Mohana Satyanarayana Moganti

San Jose, CA • mohanmoganti2023@gmail.com • +1 (669)-329-9412
https://linkedin.com/in/mohana-moganti • https://github.com/Dead-Stone

Skills

- **Languages:** Python, TypeScript, Golang, Scala, C++, Java, JavaScript, SQL
- **Web Development:** FastAPI, ReactJS, Svelte, Spring Boot, Django, Flask, GraphQL
- **LLM/GenAI:** RAG, LangChain, CrewAI, Prompt Engineering, Embeddings, A/B Testing, Fine-tuning, OpenAI, Gemini, Vertex AI
- **DevOps/MLOps:** GitLab CI, GitHub Actions, Harness, ArgoCD, Docker, Kubernetes, Prometheus, Grafana, MLflow
- **Cloud/Data:** AWS (Lambda, S3, EC2), Azure, Databricks, Snowflake, Spark, Flink
- **Databases:** Weaviate, Pinecone, Neo4j, MongoDB, PostgreSQL, MySQL
- **QA/Tools:** Selenium, Postman, JUnit, TestNG, Jira, Vite, Tailwind

Experience

Founding Engineer *Gembizz LLC — San Jose, CA* Aug 2025 – Present

- Led system architecture, full-stack development, and infrastructure setup for a community-focused platform, owning backend services, frontend workflows, and production deployment.
- Built scalable onboarding, content publishing, and engagement flows using **TypeScript**, React (Vite), **Tailwind CSS**, FastAPI, and MongoDB Atlas, supporting early growth and evolving product requirements.
- Deployed containerized backend services on AWS with Docker, configuring environments, service communication, structured logging, and basic observability.
- Established CI/CD pipelines using GitHub Actions, enabling frequent, low-risk deployments and reducing manual release overhead by over 50%.
- Delivered real-time messaging and update workflows using WebSockets and event-driven patterns, improving platform engagement and repeat usage by approximately 25%.
- Partnered directly with the founder to translate early product ideas into technical architecture, delivery milestones, and prioritized execution plans.

AI Engineer Sep 2024 – Nov 2024
Astranetix Corporation — San Jose, CA

- Built multimodal Retrieval-Augmented Generation (RAG) systems using Django, Weaviate, OpenAI APIs, and GraphQL-based service integration.
- Optimized vector search using HNSW indexing and improved embedding pipelines, reducing retrieval latency and increasing response relevance.
- Designed and executed A/B testing experiments to compare prompts, embedding strategies, and retrieval configurations, enabling data-driven model quality improvements.
- Deployed AI services on AWS Lambda and S3, lowering infrastructure and operational costs by approximately 40% while improving scalability.

AI Intern May 2024 – Aug 2024
Flatirons AI — San Jose, CA

- Designed and executed large-scale ETL pipelines using AWS Glue and PySpark to support generative AI, personalization, and recommendation workloads.
- Built and maintained knowledge graphs using Neo4j and Microsoft GraphRAG, improving contextual retrieval quality and inference accuracy across AI products.

- Defined observability and performance standards for ML services by integrating Prometheus metrics and Grafana dashboards, reducing operational issues and improving model reliability across production deployments.
- Developed requirement docs, sprint plans, and acceptance criteria by closely collaborating with engineering and data science teams to ensure timely and feasible delivery.
- Led evaluation of LLM-driven features, incorporating user feedback, telemetry data, and experiment metrics to guide roadmap decisions.

Teaching Assistant Aug 2024 – May 2025
San Jose State University — San Jose, CA

- Supported graduate-level courses in Machine Learning, Networking, and Information Security, assisting 100+ students across multiple semesters.
- Led labs and office hours covering OWASP Top 10, TCP/IP routing, and deep learning workflows using PyTorch and TensorFlow.
- Evaluated assignments and projects with detailed technical feedback, maintaining consistency and academic integrity.

Associate Software Analyst Aug 2021 – Jun 2024
Deloitte — Hyderabad, India

- Developed and maintained enterprise-scale fintech applications using React, TypeScript, and GraphQL in regulated production environments.
- Designed GraphQL APIs to enable modular data access and near real-time updates across distributed systems.
- Delivered CI/CD pipelines using Harness and GitLab CI, increasing deployment frequency and reducing rollback incidents.
- Authored unit and integration tests using JUnit, TestNG, and Selenium, reducing post-release defects by approximately 25%.
- Collaborated with global cross-functional Agile teams to deliver client-facing features under strict compliance requirements.
- Received Deloitte Spot Award for high-impact delivery on critical client projects.

Trainee Software Engineer Jan 2020 – Jul 2021
Turito / YuppTV — Hyderabad, India

- Developed scalable RESTful APIs in Scala for an EdTech platform serving over 100K monthly users.
- Built and maintained Angular front-end modules with reusable UI components to improve user experience.
- Integrated and optimized AWS DynamoDB operations, reducing API response times by up to 30%.
- Designed relational database schemas and implemented stored procedures supporting user management and enrollment workflows.
- Implemented secure authentication and authorization to enable reliable role-based access across the application.

Projects

AI Sketch Generator (Personal Project) Jan 2024 – Apr 2024

- Built an AI-powered sketch generation web app using React, Python FastAPI, and Stable Diffusion to convert text prompts into hand-drawn style sketches. Implemented prompt-to-image pipelines with control models to ensure accurate shape and style consistency.
- Designed a full-stack workflow with user authentication, job queues, and cloud-based image processing using AWS Lambda and S3. Optimized inference by caching embeddings and batching requests, reducing generation time by 45%.

Agentic Grading System (SJSU) Dec 2024 – May 2025

- Built an automated agentic grading platform using CrewAI, Weaviate, and Gemini to evaluate text and image submissions. The system uses multimodal RAG, rubric-based reasoning, and multi-agent workflows to produce consistent grading decisions.
- Designed vector search pipelines and fine-grained retrieval logic to map student answers to rubric criteria. This reduced manual grading workload by 60% and improved scoring consistency across evaluators.

Movie Ticket Booking System (SJSU) Aug 2023 – Nov 2023

- Developed a full-stack movie ticket booking platform using React, Node.js, and GraphQL for real-time seat selection and dynamic showtime updates. Implemented optimized resolvers and schema design to eliminate over-fetching and improve query performance.
- Built microservice-based backend services with MongoDB, containerized using Docker, and deployed via a CI/CD pipeline inspired by Harness automation. Added caching and rate-limiting to ensure responsiveness under high concurrent user load.

Remote Joystick (Osmania University) Nov 2019 – Jan 2020

- Built a real-time remote joystick system using React and Python with optimized TCP socket communication. The architecture achieved sub-10ms input latency, enabling responsive gameplay even on low-end client devices.
- Implemented client-side prediction and real-time state synchronization to avoid lag spikes during gameplay. This improved usability and reduced input jitter under fluctuating network conditions.

Publications

"Learning Based Approach for Hindi Text Sentiment Analysis using Naïve Bayes Classifier", International Journal of Innovations in Engineering Research and Technology (IJIERT), Vol. 7 No. 08, pp. 40–47, 2020.
https://repo.ijiert.org/.../view/161

Education

Master of Science in Software Engineering
San Jose State University, CA *May 2025*
GPA: 3.6 / 4.0