

Katta Nitish Goud

AI/ML Engineering Student

+91-9032239630 | nitishgoud69@gmail.com | github.com/Katta-Nitish | linkedin.com/in/katta-nitish-goud | katta-nitish.github.io |
Hyderabad, India

PROFILE

B.Tech CSE (AI & ML) student with hands-on experience building agentic AI systems, multi-stage RAG pipelines, and local LLM deployments. Focused on end-to-end projects with measurable outcomes — from data collection and model training to API deployment and user-facing interfaces.

TECHNICAL SKILLS

Languages: Python, SQL

Frameworks: LangChain, LangGraph, FastAPI, Streamlit, Pydantic v2, scikit-learn

AI / ML: Generative AI, LLMs, RAG, Agentic AI, Prompt Engineering, LLM Evaluation (RAGAS), Local LLM Deployment (Ollama), Regression Modelling, ML Pipelines

Libraries: NumPy, Pandas, LlamaIndex, HuggingFace, ChromaDB, FAISS, Requests

DevOps: Docker, Async Python, pytest, Git, GitHub

Fundamentals: Data Structures & Algorithms, OOP, DBMS, Operating Systems

Tools: Ollama, Tavily, Streamlit Cloud

PROJECT EXPERIENCE

Self-Improving RAG Agent — Evaluate, Reflect & Retry

2026

Python · LangGraph · FAISS · RAGAS · DeepSeek-R1:8b · Gemma3:12b · Ollama · Streamlit | [GitHub](#)

- Built a fully offline RAG system that self-corrects without human intervention — improving answer relevancy by 73% across retry attempts (RAGAS: 0.49 → 0.85) at zero API cost.
- Implemented a dual-LLM evaluation loop — DeepSeek-R1:8b for answer generation and Gemma3:12b as an independent judge — using a 4-node LangGraph StateGraph with adaptive top-k retrieval (k: 6 / 8 / 12) triggered when scores fall below 0.70.
- Tracked per-attempt RAGAS score, top-k value, and latency in real-time via a Streamlit sidebar, across up to 3 configurable retry attempts.

Agentic RAG System — Multi-Model Routing & Retrieval

2026

Python · LangGraph · FAISS · Cross-Encoders · Ollama · Docker | [GitHub](#) · [Demo](#)

- Closed a routing accuracy gap by layering cross-encoder reranking (BAAI/bge-reranker-base) on top of FAISS semantic search, validated via ablation testing — raised accuracy from 60% to 100%.
- Built a 5-action LangGraph routing agent (Retrieve, Clarify, Tool, Answer, Refuse) on a local multi-LLM architecture, separating the query-routing model from the response-generation model to reduce self-evaluation bias — grounded on 50 arXiv cs.AI papers.
- Integrated arXiv API, Python REPL, and calculator tools; Dockerized on python:3.11-slim with host Ollama bridge; pytest suite covers calculator, division-by-zero, and REPL errors.

Local LLM Performance Predictor

2026

Python · scikit-learn · FastAPI · Streamlit · Ollama · psutil · pynvml · Pandas | [GitHub](#)

- Identified a CPU-offload bottleneck where gemma3:12b ran 10x slower per GB VRAM than comparable 8B models due to VRAM spillover — caught via feature importance analysis (48% weight on model identity flag) on a self-collected 288-run benchmark across 8 local LLMs.
- Trained split regression models — Ridge (MAE 0.54s) for latency and Random Forest (MAE 7.6 MB / 0.32 MB) for RAM/VRAM — using ColumnTransformer with StandardScaler + OneHotEncoder, justified by 5-fold CV across 3 algorithms.
- Deployed a FastAPI backend serving 3 trained joblib pipelines with Pydantic input validation and a Streamlit frontend with real-time hardware risk assessment based on user's available RAM and VRAM.

EDUCATION

Marri Laxman Reddy Institute of Technology and Management

Expected May 2027

B.Tech — Computer Science Engineering (AI & ML) | Aggregate: 80% | Hyderabad, Telangana

Sri Chaitanya Junior College

Completed 2023

Intermediate (MPC) | 93% | Hyderabad, Telangana

CERTIFICATIONS

Generative AI Fundamentals — Outskill 2025

Python Programming — GeeksforGeeks 2024

IoT and Robotics — Launched Global 2023