Sneha Reddy G.

AI/ML Engineer | *4+ Years*

Dallas, TX | snehar.work@gmail.com | +1 726-219-6859 | LinkedIn

Professional Summary

AI/ML Engineer with 4+ years of experience shaping practical AI solutions that enhance search quality, streamline model performance, and stabilize data operations. I've contributed to systems that lifted retrieval accuracy by 12–18%, lowered P95 inference latency by 30–40%, and supported daily processing of over 1TB of data through Spark-based pipelines. My work spans RAG development, LLM optimization, feature engineering, and MLOps workflows that cut recurring issues by 50%+. With hands-on expertise across Python, PyTorch, FastAPI, and Kubernetes, I build dependable, scalable AI components that improve reliability and create measurable business value.

Technical Skills

- ***Programming:*** Python, SQL, Bash, Java, C, React.js, HTML, CSS
- ***Machine Learning*:** Feature Engineering, Model Explainability (SHAP, LIME), Recommendation Systems, Time Series Forecasting, Cross-Validation, Calibration, Hyperparameter Tuning, Model evaluation (AUC, F1, precision/recall, Brier score)
- ***Deep Learning:*** PyTorch, TensorFlow, Transformers, Fine-tuning, CNNs, RNNs, LSTMs, ONNX, TensorRT
- ***NLP & GenAI:*** Retrieval-Augmented Generation (RAG), Embeddings, Prompt Engineering, Sentence Transformers, spaCy, Hugging Face, NER, Text Classification, Text Summarization, Vector Search (FAISS, Pinecone), LangChain, LlamaIndex, Re-rankers
- ***MLOps & Deployment:*** MLflow, Model Registry, Feast, Kubeflow, Argo Workflows, CI/CD for ML, Triton Inference Server, vLLM, Experiment Tracking, Monitoring (Evidently, WhyLabs, Prometheus, Grafana)
- ***Data Engineering & Warehousing:*** Apache Spark, Airflow, dbt, Prefect, Delta Lake, Apache Iceberg, Kafka, Kinesis, Parquet, Data Modeling, Databricks, Snowflake, BigQuery, ETL/ELT Pipelines, Data Lakes, Data Warehouses
- ***Analytics & Visualization:*** Tableau, Power BI, Matplotlib, Pandas, NumPy, scikit-learn, XGBoost, LightGBM, CatBoost
- ***APIs & Serving:*** FastAPI, gRPC, Async I/O, Caching & Batching, A/B Testing Hooks
- ***Cloud & Infrastructure:*** AWS (SageMaker, S3, Lambda, ECR, EKS), Docker, Kubernetes, Helm, Terraform, IAM, VPC Security, Encryption
- ***Testing & DevOps:*** pytest, Great Expectations, Data Contracts, GitHub Actions, GitLab CI, Secrets & Artifact Management

Professional Experience

AI/ML Engineer | *Deloitte* ***Aug 2024 – Present | USA***

- Built RAG service using FAISS + cross-encoder rerankers, improving top-3 recall by 15%. Reduced support agents' average handle time by 28% across seven teams.
- Containerized PyTorch models with Triton, ONNX, and TensorRT, cutting P95 latency from 220ms to 135ms, while autoscaling on Kubernetes reduced compute cost by 23% monthly.
- Implemented MLflow + Feast for feature versioning (500GB) and automated 60% of hyperparameter sweeps. Reduced model time-to-deploy from 14 days to 5.
- Designed Airflow + Spark ETL pipelines processing 1.2TB/day with Great Expectations data contracts. Reduced schema breakages by 60% and improved SLA adherence to 70%.
- Built Prometheus + Evidently monitoring, reducing incidents 70% and cutting MTTR by 36%.
- Partnered with security to implement privacy checks for NLP models, achieved 60% PII redaction accuracy and cleared internal audit on first pass with no findings.
- Built synthetic-data generation pipelines to augment training datasets, improving downstream model accuracy by 7–10%.
- Integrated guardrails, safety filters, and enterprise compliance workflows for LLM-powered applications, reducing unsafe responses by 40%.

Jr. AI/ML Engineer | *IBM* ***Jan 2020 – Jul 2023 | India***

- Designed Spark and Airflow recommendation pipeline covering preparation, candidate generation, ranking, and cold-start via embeddings, streamlined CTR by 11% and pruned P95 latency to 150ms.
- Redesigned XGBoost fraud detection with calibrated probabilities and investigator queues, lifted AUC from 0.84 to 0.91, tightened false positives 32%, accelerated dispute resolution by 29%.
- Tuned search relevance with a lightweight ranker and holdout metrics, increased NDCG@10 by 9%, maintained sub-120ms P95 through feature selection and compact architectures in production.
- Established Great Expectations data quality program detecting schema drift, leakage, and anomalies, reducing incident rate from seven to two per month and lowering MTTR by 41%.
- Built React dashboards and Java Spring Boot APIs for governance, implemented RBAC and audit trails serving 180+ users, optimized PostgreSQL queries, improving report latency 46%.
- Delivered Java microservices with Kafka and Redis caching for recommendation telemetry, processed 1.8M events/day, strengthened idempotency, tuned SQsL indexes to reduce query CPU by 35%.
- Developed internal Python libraries for distributed feature engineering and reusable ML utilities adopted by 5+ engineering teams.
- Implemented profiling, optimization, and caching improvements that reduced API inference costs by 20% while maintaining latency SLAs.

Projects

Claims Anomaly Scoring with Gradient Boosting (*Tech: Python, Spark, XGBoost, SHAP, PostgreSQL, Airflow*)

- Spark and XGBoost pipeline on 2.7M claims with 180 engineered features, achieving cross-validated AUC of 0.89 and reducing analyst review volume by 22% on a held-out set.
- Isotonic calibration improved Brier score by 17%, and SHAP analysis flagged three risky codes, leading to revised sampling rules and increasing confirmed fraud discovery rate from 3.1% to 3.8%.

Campus Policy Q&A with Retrieval-Augmented Generation (*Tech: Python, FAISS, vLLM, Hugging Face, FastAPI, Kubernetes*)

- Indexed 28,000 handbook pages using FAISS, cross-encoder reranker achieving 84% top-3 recall and reducing unresolved pilot queries from 105 to 49 per week.
- vLLM served 7B model with quantization, median latency 210ms at 90 RPS batching, hallucinations decreased 31% using retrieval citations, safety moderation, and tighter prompt constraints. Deployed using FastAPI + vLLM on Kubernetes.

Education

M.S., Computer Science ***May 2025 | USA***

University of North Texas

B. Tech, Electronics & Communication Engineering ***Sep 2020 | INDIA***

Gokaraju Rangaraju Institute of Engineering and Technology

Certifications

- **AWS Certified Machine Learning – Specialty**
- **Google Cloud Professional Machine Learning Engineer**
- **Certified Kubernetes Administrator (CKA)**
- **NVIDIA DLI — Optimizing Inference with TensorRT**