KESANA SAI NARENDRA KUMAR

AI/ML Engineer| Dallas, Texas | kesana.class2024@gmail.com | (469) (494) – 4742 | LinkedIn

SUMMARY

AI/ML Engineer with 4+ years of building production orchestration systems, agentic AI pipelines, and LLM optimization frameworks on AWS. Expert Python engineer specializing in Apache Airflow orchestration, autonomous retrieval agents (LangChain), and AWS-native data engineering (ECS, EventBridge, Lambda, S3, Glue) serving 2M+ daily users. Proven track record designing context-aware systems processing Git/PR metadata, implementing prompt engineering (chaining, context packing, DSPy concepts), and developing backend services with comprehensive evaluation frameworks. Strong background in robust data pipelines with automated backfills, intelligent retries, data quality monitoring, and cross-functional collaboration delivering 15+ production systems with 99.5% uptime.

PROFESSIONAL EXPERIENCE

CAPITAL ONE|AI ENGINEER | DALLAS | TX — SEP 2023-Current

- Designed and deployed production-grade **Apache Airflow orchestration** system managing 15+ ML pipelines on AWS infrastructure, processing 2M+ daily user requests with automated scheduling, retries, and dependency management.
- Built event-driven pipelines using AWS-native services like EventBridge for triggering, ECS Fargate for containerized tasks, Lambda for lightweight functions, Glue for Spark jobs, S3 for data lake-all orchestrated via Airflow.
- Implemented robust **orchestration workflows** with comprehensive error handling, automated retries with exponential backoff, data quality checks at each pipeline stage, and alerting systems achieving 99.5% uptime
- Developed high-performance **Python services** using advanced patterns (async/await, multiprocessing, generators) to process 50K+ daily requests with sub-200ms latency.
- Leveraged modern ML tooling including **Hugging Face Transformers**, LangChain, MLflow, Weights & Biases, and DVC for version control, ensuring reproducible and scalable model development.
- Developed autonomous **retrieval agent** using LangChain that queries Git metadata, searches vector database of code embeddings, and synthesizes historical context for PR analysis—processing 50K+ queries monthly.
- Implemented systematic **prompt engineering** workflows including prompt chaining (multi-step reasoning), context packing (maximizing token efficiency), few-shot learning, and chain-of-thought prompting, improving LLM response quality by 40%.
- Applied **prompt optimization** principles aligned with **DSPy** methodology including automatic prompt tuning, signature-based prompting, and modular prompt composition for complex multi-step LLM workflows
- Optimized **context packing strategies** like dynamic truncation based on relevance scores, sliding window for long documents, hierarchical summarization for reducing token count, and smart few-shot example selection
- Created automated **workflow tracking PR** merges to production branches, computing time-to-merge metrics, identifying bottlenecks, and alerting on unusual patterns (e.g., direct commits bypassing PR process
- Implemented comprehensive **MLOps** toolkit including Docker/Kubernetes for containerization, Terraform for IaC, GitHub Actions for CI/CD, and Prometheus/Grafana for monitoring.
- **Evaluated open-source LLMs** (LLaMA 2, Mistral) for cost-sensitive tasks, benchmarking accuracy vs. GPT-3.5, measuring inference latency, and identifying use cases suitable for OSS models (30% cost reduction)
- Established Python best practices including type hints, comprehensive testing (pytest), code reviews, and PEP 8 compliance across team of 3 junior engineers.

HSBC | DATA SCIENTIST | India — JUL 2020-JUL 2022

- Built 20+ predictive models (regression, classification, clustering) using **Scikit-learn and XGBoost**, applying strong mathematical foundations to forecast customer behavior with 85% accuracy and 15% retention improvement
- Built data pipelines using modern stack like Apache Airflow for orchestration, PySpark for distributed processing, Delta Lake for data versioning, and Kafka for real-time streaming
- Built data ingestion pipelines processing Git metadata, PR payloads, code diffs, commit histories, and branch relationships from GitHub API, tracking 10,000+ pull requests monthly for lineage and impact analysis
- Performed rigorous **A/B testing and statistical analysis** on product features, validating hypotheses through experimental design and providing actionable recommendations that improved conversion rates by 18%
- Implemented **user feedback interface** with thumbs up/down on predictions, optional comments for false positives/negatives, and confidence ratingscollecting 1000+ feedback samples monthly.
- Implemented Open Telemetry for distributed tracing, propagating trace context through Airflow→ API → database queries, enabling end-to-end latency analysis and bottleneck identification
- Implemented statistical anomaly detection across data pipelines: Interquartile Range (IQR) for outliers, Z-score for statistical deviations, time-series forecasting for expected values, and schema drift monitoring.
- Created automated training data pipeline extracting features from feedback-annotated PRs, generating balanced datasets, performing **feature engineering**, and versioning training data using DVC for reproducibility

SKILLS

Languages: Python, R, SQL
ML & Frameworks: TensorFlow, PyTorch, Keras, Scikit-learn, XGBoost, Hugging Face Transformers, NumPy, Pandas, Matplotlib, Seaborn, Model Evaluation, Hyperparameter Tuning, Topic Modeling, A/B testing, Time-series Analysis.
Deep Learning & Neural Networks: CNNs, RNNs, LSTMs, GANs, Transformers, BERT, GPT, Attention Mechanisms, Transfer Learning, Fine-tuning
NLP & computer Vision: NLTK, SpaCy, Gensim, Named Entity Recognition (NER), Sentiment Analysis, Text Classification, Question Answering, Language Models (LLMs), Image Segmentation, Face Recognition, OCR
Agentic AI & LLM Frameworks: LangChain, LangGraph, Multi-Agent Systems, Autonomous AI Agents, Agent Orchestration.
GenAI & LLM Technologies: OpenAI API, LlamaIndex, Prompt Engineering (Prompt Chaining, Context Packing, DSPy Concepts) RAG, Vector Search, Fine-tuning LLMs (LoRA & QLoRA)
Data Modeling & Engineering: PR/Commit Metadata Modeling, Branch/Merge Tracking, Code Artifact Linking, Graph-Based Relationships, PySpark, Kafka, Delta Lake
Orchestration & Cloud: Apache Airflow, Dagster, MLflow, CI/CD, Automated Retries, Scheduling, Backfills, Workflow Monitoring, AWS (SageMaker, Lambda, ECS, EventBridge, S3, Glue, EC2), Infrastructure as Code (Terraform), Docker, Kubernetes
Data Engineering & ETL: PySpark, Spark, Hadoop Ecosystem, Databricks (Delta Lake), Kafka, REST API Pipelines.
Databases: PostgreSQL, MySQL, MongoDB, Redis, Elasticsearch, Pinecone, ChromaDB, Vector Databases.
Soft Skills: Problem Solving, Analytical Thinking, Cross-Functional Collaboration, Mentorship, Agile/Scrum Practices.

EDUCATION & CERTIFICATION

- **Master of Science in Business Analytics & AI** *University of Texas at Dallas, Richardson, TX* **May2024**
- **Bachelor of Technology, Ceramic Engineering** *Indian Institute of Technology, Varanasi, India* **May2021**
- AWS Certified Machine Learning – Associate - 2025
- Python for Data Science, AI (IBM and Coursera) – 2024