SAILAJA MORRENNAGARI

AI/ML Engineer | Generative AI, Retrieval Systems & MLOps

morrennagarisailaja@gmail.com | sailaja-morrennagari.github.io | github.com/Sailaja-ASB | linkedin.com/in/sailajamorrennagari | leetcode.com/u/Its_Sailaja

PROFESSIONAL SUMMARY

AI/ML engineer with 6+ years of experience spanning distributed data engineering, machine learning, and production AI systems. Builds and deploys training and inference workflows, retrieval-augmented generation applications, and scalable ML infrastructure using Python, PyTorch, Spark, Kubernetes, and AWS. Delivered measurable improvements in model accuracy, search speed, deployment cycles, batch processing, and data quality across enterprise platforms. Open-source contributor to the Hugging Face ecosystem.

TECHNICAL SKILLS

Programming	Python, SQL; data structures, algorithms, complexity analysis, performance optimization
Machine Learning	PyTorch, TensorFlow, Scikit-learn, XGBoost; training, inference, evaluation, optimization
Generative AI	Large language models, RAG, LangChain, FAISS, ChromaDB, embeddings, agentic workflows, prompt engineering, LLM evaluation, benchmarking
Data Systems	PySpark, Apache Spark, Kafka, Hadoop, Delta Lake, Airflow, ETL/ELT, Snowflake
ML Infrastructure	AWS SageMaker, S3, Glue, Redshift, EMR; MLflow, Docker, Kubernetes, GitHub Actions CI/CD, FastAPI model serving, Ollama
Observability	Prometheus, Grafana, Great Expectations, PyTest

PROFESSIONAL EXPERIENCE

AI/ML Engineer (Contract) | UnitedHealth Group, USA **Jun 2025 – Present**

- Designed and deployed production ML training and inference workflows for clinical risk scoring using AWS SageMaker, PyTorch, and TensorFlow, improving model accuracy from 70% to 82.6%.
- Developed retrieval-augmented generation applications with LangChain and FAISS, reducing clinical document search time by 65% across 10K+ unstructured documents.
- Built scalable batch and streaming pipelines with Kafka, PySpark, and Apache Airflow to supply reliable, high-volume data to ML training and inference workloads.
- Standardized model packaging and deployment with MLflow, Docker, Kubernetes, and GitHub Actions, reducing deployment cycle time by 40% through reproducible CI/CD automation.
- Instrumented production ML services and pipelines with Prometheus and Grafana for monitoring, alerting, and incident response, strengthening reliability and root-cause analysis.
- Partnered with clinicians, data scientists, and engineers to translate requirements into production-grade AI/ML solutions.

Environment: Python, AWS SageMaker, PyTorch, TensorFlow, Kafka, PySpark, MLflow, Docker, Kubernetes, Prometheus

Data Scientist | Tata Consultancy Services, India **Mar 2022 – Dec 2023**

- Built and evaluated classification and predictive models in Python, applying feature engineering, model selection, validation, and performance analysis across repeatable experimentation workflows.
- Performed exploratory data analysis and feature engineering across multiple datasets, reducing data preparation time by 30%.
- Replaced manual SQL reporting with recurring Snowflake-based queries, reducing turnaround time by 40% across 10+ reports.
- Developed scalable ML-ready feature and dataset pipelines using AWS Glue, S3, and PySpark to support repeatable training, validation, and analytical workloads.

Environment: Python, Scikit-learn, SQL, Snowflake, AWS Glue, S3, PySpark

Data Engineer | Hexaware Technologies, India **Jun 2018 – Feb 2022**

- Developed distributed data-processing pipelines in PySpark and SQL for high-volume structured and semi-structured workloads.
- Resolved Spark performance bottlenecks through partitioning and query optimization, reducing batch-processing time by 35%.
- Engineered reusable PySpark and Spark SQL components to standardize transformation logic and improve pipeline maintainability.
- Implemented data quality checks and validation rules, reducing data-related defects by 30%.

Environment: PySpark, SQL, Apache Spark, Hadoop (HDFS, Hive), Airflow, Git, Linux

SELECTED PROJECTS

Local Agentic Copilot – Agentic RAG & Retrieval Benchmarking **github.com/Sailaja-ASB/local-agentic-copilot**

Python | Ollama (Qwen3) | ChromaDB | Sentence-Transformers | BM25 | Cross-Encoder

- Built a local-first agentic RAG copilot with Researcher, Coder, and Reviewer agents, deterministic guardrails, and a bounded self-correction loop, running fully on-device via Ollama.
- Engineered hybrid retrieval (dense semantic + BM25 fusion) with cross-encoder reranking and page-aware PDF citations.
- Built a reproducible 47-question retrieval benchmark comparing dense, BM25, hybrid and reranked search; hybrid retrieval (RRF) raised Top-1 accuracy from 91.5% to 100% over either method alone, while a cross-encoder reranker added ~50 ms/query with no further gain, so it was left out.
- Documented the build in a public technical write-up, "Building a Local Agentic RAG Copilot: Retrieval, Citations, and Self-Correction" (LinkedIn & Medium, Oct 2026).

LLMOps Platform – Production LLM Serving (in progress) **github.com/Sailaja-ASB/llmops-platform**

Python | FastAPI | Kubernetes (kind) | Helm | Docker | Prometheus | Locust | BM25 + Embeddings | Ollama (Qwen2.5-3B)

- Built a production-style LLM platform on Kubernetes with Helm: a FastAPI gateway with API keys, rate limiting, circuit breaker, SSE streaming, load shedding and Prometheus metrics, plus a separate hybrid-retrieval (BM25 + dense, RRF) RAG service.
- Load-tested with Locust (1–20 users); found that liveness probes restarted healthy pods under load, and fixing probes and graceful shutdown cut stress-test errors from 30% to 0%; showed gateway autoscaling cannot fix a single-model-server bottleneck.
- Added RAG with gateway-verified citations and layered prompt-injection guardrails; a prompt change raised citation accuracy from 40/47 to 47/47, and a red-team eval showed 0 of 12 attacks succeeding with no false positives.

Generative AI – RAG Chatbot

LangChain | FAISS | Vector Embeddings

- Designed and implemented a retrieval-augmented generation pipeline using LangChain, vector embeddings, and FAISS, with ingestion, retrieval, and LLM orchestration workflows tuned via prompt engineering and embedding optimization.

Deep Reinforcement Learning for Stock Trading **github.com/Sailaja-ASB/Deep-RL-Stock-Trading**

Python | TensorFlow | PPO | A2C | DQN

- Built a custom trading environment and trained PPO, A2C, and Double DQN agents on real stock-market data, mitigating overfitting through architectural tuning and reward shaping.

Open-Source Contributions — Hugging Face Ecosystem

Open pull requests: huggingface/evaluate #830 · huggingface/datasets #8794 · huggingface/datasets #8805

- Implemented the documented zero_division parameter for the r_squared metric in huggingface/evaluate (PR #830, fixes #803).
- Fixed the JSON loader incorrectly parsing plain strings in inferred JSON columns (PR #8794, fixes #8713) and a SplitDict.copy() crash on SplitInfo objects (PR #8805, fixes #8617) in huggingface/datasets.

EDUCATION

Master of Science in Computer Science (Machine Learning) **May 2025**

University of South Dakota, USA

Master of Computer Applications (MCA) **Jun 2021**

Dr. Ambedkar Institute of Technology, India

Bachelor of Science – Mathematics, Physics & Computer Science **May 2018**

Sri Venkateswara University, India

CERTIFICATIONS

- AWS Certified Solutions Architect – Associate
- IBM Generative AI (LangChain RAG) – Capstone
- Machine Learning & Deep Learning – REGEX Software