Satya Akhil Galla

+1 8575404550 | akhilg@bu.edu | linkedin.com/in/satya-galla | github.com/satyagalla

Summary

AI Engineer specializing in End-to-End Generative AI, from data engineering to cloud deployment. Expertise includes architecting scalable RAG systems on AWS, fine-tuning Large Language Models (Llama-3, Phi-3) using QLoRA, and engineering autonomous AI Agents for edge latency. Proven researcher with IEEE publications in multimodal fusion and graph convolutional networks.

Education

Boston University **Jan 2026**
Master of Science in Artificial Intelligence *GPA: 3.7/4.0*

IIIT Sri City **May 2024**
Bachelor of Technology in Computer Science *GPA: 3.5/4.0*

Technical Skills

Generative AI & LLMs: Llama-3, Phi-3 (Quantization), RAG Architectures, AI Agents (LangGraph), Neuro-Symbolic AI, Vector DBs
Cloud & Engineering: AWS (EC2/S3), Azure, Docker, CI/CD, FastAPI, SQL, TCP/IPC Sockets, Bash Scripting, Linux, Git
Data Science & ML: XGBoost, GCNs, Time-Series Forecasting, Pandas, NumPy, Scikit-learn, OpenCV, Regex Parsing

Experience

ISRO (Indian Space Research Organization) **Dec 2022 – May 2023**
Machine Learning Engineer
- Engineered a **Latent Graph Convolutional Network (GCN)** to process hyperspectral data, utilizing an Autoencoder for dimensionality reduction on the Chandrayaan-2 dataset.
- Devised a **Spectral-Spatial Non-Linearity** graph formulation to handle irregular lunar topography, achieving **90.1% accuracy** and validating results via an **IEEE WHISPERS 2023** publication.

Aaizel International Tech **Feb 2024 – May 2024**
Machine Learning Engineer
- Architected a **multimodal fusion pipeline (IFCNN)** merging IR+RGB satellite imagery and constructed the domain's **first remote sensing scene graph dataset**.
- Fine-tuned a **RelTR Transformer** for geospatial relationship extraction, deploying the inference engine on **Microsoft Azure** for real-time semantic visualization.

Terrafic Inc **June 2023 – Nov 2023**
Computer Vision Engineer
- Integrated the **Segment Anything Model (SAM)** to automate maritime feature extraction, delivering a production-ready MVP that secured initial client pilots.
- Implemented **Super-Resolution (SR) algorithms (4x/6x)** to enhance object detection performance on low-resolution defense mapping data.

Boston University **Sep 2025 – Jan 2026**
Graduate Teaching Assistant – CS 365 Foundations of Data Science
- Led technical labs on **Python optimization** and **statistical modeling**, mentoring 50+ students in data science fundamentals and algorithmic best practices.
- Facilitated weekly code reviews to enforce production-standard coding styles and debugging techniques.

Key Projects

Enterprise RAG System for Clinical Decision Support | *Llama-3, AWS, Docker* | GitHub **Sep 2024 – Dec 2024**
- Architected a **Hybrid RAG pipeline** on **AWS EC2** to eliminate clinical hallucinations, orchestrating ingestion for **50k+ documents** via **LangChain** and **PostgreSQL**.
- Engineered a dual-retrieval mechanism combining **Dense Vector Search** and **Best Matching 25 (BM25)** with **Cross-Encoder Re-ranking** to strictly filter context and capture precise medical nomenclature.
- Fine-tuned **Llama-3-8B** using **QLoRA** on synthetic instruction sets, achieving an **87% SBERT alignment score** (+38% over baseline) and deploying via **Docker/FastAPI**.

CoCo: Neuro-Symbolic Desktop Companion | *Phi-3, PyTorch, TCP/IPC* | GitHub **Dec 2025 – Jan 2026**
- Developed a **bio-mimicking Agentic AI architecture** to replace static desktop pets, designing a dual-layer system that separates high-speed **spinal reflexes** (60Hz reactive loop) from slow **cortical reasoning** (LLM).
- Integrated **Phi-3 Mini (4-bit)** as the cognitive brain, using asynchronous **TCP sockets** to fetch complex reasoning without blocking the real-time animation loop.
- Implemented a **Leaky Integrate-and-Fire (LIF)** spiking neural network to model internal emotional states, enabling the agent to dynamically react to user activity patterns.

Generative Audio: Transformer vs. Diffusion | *AudioLDM* | GitHub **Jan 2025 – May 2025**
- Conducted a comparative analysis of **8 hybrid architectures**, identifying the optimal trade-off between semantic alignment (Transformers) and audio fidelity (Diffusion).
- Engineered a cascaded pipeline achieving a peak **CLAP score of 0.35** and **FAD of 6.55**, validating the superiority of separate structure-texture generation phases.

Multimodal Skin Cancer Detection System | *PyTorch* | GitHub **Sep 2024 – Dec 2024**
- Designed a **Voting Ensemble** fusing CNNs (image) and tabular models (metadata) to detect malignancy in a dataset with extreme **0.1% class imbalance**.
- Surpassed single-modal baselines by **15 points (0.96 ROC-AUC)** by applying **synthetic oversampling (SMOTE)** and rigorous cross-validation strategies.

Select Publications

Efficient Graph Formulation and Latent Space Integration for Lunar Hyperspectral Image Classification – *IEEE WHISPERS 2023*
- Proposed a **Latent Graph Convolutional Network (GCN)** utilizing an **Autoencoder** for dimensionality reduction, achieving **90.1% accuracy** on Chandrayaan-1 benchmarks by engineering a spectral-spatial non-linearity function.

Enhancing Hyperspectral Classification through GCNs with Adaptive Graph Construction – *NASA NESF 2023 (Poster)*
- Presented a comparative analysis of **adaptive graph constructions** for Chandrayaan-2 IIRS data, identifying non-linear spectral-spatial connections as optimal for handling irregular lunar topography.