



SAI KRISHNA MANOHAR BALUSU

Software Engineer | Applied AI | Agentic Systems | Distributed Systems

Tampa, FL | +1 656-214-9970 | saikrishnamanohar@outlook.com | LinkedIn | GitHub | saibalusu.com

PROFESSIONAL SUMMARY

Software and AI Engineer with 6+ years across production software, cloud, data, and applied AI. Builds agentic workflows, RAG systems, evaluation infrastructure, APIs, and event-driven distributed services. Production experience includes inference APIs serving 10,000+ predictions/day and pipelines processing 500+ GB. Strong in Python, Java, TypeScript/JavaScript, FastAPI, LangGraph, AWS, PostgreSQL, Redis, Docker, Kubernetes, CI/CD, observability, and secure enterprise workflows.

TECHNICAL SKILLS

AI/Agents: OpenAI APIs, GPT, Amazon Bedrock, LangGraph, LangChain, MCP, RAG, tool calling, structured outputs, embeddings, multi-agent systems, evals, guardrails, human-in-the-loop

Software/Systems: Python, Java, TypeScript/JavaScript, SQL, FastAPI, Node.js, React, REST APIs, microservices, distributed systems, async processing, Kafka, RabbitMQ

Cloud/Data: AWS Lambda, ECS, API Gateway, SQS, EventBridge, GCP, PostgreSQL, pgvector, Redis, Pinecone, Airflow, Pandas, NumPy

Reliability/Security: Docker, Kubernetes, Terraform, Jenkins, GitHub Actions, CI/CD, OpenTelemetry, RBAC, OWASP, audit logging, testing, incident troubleshooting, code reviews

PROFESSIONAL EXPERIENCE

University of South Florida - Technology Transfer Office - Tampa, FL | *AI Engineer, Research and Innovation* | Sep 2025 - Present

- Architect Python-based AI applications and agentic workflows with LangGraph, LangChain, RAG, APIs, retrieval, structured outputs, access controls, and cloud services, translating ambiguous requirements into maintainable systems.
- Design grounded retrieval with Pinecone, PostgreSQL/pgvector, embeddings, and approved knowledge sources; develop evaluation and regression tests to catch retrieval, workflow, and safety failures before release.
- Implement contextual state, RBAC, human approvals, prompt-injection defenses, and content filtering; use application data and stakeholder feedback to drive iterative improvements.

Deloitte USI - Hyderabad, India | *Business Technology Analyst - Software, Cloud and Data Engineering* | Jan 2021 - Aug 2024

- Developed Python backend services and REST APIs for enterprise, data-intensive applications using cloud-native, microservice, and asynchronous processing patterns.
- Designed event-driven AWS workflows with Lambda, ECS, API Gateway, SQS, and EventBridge; built Airflow/Kafka pipelines across AWS and GCP for downstream applications, analytics, and ML workloads.
- Containerized services with Docker/Kubernetes, automated infrastructure with Terraform, and implemented Jenkins/GitHub Actions CI/CD for repeatable build, test, and deployment workflows.
- Supported production systems through monitoring, debugging, incident/root-cause analysis, releases, code reviews, and cross-functional technical design discussions.

Zee Entertainment Enterprises - Hyderabad, India | *Software Engineer* | Feb 2020 - Jan 2021

- Built Python inference APIs serving 10,000+ predictions/day at approximately 99% availability; evaluated and integrated PyTorch model variants into production services.
- Developed Python/SQL pipelines on GCP processing 500+ GB of structured and unstructured data; debugged API and data-processing failures across production workflows.

SELECTED ENGINEERING PROJECTS

Enterprise Agentic Workflow Platform | Python, FastAPI, LangGraph, PostgreSQL, Redis, Docker, AWS

- Designed a distributed agent workflow service for retrieval, policy validation, risk analysis, tool execution, and human approvals with persistent checkpoints, retries, failure recovery, RBAC, structured outputs, audit logging, and asynchronous execution.

AI Agent Evaluation and Observability Platform | Python, FastAPI, PostgreSQL, OpenTelemetry, CI/CD

- Built reusable eval/regression infrastructure with experiment history, configuration comparisons, tracing, metrics, latency diagnostics, and automated CI checks for safer AI/software releases.

INNOVATION AND INTELLECTUAL PROPERTY

Sole inventor - two provisional patent applications filed through USF Technology Transfer Office: distributed LLM inference with dynamic peer-to-peer resource coordination (27T008PR-CS); methods addressing bias and precision in language-model systems (27T007PR-CS).

EDUCATION AND CERTIFICATIONS

University of South Florida - M.S., Data Intelligence and Artificial Intelligence, GPA 3.65/4.00 | Aug 2024 - Aug 2026

Lovely Professional University - B.S., Information Technology | Aug 2015 - Aug 2019

Certifications: AWS Certified DevOps Engineer - Professional | Google Professional Cloud DevOps Engineer | CodePath Foundations of AI Engineering