pipelines, and computer vision applications. Conducting independent research on multi-agent LLM reasoning for medical false alarm reduction, targeting IEEE Access. Seeking ML/AI Internship roles.

Education

Guru Gobind Singh Indraprastha University, Delhi

Expected 2028

B.Tech — Artificial Intelligence & Machine Learning CGPA: 9.19/10.0

Coursework: DSA, Machine Learning, Deep Learning, NLP, Computer Vision, DBMS, Probability & Statistics

Technical Skills

Languages	Python (Advanced), C, C++, SQL
ML / DL	Regression, Classification, CNN, RNN, Transformers, Transfer Learning, Model Evaluation
GenAI & LLMs	RAG, LangChain, LangGraph, Prompt Engineering, HuggingFace, Groq API, Semantic Search, Embeddings
Computer Vision	YOLOv8, YOLOv8-Pose, Object Detection, Pose Estimation, OpenCV, Data Augmentation
Tools & DBs	FastAPI, Streamlit, Gradio, ChromaDB, FAISS, Git, GitHub, HuggingFace Spaces, VS Code
Libraries	TensorFlow, Keras, PyTorch, Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn

Research

MedGuard: Multi-Agent LLM Reasoning for False Alarm Reduction in Emergency Detection

LangGraph · YOLOv8-Pose · Gemini 2.0 Flash · FastAPI

Manuscript in preparation — IEEE Access

Designing a 5-agent LangGraph pipeline (Perception → Context Memory → LLM Reasoning → Decision → Action) to suppress false positives in camera-only out-of-hospital emergency detection, replacing brittle rule-based threshold logic
Gemini 2.0 Flash performs chain-of-thought reasoning over pose trajectories and scene context; asymmetric thresholds ($T_{escalate} = 0.60$, $T_{suppress} = 0.85$) minimise false negatives in safety-critical escalation decisions

Projects

AI Research Agent LangGraph · Tavily · Mistral AI · Streamlit

GitHub Live Demo

Architected a 4-node stateful LangGraph pipeline — Search, Scrape, Write, Critic — with typed state schema and conditional edge routing for autonomous web research and structured report synthesis
Critic agent scores drafts against a structured rubric via Mistral AI and triggers automatic rewrites; Streamlit frontend delivers token-level streaming output with downloadable final reports

PDF Chat Assistant LangChain · ChromaDB · Groq · Gradio

GitHub Demo

Full RAG pipeline: recursive text splitting → HuggingFace sentence-transformer embeddings → ChromaDB vector storage → top-k retrieval → LLaMA 3 (Groq) generation with windowed `ConversationalRetrievalChain`
Deployed to HuggingFace Spaces with persistent storage; resolved Gradio 4.x breaking changes and migrated from Gemini to Groq for reduced latency

AI Medical Emergency Detector Python · YOLOv8-Pose · OpenCV · FastAPI · aiomsmtplib

GitHub

Modular 5-agent architecture (Perception → Context Memory → LLM Reasoning → Decision → Action) orchestrated via LangGraph; each agent isolated in its own module with shared Pydantic state (`MedGuardState`)
YOLOv8n-Pose extracts keypoints per frame; a rolling 10-frame buffer in `a2_context_memory.py` computes pose trajectories fed to Gemini 2.0 Flash for chain-of-thought reasoning in `a3_llm_reasoning.py`
Async alert system (aiomsmtplib + FastAPI) dispatches annotated frames with event metadata on confirmed detection; daily JSONL decision logs and a 30-scenario evaluation harness support reproducible benchmarking

Certifications

Intermediate Machine Learning — Kaggle

Data Science & Analytics — HP (Hewlett-Packard)

Achievements

Authored research paper "MedGuard: A Multi-Agent LLM Reasoning Pipeline for False Alarm Reduction in Camera-Based Out-of-Hospital Medical Emergency Detection" — empirically comparing traditional rule-based AI against a 5-agent LangGraph system; targeting IEEE Access publication
Maintained CGPA 9.19/10 in B.Tech AIML (2nd year), consistently ranking among top performers in the batch
Built and deployed 3+ end-to-end AI/ML projects with live demos, GitHub documentation, and production-ready codebases