Satya Akhil Galla

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

Summary

ML engineer specializing in agentic systems, LLM inference optimization, and RAG pipelines. Shipped a medical inference engine achieving **0.529s TTFT** via AWQ profiling and a productivity plugin with **200 active daily users**. IEEE-published in graph neural networks; available immediately on OPT.

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), Multi-Agent Orchestration, vLLM, Neuro-Symbolic AI, Vector DBs
Cloud & Engineering: AWS (EC2/S3), Azure, Docker, CI/CD, FastAPI, SQL, Bash Scripting, Linux, Git
AI-Assisted Development: Cursor, GitHub Copilot, Claude Code
Data Science & ML: Pandas, NumPy, Scikit-learn, OpenCV

Experience

Boston University **Sep 2025 – Jan 2026**
Graduate Teaching Assistant – CS 365 Foundations of Data Science
- Led technical labs on Python optimization and statistical modeling for 50+ students; students improved from writing slow iterative code to vectorized NumPy/Pandas implementations.
- Ran weekly code reviews focused on debugging and production-standard practices, reducing common submission errors by catching issues before final submission.

Aaizel International Tech **Feb 2024 – May 2024**
Machine Learning Engineer (Internship)
- 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 and deployed the inference engine to Microsoft Azure, enabling real-time semantic visualization.

Terrafic Inc **June 2023 – Nov 2023**
Computer Vision Engineer (Internship)
- Integrated the Segment Anything Model (SAM) and OpenCV to automate maritime feature extraction; delivered 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.

ISRO (Indian Space Research Organization) **Dec 2022 – May 2023**
Machine Learning Engineer
- Boosted hyperspectral classification accuracy on the Chandrayaan-1 dataset by **13.5%** compared to CNN baselines by engineering a **Graph Convolutional Network (GCN)** utilizing Latent Space projections.
- Devised a **Spectral-Spatial Non-Linearity** graph formulation to handle irregular lunar topography, achieving **91% accuracy** on Chandrayaan-2 data; published at **IEEE WHISPERS 2023** | Google Scholar (NASA NESF 2023 Poster).

Key Projects

Deep Research Agent | *Python, Claude API, MCP* | GitHub
- Built an autonomous agent that independently researches any topic end-to-end: searches, reads, cross-references, and synthesizes complex multi-step tasks without human input; shipped from scratch in 5 days.
- Engineered **51 tools across 10 namespaces** with subagent orchestration via agent-as-tool (isolated contexts, structured outputs), KV-cache-aligned context management, LLM-based eval harness, observability, exponential backoff, and rate limiting.
- Designed composable tool chains (search → extract → synthesize) where tools consume structured outputs of other tools; validated end-to-end coherence via LLM-based eval harness across 20+ tool call sequences.

Serverless Clinical AWQ Llama Engine | *Llama-3.1, Modal, vLLM, Pinecone, Hugging Face* | GitHub | HuggingFace Spaces
- Built a medical AI assistant that answers clinical questions from 10,000 documents in under 0.5 seconds, fast enough for real clinical use, by training a specialized model and deploying it on optimized cloud infrastructure.
- LoRA fine-tuned Llama-3.1-8B on 34k medical samples; architected a 384-dimensional Hybrid RAG pipeline (Dense BGE-Small + Sparse BM25) with MiniLM Cross-Encoder reranking, achieving **0.529s TTFT** and ~60 tokens/sec on an A10G GPU.
- Profiled AWQ vs. NF4: found NF4 bottlenecked at 35% GPU utilization from dequantization overhead; switched to AWQ fused kernels for near-100% utilization; enforced BF16 after SwiGLU activations exceeded FP16 bounds, yielding **14% USMLE accuracy gain**.

Morning OS | *TypeScript, Obsidian API, LLM APIs* | GitHub
- Shipped a productivity plugin used by **200 active users daily** that generates a personalized AI briefing each morning from notes, tasks, and goals; iterated on features through a live Discord community based on real user feedback.
- Implemented carry detection with fuzzy matching, multi-provider LLM support, graceful degradation without API keys, and full mobile compatibility without native dependencies.

CoCo: Neuro-Symbolic Desktop Companion | *Phi-3, PyTorch, TCP/IPC* | GitHub
- Built a desktop AI agent that watches your activity in real time and adapts its behavior to your emotional state, running a fast reactive layer and a slow reasoning layer simultaneously without either blocking the other.
- Integrated Phi-3 Mini (4-bit quantized) via async TCP sockets for non-blocking LLM reasoning; implemented a Leaky Integrate-and-Fire (LIF) emotional state model driving dynamic behavioral responses at 60Hz.