

Yadlapalli Sankar Sai Narayana

sankarsainarayana@gmail.com | +91-7207404856 | [LinkedIn](#) | [Github](#) | [Portfolio](#)

CAREER SUMMARY

AI & Cloud-focused Software Engineer with hands-on experience building **production-grade LLM applications, AI agents, and RAG systems. Certified Kubernetes Administrator (CKA)** with expertise in **AWS, Kubernetes, Docker, and cloud-native architectures**. Experienced in developing **MCP-based intelligent automation platforms**, AI-driven incident analysis solutions, and scalable infrastructure. **passionate about combining AI Engineering and DevOps to design, deploy, and maintain production-grade AI applications.**

TECHNICAL SKILLS

Programming Languages:	Python, TypeScript, Golang
BI & Visualization Tools:	Microsoft Power BI, ETL, Microsoft
AI/ML & LLM Frameworks:	LangChain, LangGraph, TensorFlow, Keras, PyTorch, Scikit-learn, HuggingFace
AI Tools & Agentic Development:	Claude Code, OpenAI Codex, Prompt Engineering, Retrieval-Augmented Generation (RAG), Vector Databases, MCP, AI Agents, Tool Calling
Backend & APIs:	Flask, FastAPI, FastMCP, REST APIs
Cloud & DevOps:	Linux, AWS (EKS, EC2, IAM, S3), Docker, Kubernetes (CKA), GitHub Actions, CI/CD
Databases:	MySQL, PostgreSQL
Core CS Fundamentals:	Data Structures & Algorithms, DBMS

EDUCATION

JNTU-GV College Of Engineering
(B.Tech in Information Technology (Honors))

November 2021 - May 2025
Vizianagaram ,Andhra Pradesh

PROFESSIONAL EXPERIENCE

DevOps Engineer Intern – Yellow AI (Yellow messenger)

Bangalore | 03/2025 – Present

- Managed production Kubernetes environments on AWS, including **cluster upgrades, monitoring, and incident response (on-call support)**, ensuring high availability and minimal downtime.
- Performed **cluster version upgrades, patching, and compatibility checks**, reducing deployment risks and improving system stability.
- Resolved more than **5 incidents** and Actively handled on-call rotations, troubleshooting real-time production issues, analyzing logs/metrics, and implementing quick remediation strategies.
- Operated **cloud-native infrastructure on AWS using Docker and Kubernetes (EKS)**, enabling scalable, containerized deployments with improved reliability and faster release cycles.

AI / Platform Engineering Contributions

- Architected an **on-demand intelligent tool-calling system** for observability and incident analysis.
- Developed an AI-driven solution that **interprets user queries (e.g., trace_id, region)** and dynamically selects the appropriate tools for execution.
- Unified **11 MCP tools** into a unified system, enabling automated **root cause analysis (RCA) and actionable recommendations without manual intervention**.
- Integrated **decision-making logic using LLM-based orchestration**, allowing the system to autonomously determine which tools to invoke based on query context.
- Accelerated debugging time significantly by enabling **instant insights into distributed traces, failures, and system anomalies**.

PROJECTS

KubeAtlas – AI-Powered Kubernetes Incident Investigation [GitHub Link](#)

June 2026

- An AI-powered Kubernetes incident investigation platform that continuously monitors cluster resources, detects unhealthy workloads, and generates structured incidents with severity and health context.
- Designed a polyglot microservices architecture using **Go, Python/FastAPI, PostgreSQL, and React**, integrating LLM-powered root-cause analysis and remediation workflows.
- Implemented an end-to-end incident pipeline covering **resource discovery, health evaluation, log/event collection, AI-driven RCA, and kubectl-based remediation recommendations** with human approval workflows
- Delivered an operator-focused dashboard with live incident updates, searchable resource views, and actionable remediation steps to reduce troubleshooting effort and time from **60 minutes to 5 minutes**

Skills: Go, Python, FastAPI, LangChain, Groq, Kubernetes, PostgreSQL, React, Docker, Incident Management

Self-Healing RAG Platform [Live](#) [GitHub Link](#)

May 2026

- Created a **self-healing Retrieval-Augmented Generation (RAG) platform** with adaptive retrieval, answer verification, and recovery workflows to improve response reliability and factual accuracy.
- Enabled **research-grade retrieval pipelines** using multi-query fusion, hybrid search (Vector + BM25), reranking, and multi-stage document/chunk filtering to enhance retrieval quality.
- Built autonomous validation mechanisms that detect weak retrievals, hallucinations, and unsupported claims, triggering re-retrieval and response regeneration with citation-backed verification.
- Deployed a production-ready API using **FastAPI, Docker, and Vector Databases**, incorporating structured logging, rate limiting, health monitoring, memory management, and scalable document ingestion workflows

Skills: Python, LangGraph, LangChain, RAG, Vector Databases, FastAPI, Docker, AI Agents, Prompt Engineering

CERTIFICATIONS

- **Certified Kubernetes Administrator (CKA)** — Linux Foundation, USA [View Certificate](#)