SHRINIWAS RAMESH SURAM

+1 940-279-6746 |shriniwassuram@gmail.com |linkedin.com/in/shriniwassuram |github.com/Shriniwas410

ABOUT ME

Lead AI Software Engineer specializing in Enterprise-scale Generative AI system architecture and distributed ML infrastructure. Proven track record of spearheading "0-to-1" organizational AI adoption, model optimization, fault-tolerant data pipelines, and Agentic frameworks. PhD student with active research in AI-driven HCI, enterprise AI adoption, and applied data intelligence.

EXPERIENCE

U.S. News & World Report Sep 2024 – Present

Lead AI Software Engineer

- **Acted as the Lead AI Engineer and primary Point of Contact (POC)** for cross-functional product teams across U.S. News and its subsidiaries. Led the end-to-end design, architectural planning, and execution of a centralized AI hosting platform, enabling seamless integration of GenAI features across 5+ product lines.
- **Directed a distributed engineering pod** consisting of 2 SWEs and an external vendor team. Defined the long-term AI product strategy, delegated development tasks, established rigorous Agile sprint cadences, and enforced system health standards through comprehensive Python code reviews.
- **Engineered a multi-modal Generative AI system for** the paid Academic Insights platform. Actively gathered requirements from stakeholders and utilized synthetic data generation for robust model evaluation and continuous tuning, **and scaled 100K+ active daily users, driving a projected ~$4M in new annual revenue.**
- **Grew a production Live AI Agent from 1M+ daily user interactions,** orchestrating model iteration, infrastructure scaling, and real-time monitoring to sustain sub-second response times and high user satisfaction at a massive scale.
- **Fine-tuned a LLM model and optimized inference performance** by implementing quantization (INT8/FP16) and knowledge distillation on GPU clusters, reducing serving latency by 45% and cloud compute costs by 30%.
- **Pioneered an original framework for Explainable AI (XAI) and Model Evaluation,** developing automated offline/online metric tracking to evaluate LLM alignment and robustness against distribution shifts.
- **Architected and scaled bespoke AI data pipelines** on the cloud, prioritizing collaborations with subsidiary CollegeAdvisor to identify requirements and accelerate their internal onboarding operations by 40%.
- **Developed and deployed AI Agent operational applications** to optimize internal workflows, automate task assignments, and generate weekly summaries from Jira for enhanced analysis and transparency, 50% reduction in time spent on toilsome tasks, achieving a ~2000+Hr in operations, amounting to ~$80k+ in annual opex budget.
- **Developed custom embedding techniques** for improved knowledge representation and reasoning in AI systems, reducing hallucinations and improving context understanding.

Greedy AI Feb 2024 – Sep 2024

AI Engineer

- **Architected robust enterprise Agentic Frameworks** (LangGraph, standardizing Tool Use) combined with Knowledge Graph RAG, resolving complex, multi-step user queries autonomously with sub-second latency.
- **Owned the Developer UI and Full-stack integrations for educational institutions,** leveraging TypeScript, React, and GraphQL for real-time visual debugging to monitor agent logic, driving a 25% increase in developer velocity.
- **Engineered fault-tolerant MLOps and Data Processing pipelines** for large-scale, real-time data ingestion while implementing proactive data lineage, automated anomaly detection, and bias-rebalancing schemas.
- **Scaled model training infrastructure** utilizing MLOps, advanced data, and tensor parallelism across distributed TPU/GPU clusters. Reduced training epochs for multi-objective RLHF models by 35% without degrading model quality.
- **Co-authored research papers for the Army Research Laboratory** on Data and AI applications for BAA funding projects, receiving recognition from military research communities for innovative approaches to AI implementation.
- **Implemented original cloud cost optimization strategies** that saving 6 figure cost annually, receiving industry recognition for the exceptional financial impact of technical solutions.

Forbes Jul 2022 – Nov 2023

Data/ML Engineer

- **Developed and deployed deep learning propensity models** spanning 1PB+ of BigQuery data using TensorFlow and Python, yielding a 60% spike in subscription conversions.
- **Architected high-throughput, low-latency distributed microservices** to manage thousands of QPS. Optimized networking payloads and storage I/O to ensure strict millisecond SLAs for real-time inference.
- **Spearheaded the deployment of an AI-powered multi-turn LLM chatbot,** maintaining strict conversational context windows and safety guardrails, which improved active user retention by 25%.
- **Led technical-debt reduction initiatives** by architecting robust CI/CD pipelines (Kubernetes, Airflow, Terraform). Enforced strict unit and integration testing to guarantee faultless, high-availability deployments.

Forbes Jul 2021 – Dec 2021

Cloud DevOps Intern

- **Automated critical deployments of large-scale distributed systems** using Terraform, Docker, Ansible, Jenkins, and Go within Kubernetes (GCP and AWS).
- **Set up disaster recovery and fault mitigation in IAC deployments,** ensuring 99.99% system availability through automated alerting and telemetry integration.

PHD RESEARCH & PUBLICATIONS

Research Focus: *Framework for Assessing the Impact, Usage, and Risk of Artificial Intelligence in Cybersecurity and IT Compliance*

- Design Thinking, Human-Computer Interaction, and AI
- Development and Integration of AI-Powered Chatbots
- Big Data Analytics and Innovation: Linking Enterprise Data Strategy to AI-Driven Outcomes
- Information Governance and Compliance for AI-Automated Data Landscapes
- Enterprise Risk Management Framework for Cloud Computing Security

PROJECTS

Hybrid AI Development & Inference Platform | *On-Premise, Kubernetes, Kafka, OpenAI*

- **Deployed on-premise AI environment serving fine-tuned models** (Gemma 7B, Qwen) with a high-throughput, multi-modal data ingestion pipeline (Kafka, Spark) processing 1TB+ of unstructured data daily from Jira, Git, and Confluence to support continuous model fine-tuning and all internal CI/CD and AI workloads.
- **Designed a dynamic routing layer that intelligently bursts complex or high-priority queries** from local servers to managed APIs (Vertex AI, OpenAI), providing developers seamless access to state-of-the-art models while reducing overall inference costs by 40%.

Autonomous Multi-Agent System for Engineering Operations | *Agentic AI, RLHF, System Design*

- **Engineered a multi-agent "digital twin"** that became the company-wide standard for AI agents, incorporating built-in circuit breakers and human-in-the-loop validation for high-risk actions to ensure zero critical production incidents from automated tasks.
- **Engineered a unified RLHF pipeline leveraging cross-domain feedback** (code reviews, security triage) to fine-tune agents, improving vulnerability detection by 30%, reducing security incidents by 50%, and built a comprehensive back-testing/evaluation framework that validated agent accuracy and delivered a verifiable reduction in escaped defects.

TECHNICAL SKILLS

Cloud Technologies: Google Cloud Platform (GCP – Vertex AI, GKE, BigQuery), Amazon Web Services (AWS – SageMaker, S3, Lambda, Route 53, RDS Aurora, DynamoDB), Microsoft Azure.

Languages: Python, TypeScript, JavaScript, SQL, R, GraphQL, HTML, CSS.

AI/ML: Generative AI (LLMs, LVMs), Recommender Systems, LangGraph, LangChain, RAG, Embeddings, Prompt Engineering, Knowledge Graphs, Transformers, OpenAI, HuggingFace, TensorFlow, Keras, PyTorch, Scikit-learn, Model Optimization (Quantization, Distillation, TensorRT), Distributed Training, MLOps.

Technologies: Kubernetes (GKE), Docker, Terraform, Airflow, Jenkins, Git, Linux, Shell, IAM, Distributed Systems.

Frameworks: React, Next.js, Node.js, Django, Pandas, Spark, PySpark, Android Studio.

Databases: Kubernetes (GKE), Docker, Terraform, Airflow, Jenkins, Git, Linux, Shell, IAM, Distributed Systems.

CERTIFICATIONS

Google: TensorFlow Developer Certificate; Machine Learning Engineer; Data Engineer;

AWS: Solutions Architect, Machine Learning

EDUCATION

University of Cumberlands - *PhD. in Information Technology, Specializing in CyberSecurity and Artificial Intelligence*

The University of Texas at Dallas - *Master of Science in Information Technology and Management*

Pune University - *Bachelor of Engineering: Mechanical Engineering*