Srujay Reddy Vangoor

Location: Sacramento, CA
Phone: +1 (916) 940-5986
Email: srujayreddyv@icloud.com
LinkedIn: linkedin.com/in/srujayreddyv
GitHub: github.com/srujayreddyv

Summary

- Software Engineer with 6+ years of experience building scalable web applications and cloud native systems. Specialized in Python, FastAPI, TypeScript, React, and AWS, with deep expertise in system design and resilient distributed architectures. Known for modernizing legacy platforms, designing high throughput microservices, and building reusable production GenAI systems including RAG pipelines, multi agent orchestration, and evaluation frameworks.

Skills

- **AI & LLM Systems**: AWS Bedrock, RAG Pipelines, LangChain, CrewAI, Hugging Face Transformers, Vector Search (FAISS, Pinecone, OpenSearch), MLflow, PyTorch, TensorFlow, Scikit learn
- **Backend & APIs**: Python (FastAPI, Django), Node.js (Express), Go (Golang), C# (ASP.NET Core), REST, GraphQL, WebSockets
- **Cloud & DevOps**: AWS, Azure, Docker, Kubernetes, Terraform, CI/CD (GitHub Actions, Azure DevOps), CloudWatch, Grafana, Prometheus
- **Data & Analytics**: PostgreSQL, Microsoft SQL Server, DynamoDB, Redis, Snowflake, Amazon Redshift, Airflow, Kafka, Spark, Elasticsearch, Power BI, Tableau
- **Frontend & Mobile**: React, Next.js, TypeScript, JavaScript, React Native

Work Experience

- **AI Software Engineer** Sacramento, CA (Hybrid)
  California Department of Developmental Services (DDS) *Feb 2025 – Present*
  - Built a reusable conversational RAG framework on AWS Bedrock, reducing analysis time by 85%, latency by 30%, and hallucinations by 45% through grounded retrieval, guardrails, and structured evaluation loops.
  - Engineered a CrewAI multi agent orchestration layer with deterministic execution, structured output validation, and persistent memory across four knowledge bases, reducing multi step failures by 25%.
  - Deployed a knowledge graph driven semantic retrieval layer with subgraph expansion, narrative aware embeddings, and relevance pruning, increasing hit@1 accuracy by 78% (0.358 to 0.639) over baseline vector search.
  - Scaled FastAPI microservices handling 20K+ daily requests across distributed ECS and Lambda infrastructure with Cognito and EventBridge, sustaining 99% uptime for healthcare integrations.
  - Developed a React and TypeScript AI workflow intake portal on AWS provisioned via Terraform, reducing submission time from 30 to 10 minutes through automated review pipelines.
  - Modernized COBOL, JCL, and Db2 workflows into event driven Python services aligned with FHIR standards, reducing nightly processing time by 60% and eliminating 200+ manual steps.
- **Software Engineer** Sacramento, CA (Remote)
  California Department of Conservation (DOC) *Jan 2024 – Feb 2025*
  - Engineered a seismic modeling service on AWS integrating a PHP application layer, Python OpenQuake, and Perl services, generating ground motion prediction curves with under 3 second compute time per request.
  - Architected Kafka ingestion pipelines processing 10+ GB daily from earthquake monitoring stations, routing time series to InfluxDB and migrating analytics workloads from PostgreSQL to Druid, reducing retrieval latency by 50%.
  - Implemented two factor authentication and RBAC for 60+ monitoring stations using Azure Entra ID and IAM integrated API Gateway, passing compliance audits with zero findings.
  - Resolved React frontend rendering and performance issues in geospatial applications, optimizing Leaflet and GeoServer integration to improve stability and reduce downtime.
- **Software Application Developer Intern** Sacramento, CA
  Population Research Center (PRC), Sacramento State *Jan 2023 – Jun 2024*
  - Built a staff management system using C#, Entity Framework, and SQL Server, optimizing stored procedures and indexing to reduce administrative workload by 40%.
  - Designed and automated health data ETL pipelines with .NET, SQL Server, and PowerShell, migrating Access and Oracle datasets into centralized SQL analytics stores for Power BI within HIPAA compliant workflows.

- Developed data entry tools and analytics dashboards for statewide public health studies, increasing collection efficiency by 60% through automated validation and data quality checks.
- Supported high volume CATI survey operations by implementing system upgrades and PowerShell diagnostic utilities, reducing survey delays by 30%.

- **Software Engineer** — Hyderabad, IN
  Human Sciences Research Group (HSRG), IIIT-H — *Jan 2021 – Aug 2022*
  - Architected a cloud native archival and analytics platform for social media and news streams, enabling real time analysis of 1M+ data points during the Indian Farmers' Protests.
  - Fine tuned multilingual BERT and IndicBERT models using Hugging Face Transformers for sentiment classification and topic modeling, improving accuracy by 8% over baseline models.
  - Built secure REST APIs with C# and ASP.NET Core to support high volume data ingestion and metadata management, with React based dashboards for exploratory analysis.
  - Containerized services with Docker and implemented CI/CD pipelines using GitHub Actions, reducing release cycles by 40% and improving deployment reproducibility

- **Software Engineer** — Hyderabad, IN
  Change and Continuity in Spiti Valley, ICSSR Sponsored Project — *Jul 2018 – Dec 2020*
  - Engineered a cross platform GIS system using Python, Django, PostgreSQL, and PostGIS to map 280+ heritage sites across 25 villages, enabling spatial search, analytics, and interactive visualization.
  - Built responsive React web and offline first React Native mobile clients with Leaflet, achieving 60% faster load times in low bandwidth field environments.
  - Designed and deployed REST APIs with Django REST Framework supporting spatial and attribute queries with pagination and caching, reducing query latency and improving scalability.
  - Developed Python based ETL and spatial processing workflows integrating survey, census, and archival datasets into a unified geospatial repository, improving metadata accuracy by 40% and enabling near real time updates.

Education

- **California State University, Sacramento** — Sacramento, California
  Master of Science (MS) in Computer Science — *Aug 2022 – Jan 2025*
- **International Institute of Information Technology (IIIT-H)** — Hyderabad, India
  Bachelor of Technology (BTech) in Computer Science — *Aug 2015 – July 2019*

Relevant Projects

- **Buddhira – Personal Knowledge Management App**
  - Architected a full stack second brain application using Next.js with TypeScript, FastAPI, and Supabase, enabling secure note capture, search, tagging, and archive workflows with strict per user data isolation.
  - Implemented JWKS based JWT authentication, consistent API error handling, rate limiting, automated GitHub Actions CI/CD pipelines, and workflow level test coverage to ensure production reliability.
- **CA DMV RAG System – Retrieval Augmented Question Answering Platform**
  - Built a RAG system using FastAPI, FAISS, and Sentence Transformers over the California DMV handbook, supporting streaming responses with citations, reranking, multi document filtering, and confidence gating.
  - Deployed with Docker on Render, adding structured observability with metrics and request IDs, CI/CD via GitHub Actions, and comprehensive PyTest coverage across ingestion and retrieval workflows.
- **FastChat – Real Time Chat Application**
  - Developed a Dockerized full stack system using FastAPI, React with TypeScript, and PostgreSQL, supporting WebSocket based real time messaging, presence tracking, and typing indicators.
  - Integrated JWT authentication, rate limiting, structured logging, CI/CD driven integration testing, and performance monitoring to maintain reliability under concurrent traffic.
- **Sacverse - Augmented Reality (AR) Campus Tour App**
  - Designed an augmented reality campus navigation prototype with geospatial search and route planning optimized for low bandwidth mobile environments.
  - Conducted structured usability testing with 50+ pilot users, improving spatial overlay accuracy and interaction flows, resulting in a 40% improvement in measured usability.