Sherla Shiva Sai

Hyderabad, India | +91 7780734852 | shivasaisherla9@gmail.com
GitHub: github.com/sherlashivasai | LinkedIn: linkedin.com/in/shivasai

Objective

Final-year AI/ML undergraduate with published research in precision agriculture and cybersecurity. Experienced in building production-grade AI/ML systems using LLMs, NLP, Deep Learning, and Generative AI models. Passionate about cloud-native development, AI agent orchestration, scalable inference pipelines, and DevOps for machine learning. Eager to contribute to NVIDIA's LLM NIM Engineering team with a focus on containerized ML deployment, optimization of transformer-based models, and cutting-edge research in multi-agent systems.

Education

B.Tech in Computer Science (AI & ML)

Siddhartha Institute of Engineering & Technology (Affiliated to JNTU-H)
Dec 2021 – Jun 2025

- Best Student Research Paper Award (ICMDRI-2025)
- Relevant Courses: Deep Learning, Natural Language Processing, Statistical Inference, Speech & Video Processing, Artificial Intelligence

Skills

- **Programming Languages**: Python, SQL, C
- **AI/ML & LLMs:** Supervised Learning, Deep Learning, Generative AI, Natural Language Processing, Transformers, RAG, LangChain, LangGraph, Groq API, Hugging Face, OpenAI APIs, CrewAI, Ollama
- **Frameworks & Libraries:** TensorFlow, Scikit-learn, Flask, Streamlit, STT, TTS, NLTK, REST APIs
- **Cloud & DevOps:** Google Cloud Platform (GCP), Docker, Git, GitHub, CI/CD, System Design
- **Data Tools & Visualization:** Power BI, Pandas, NumPy, Data Cleaning, Feature Engineering

Projects & Research

MCP-AI: Multi-Agent Control Platform for Autonomous Agriculture

GitHub: github.com/sherlashivasai/MCP-AI

- Engineered a modular, agentic AI system integrating Groq LLMs, IoT sensors, and CrewAI for crop management.
- Implemented real-time data ingestion, predictive analytics, and decision-making with Streamlit dashboard UI.

Conversational AI Assistant

GitHub: github.com/sherlashivasai/Conversational-AI-assistant

- Designed a real-time voice assistant using GPT-3.5, LangChain, Whisper STT, and TTS pipelines.
- Deployed via GCP microservices with containerized modules; 87% STT accuracy and prompt response time.

Medical QA Chatbot using RAG + Gemma3 LLM

GitHub: github.com/sherlashivasai/Medical-Chatbot.git

- Developed a Retrieval-Augmented Generation (RAG) system for medical queries using Gemma3-9B.
- Integrated with LangChain, FAISS vector store, and Streamlit frontend; achieved 82% answer relevance.

Crime Prevention via CCTV Network Analysis

GitHub: github.com/sherlashivasai/Smart-City-Surveillance.git

- Created anomaly detection pipeline for smart surveillance using object tracking and predictive modeling.
- Delivered 84% threat detection accuracy for real-time public safety monitoring.

Research Publications

- MCP-AI: Multi-Agent Control Platform for Autonomous Crop Management – ICMDRI-2025
- Phish Catcher: Client-Side Defence Against Web Spoofing Using ML – ICMDRI-2025

Hackathons & Competitions

- Participated in 10+ hackathons including IIT-level AI/ML innovation sprints
- Focused on LLMs, generative AI applications, ML deployment, and inference optimization

Certifications

- Deep Learning – Neuralearn.AI
- Generative AI Foundations – KrishAI Technologies (Udemy)
- Agentic AI Systems – Ed Doner (Udemy)
- Computer Vision & Image Recognition – Rajeev D Ratan
- AI & Robotics Workshop – IIT Hyderabad (TechnoGyan)