

Santosh Muruggu

Dubai, UAE | Immediately Available in UAE — Open to Relocate within UAE

+971 55 340 5571 | santoshmuruggu@gmail.com | github.com/santoshmuruggu | linkedin.com/in/santoshmuruggu

Summary

AI/ML Engineer specializing in Generative AI and LLM applications, with hands-on experience designing and deploying retrieval-augmented generation (RAG) systems, autonomous multi-agent pipelines (LangGraph), and computer vision models. Shipped two production GenAI systems with rigorous evaluation harnesses (deepeval, LLM-as-judge) and full observability via Langfuse tracing. Strong foundation in Python, deep learning, and full-stack deployment (FastAPI, Streamlit, Docker). Co-authored ML research on constraint-driven PCB autorouting. Seeking an AI/ML Engineer or GenAI Engineer role in the UAE.

Skills

Programming Languages: Python, Java, JavaScript, TypeScript, SQL

AI & Machine Learning: Machine Learning, Deep Learning, Computer Vision, NLP, LLMs, RAG, Agentic AI, Federated Learning, Multi-Agent Systems, Knowledge Graphs

Frameworks & Libraries: PyTorch, TensorFlow, Scikit-Learn, LangChain, LangGraph, FastAPI, React, Next.js, Node.js

Databases & Vector Stores: PostgreSQL, MongoDB, SQLite, Supabase, ChromaDB

MLOps & Tools: Git, GitHub, Docker, Linux, Ollama, Langfuse, deepeval, REST APIs

Core Concepts: Data Structures & Algorithms, OOP, DBMS, Operating Systems, Computer Networks

Flagship Projects

Sakan — Grounded RAG Assistant for Dubai Tenancy Law | [Live Demo](#) · [GitHub](#)

- Built a hybrid retrieval-augmented generation system that answers Dubai tenancy law questions with citations to the exact legal article, merging vector search (ChromaDB) and keyword search (BM25) via Reciprocal Rank Fusion over law text chunked by article boundary.
- Resolved a legal-correctness edge case where an amendment law superseded specific articles by tagging every chunk with instrument, article number, and currency status, and excluding superseded text from the retrieval space entirely to prevent outdated-article citations.
- Built a citation-grounding verifier that checks every cited article against retrieved source text, catching hallucinated citations before they reach the user.
- Evaluated on 36 hand-written test questions (factual, edge-case, adversarial), achieving **92% refusal accuracy**, **100% citation grounding**, and a **0.99 average faithfulness score** (deepeval FaithfulnessMetric). Deployed publicly on Streamlit Cloud using Claude Haiku for cost-efficient, context-grounded generation.
- *Stack: Python, ChromaDB, BM25, Anthropic API, Streamlit, deepeval.*

Scout — Autonomous Multi-Agent Company Due-Diligence Tool | [Live Demo](#) · [GitHub](#)

- Designed a 5-node LangGraph state machine (plan → act → cluster → synthesize → critique) that researches a company across web/news search and regulatory registers (DFSA, DIFC), producing a structured, Pydantic-typed due-diligence report.
- Implemented a deterministic verification layer requiring 2+ independent source domains to mark a claim “verified” vs. “flagged,” deliberately decoupled from LLM judgment to keep the trust decision auditable.
- Instrumented full request tracing with Langfuse (per-node spans, cost/latency capture) and built retry/backoff handling for flaky search calls and malformed structured outputs.
- Evaluated on 15 UAE companies: **100% tool-call success rate**, **78.8% factual recall** against hand-verified ground truth, **~93s average latency**, and **~\$0.15 LLM cost per report**; deployed as a FastAPI backend + Streamlit UI across two services on Render.
- *Stack: Python, LangGraph, Pydantic, FastAPI, Tavily API, Langfuse, Claude Sonnet.*

Experience

InTrain Tech

Dec 2024 – Feb 2025

AI/ML Intern

- Developed an AI-powered gunny bag detection system for rice warehouses using YOLO-based object detection models.
- Created and annotated custom datasets of rice gunny bags and applied data augmentation techniques to improve model generalization.
- Trained, evaluated, and fine-tuned deep learning models to achieve robust detection performance across varying warehouse environments.
- Designed a real-time detection pipeline for automated inventory estimation and warehouse monitoring, reducing manual inspection and counting effort.

Research

SiRouteNet: Learning Signal-Integrity-Aware Cost Maps for Constraint-Driven PCB Autorouting

- Co-authored an AI-assisted PCB autorouting framework combining a U-Net-based signal-integrity prediction model with deterministic A* routing.
- Designed and evaluated deep learning models for signal-integrity risk estimation in PCB layouts, demonstrating reduced routing risk while preserving geometric efficiency.

Achievements

- Finalist — BIMSTEC AI Road Safety Hackathon 2026, Centre of Excellence for Road Safety (CoRES), IIT Madras (Jul 2026)
- 1st Place — Smart Automation, NRIIT National Hackathon 2024
- 3rd Place — Hack-to-Detect AI Surveillance System Hackathon
- Top 10 Finalist — International VIT-AP HackByte Hackathon

Education

Vellore Institute of Technology, Andhra Pradesh (VIT-AP)

Bachelor of Technology — Computer Science & Engineering (AI & ML Specialization), CGPA: 8.2/10

2022 – 2026

Amaravati, India