

Ujjwal Reddy Kalvolu Sreenivasa Reddy

+1 812 322 3743 | ujjwalreddyks@gmail.com | ujjwalreddyks.com | linkedin.com/in/ujjwalreddyks | github.com/ujjwalredd

SUMMARY

AI Engineer building production agentic systems. Shipped a 13-agent LangGraph control tower with HMAC-signed guardrails, a hybrid RAG pipeline at 85.5% composite confidence on 511K documents, and a zero-trust MCP gateway. Strong across LangGraph, FastAPI, MCP, Anthropic and OpenAI APIs, vector search, Kafka, and OpenTelemetry.

SKILLS

AI, Agents, and LLMs: LangGraph, Model Context Protocol (MCP), OpenAI Swarm, Anthropic API, Ollama, multi-agent workflows, tool calling, SSE and WebSocket streaming, prompt engineering, prompt injection defense
RAG and Retrieval: Retrieval-Augmented Generation, ChromaDB, BM25, hybrid search, Reciprocal Rank Fusion, dense vector search, chunking strategies, LLM-as-judge evaluation, citation verification
Backend and Data: Python, FastAPI, Pydantic, Streamlit, PostgreSQL, REST APIs, Kafka, PySpark, Dagster, dbt, Delta Lake, Neo4j, TigerGraph, PuppyGraph, GraphFrames
ML and Reliability: XGBoost, Random Forest, MLflow, ROC AUC, model promotion gates, HMAC-signed corrections, audit logs, human-in-the-loop, pytest
DevOps and Observability: Docker, OpenTelemetry, Grafana, Git

WORK EXPERIENCE

Research Assistant

Indiana University Bloomington, Prof. Da Yan

Bloomington, IN

Dec 2025 to May 2026

- Built an AI-ready graph pipeline supporting a **200+ GB** StackOverflow benchmark with 9 node types, 19 edge types, and **782M edges** across Neo4j, TigerGraph, PuppyGraph, and GraphFrames.
- Reduced projected Neo4j load time from **2–3 days to under 10 minutes** using parallel batching, memory tuning, relationship ordering, and validation gates while preserving benchmark repeatability.
- Hardened preprocessing for malformed CSVs, null IDs, duplicate keys, embedded newlines, encoding issues, and timestamp drift, improving reliability for downstream graph analytics and AI search use cases.

Deep Learning Researcher

AIR Center, VIT AP University

Amaravati, India

Feb 2022 to May 2022

- Co-authored a **10-year** GAN survey covering 100+ architectures, training instability, mode collapse, evaluation metrics, and applications across vision, NLP, medical AI, and time series; published in IOP MLST.
- Developed a static Android malware detection pipeline on **1,000** APKs with feature extraction and Random Forest classification, achieving **95%** accuracy and publication in Springer ICISML.

PROJECTS

OmniCore | FastAPI, Python, MCP, OpenAI Swarm, OpenTelemetry, Streamlit, Pydantic, pytest

- Built a **zero-trust agentic gateway** over local LLMs with a three-stage guardrail stack (regex deny-list, LLM-judge classifier, execution rail) blocking prompt injection and shell exploits before MCP dispatch.
- Designed a **multi-agent Swarm engine** with a 4-strategy JSON repair pipeline for malformed small-model output and a Pydantic-validated MCP stdio server with a sandboxed read-only command executor.
- Instrumented every agent step with **OpenTelemetry spans** into a local flywheel log and built a Streamlit audit dashboard with WebSocket streaming and a **continuous fine-tuning pipeline** scoring run quality.

Autonomous Supply Chain Control Tower | LangGraph, FastAPI, Kafka, Dagster, XGBoost, MLflow, Grafana

Live

- Built a **13-agent** LangGraph and FastAPI control tower that converts raw supply-chain CSV uploads into schema-inferred pipelines, ML features, model promotion, and refreshed Grafana dashboards in **15–20 minutes**.
- Integrated Anthropic API streaming, Kafka, PySpark, dbt, Delta Lake, and Dagster workflows for agentic logistics automation across ingestion, validation, forecasting, promotion, and monitoring tasks.
- Implemented production guardrails including Kafka Guardian and MLflow Guardian agents, ROC AUC **0.60** promotion gates, HMAC-signed corrections, prompt injection defense, immutable audit logs, and human escalation.

RAG Pipeline with Hybrid Search | FastAPI, ChromaDB, BM25, Ollama, Streamlit, Docker

- Engineered a hybrid retrieval system combining **dense vector search** (ChromaDB) with **BM25 sparse search** using Reciprocal Rank Fusion, achieving **85.5% composite confidence** with inline citation verification across **511K enterprise documents**.
- Built an LLM-as-judge evaluation framework measuring **5 automated metrics** (correctness, faithfulness, retrieval recall, retrieval precision, citation accuracy) across **500 golden Q&A pairs** with chunking strategy comparison benchmarks.

EDUCATION, PUBLICATIONS, AND CERTIFICATIONS

Indiana University Bloomington

MS, Data Science | GPA: 3.52 / 4.0

Bloomington, IN

VIT AP University

B.Tech, Computer Science with AI Specialization | GPA: 8.48 / 10.0

Amaravati, India

- **Publications & Certifications:** Ten Years of GANs, IOP MLST 2024; Safeguarding Public Spaces using Edge Impulse, IEEE RMKMATE 2023; Android Malware Detection Using ML, Springer ICISML 2022, Anthropic Model Context Protocol, Mar 2026; Developing LLM Applications with LangChain, DataCamp, May 2025; Microsoft Azure ML, Jun 2023; DeepLearning.AI ML, Jun 2023; NVIDIA Fundamentals of Deep Learning, Sep 2022.