Sarath Chandra Lingareddy

+1 (214)-899-0983 | sarathchandralingareddy@gmail.com | LinkedIn | Github

PROFESSIONAL SUMMARY

AI Engineer with hands-on experience in Python, Machine Learning, and Generative AI, focused on building cloud-native and production-ready AI systems. Experienced in developing Retrieval-Augmented Generation (RAG) pipelines, LLM agent workflows using LangChain and LangGraph, and deploying scalable inference services with FastAPI on AWS. AWS Certified with strong foundations in MLOps, model deployment, and real-world AI application development.

EXPERIENCE

AgilePath LLC — AI Engineer Intern **May 2025 – July 2025**

- Developed a two-stage **multimodal RAG system** (Muvera + ColPali) using **Qdrant** and **AWS Bedrock**, reaching **90% retrieval relevance** across 2,000+ docs.
- Optimized **multi-vector retrieval** and metadata filtering to reduce manual information lookup time by **50%** for engineering and support teams.
- Supported **MLOps workflows** by integrating **foundation models** with custom logging and versioned prompt configurations, maintaining **99%+** service availability.
- Leveraged **AI-powered code assistants** to accelerate development and applied code sanitization and secret scanning, reducing **security vulnerabilities in LLM-backed services**.

PROJECTS

Vigil3D – Video Violence Detection Platform [Demo] | *Python, PyTorch, FastAPI, Docker, AWS EC2, S3, React, TypeScript, Vercel*

- Trained a 3D CNN video classification model for violence detection, achieving **∼93%** validation accuracy on the RWF-2000 dataset.
- Built a production-grade FastAPI inference service with **under 900ms inference time**, configurable thresholds, and health checks.
- Deployed containerized inference on AWS EC2, loading 100% of model artifacts from S3 for scalable, stateless serving.
- Developed a React + TypeScript frontend enabling video upload, end-to-end inference pipeline, and confidence-based result visualization.

Reducing Hallucinations in QA Chatbots | *Python, Hugging Face Transformers, RLHF, DPO, SQuAD 2.0*

- Fine-tuned a TinyLLaMA-based QA chatbot using **Direct Preference Optimization (DPO)** to explicitly learn when to refuse on unanswerable questions using the SQuAD 2.0 dataset.
- Designed a **balanced preference-judging pipeline** distinguishing hallucinated answers vs valid refusals, reducing hallucination rate on impossible questions from **∼18.6% to ∼4.6%** (**∼75% relative reduction**).
- Evaluated SFT vs DPO using refusal accuracy, hallucination rate, and answer accuracy, identifying an optimal DPO configuration with the best trade-off between correctness and conservatism.

TECHNICAL SKILLS

AI/ML Engineering: Supervised & Self-Supervised Learning, Deep Learning (PyTorch, Tensorflow), Reinforcement Learning, NLP, Computer Vision (YOLO), Fine-Tuning (LoRA, PEFT), Transfer Learning.

Production ML Systems: Data Pipelines, Training Pipelines & Evaluation Frameworks, Scalable Inference APIs (FastAPI, REST), Model Deployment & Serving, A/B Testing, Monitoring & Logging, Drift Detection.

LLM Systems & Agentic AI: LLM Orchestration (LangChain, LangGraph), LLM Inference Platforms (AWS Bedrock), Retrieval-Augmented Generation, Prompt Engineering, Multi-Agent Systems, Model Context Protocol, Guardrails, Hugging Face Ecosystem, Workflow Automation (n8n).

Cloud & Infrastructure: AWS (EC2, S3, Lambda, IAM, SageMaker), Docker, CI/CD (GitHub Actions), Terraform (IaC).

Programming & Databases: Python, Go, TypeScript, SQL (PostgreSQL/MySQL), MongoDB, Pinecone, Qdrant.

CERTIFICATION

AWS Certified Machine Learning – Specialty **Sep 2025 – Sep 2028**
AWS Certified Developer – Associate **Aug 2025 – Aug 2028**
HashiCorp Certified: Terraform Associate (003) **Dec 2025 – Dec 2027**

EDUCATION

University of North Texas, Denton, TX **Aug 2024 – Dec 2025**
Master of Science in Artificial Intelligence
Mahindra University, Hyderabad, India **Aug 2020 – May 2024**
Bachelor of Technology in Artificial Intelligence