Sree Vardhan Reddy K

AI Engineer

Hyderabad | ksvr122002@gmail.com | 9133469003 | www.linkedin.com/in/sreevardhanreddy-k
https://github.com/Sree-Vardhan-Reddy-K

Summary

Entry level AI Engineer with foundations in Python, SQL, Machine Learning, and NLP. Built end-to-end AI systems, including retrieval-augmented pipelines, with emphasis on correctness, evaluation, and reliability. Hands-on experience developing and deploying FastAPI-based APIs on AWS EC2 for applied AI use cases.

Areas of Expertise

Programming: Python, SQL

Data Science: Statistics, Machine Learning, Deep Learning, Natural Language Processing

GenAI and LLM: LangChain, RAG Concepts, Vector Stores (FAISS / Chroma)

API and Deployment Exposure: FastAPI, Docker, AWS EC2

Education

Chaitanya Bharathi Institute of Technology (Graduation) BE in Computer Science. 2020 - 2024

- GPA: 9.35/10.0

Certifications

- Data Science Certification, ExcelR Solutions
- Complete Data Science, Machine Learning, Deep Learning and NLP Bootcamp Certification - Udemy

Projects

Verbalens – AI Document Intelligence System | GitHub: github.com/Sree-Vardhan-Reddy-K/Verbalens_project

- Built a high-precision document intelligence pipeline with hybrid retrieval, combining dense vector search (FAISS) and Cross-Encoder re-ranking to maximize recall while selecting only the most semantically relevant document chunks.
- Implemented LLM-based semantic guardrails to enforce document-centric reasoning, blocking out-of-scope queries and external entity leakage through prompt-level YES/NO gating and hard constraints.
- Applied automated quality filtering and post-generation deduplication to remove noisy, fragmented, or redundant context, ensuring clean inputs to the LLM and concise, non-repetitive final responses.
- Evaluated system quality using RAGAS metrics, achieving 0.75 context recall, 0.75 context precision, and 0.96 answer relevancy to validate answer grounding and reduce hallucinations.
- Deployed a containerized FastAPI application on AWS EC2 using Docker, with separate ingestion and inference pipelines for reliable AI service operation.

Enterprise Analytics Copilot with Business Logic Enforcement

GitHub: github.com/Sree-Vardhan-Reddy-K/enterprise-analytics-copilot

- Designed a correctness-first analytics system with metric-owned semantics (versioned KPIs, enforced grain, authoritative time columns, allowed/forbidden filters), achieving deterministic rejection of ambiguous or invalid queries by design.
- Implemented fail-fast intent and semantic validation using closed enums and schema enforcement, eliminating LLM hallucinations and best-effort guessing in KPI computation.
- Built parser-based SQL safety using SQLGlot, enforcing SELECT-only execution, mandatory time filters, and LIMITs, resulting in no unsafe query execution paths across tested scenarios.
- Developed deterministic intent-based caching, achieving O(1) latency and cross-version data correctness.
- Enabled versioned business logic with automated semantic checks, allowing safe KPI evolution while preserving historical financial correctness in a FastAPI + MySQL production-grade service.