SAI CHARAN J

saicharanj1410@gmail.com | +91 9743395352 | Bengaluru, India | https://www.linkedin.com/in/sai-charan-j/

SUMMARY

Agentic AI / Software Engineer with 4.5 years of experience designing, building, and scaling production AI systems. Architected an enterprise-wide LLM observability platform monitoring 150+ applications and ~4 million events daily. Built agentic pipelines with human-in-the-loop workflows on AWS AgentCore, implemented hybrid search and reranking for RAG retrieval, and developed LLM-as-Judge evaluation frameworks for automated quality assurance. Strong foundation in Python, FastAPI, pgvector, Azure AKS, Docker, and CI/CD — with a consistent track record of end-to-end production ownership.

EXPERIENCE

AI Engineer, Emerging Lead — State Street
Apr 2026 – Present | Bengaluru, India

- Built 'Ask SS', an agentic chatbot MVP on AWS AgentCore that automates triage of alerts from an internal alert monitoring system - the agent analyzes SOPs to identify the root cause and recommend a resolution, with a human-in-the-loop approval step before action is taken, reducing alert closure time by 50%.
- Lead the enterprise LLM observability and evaluation platform, extending its adoption to new AI teams and use cases.

AI Engineer, Senior Associate — State Street
Apr 2025 – Mar 2026 | Bengaluru, India

- Architected an enterprise-wide LLM observability system adopted as the standard monitoring layer across 150+ production applications — LLM agents, RAG pipelines, and Responsible AI workloads — processing ~4 million traces per day in near real-time.
- Built an LLM-as-Judge evaluation framework to run controlled A/B tests of prompt and pipeline changes against gold datasets, automating detection of hallucinations, quality regressions, and toxicity across the enterprise model portfolio.
- Owned end-to-end Arize AI integration: built custom Python SDK wrappers, led cross-functional onboarding across engineering teams, and served as the primary adoption point for AI observability tooling.
- Delivered FinOps, Operational, and Enterprise Risk Management (ERM) dashboards providing real-time visibility into latency, token cost, error rates, and compliance metrics across the production AI portfolio.
- Redesigned distributed telemetry infrastructure to eliminate multi-hop trace loss, achieving zero-loss end-to-end tracing and significantly improving production incident diagnosis speed.
- Validated infrastructure reliability via Locust load testing, confirming trace completeness, concurrency handling, and zero-drop behavior under peak production load.

AI Engineer, Associate 2 — State Street
Jan 2022 – Mar 2025 | Bengaluru, India

- Built core components of an internal RAG-as-a-Service platform, implementing hybrid search and reranking over pgvector embeddings to improve retrieval relevance beyond naive similarity search.
- Designed ingestion and chunking pipelines for structured data (Excel/CSV), and instrumented the end-to-end retrieval pipeline for full observability coverage.
- Built the Responsible AI observability dashboard on Azure Databricks, tracking model usage, input parameters, and ethical AI metrics to support transparency and regulatory compliance.
- Built high-performance RESTful APIs using FastAPI, optimized for scalability and low-latency responses in data-intensive AI workflows.
- Designed and implemented a Medallion Architecture (Bronze→Silver→Gold) data pipeline on Azure Databricks for observability trace data, and owned Azure DevOps CI/CD pipelines across all environments.

SKILLS

Python & FastAPI · Agentic AI (AWS AgentCore) · RAG Pipelines · Vector Search (PostgreSQL/pgvector) · LLM Evaluation (LLM-as-Judge) · Arize AI · OpenTelemetry / OpenInference · LangChain · Kubernetes (Azure AKS) · Docker · Azure & Azure Databricks · CI/CD (Azure DevOps) · System Design & Scalability · Locust Load Testing

EDUCATION

B.E. (Honours), Electronics & Communication Engineering — Bangalore Institute of Technology
2018 – 2022 | Bengaluru, India | CGPA: 8.72 / 10 · Graduated with Honours