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 specialized experience in designing, deploying, and optimizing production-grade artificial intelligence and machine learning solutions that drive measurable business impact across financial services and enterprise environments. Expert in building scalable agentic AI systems using LangGraph and LangChain, with proven success orchestrating multi-agent architectures that automate complex workflows and reduce manual processing time by up to 65%.

PROFESSIONAL EXPERIENCE

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

- Architected and deployed **multi-agent AI systems** using LangGraph and LangChain, orchestrating 5+ specialized agents that automated complex workflows, reducing manual processing time by 65% and serving 50,000+ monthly queries with 94% accuracy
- Engineered production-grade **RAG (Retrieval-Augmented Generation)** pipelines integrating Pinecone vector database, OpenAI embeddings, and hybrid search strategies, achieving 40% improvement in response relevance and reducing hallucination rates from 23% to 8% across enterprise knowledge base.
- Designed and productionized **agentic AI workflows on AWS** (SageMaker, Lambda, Bedrock) using LangGraph's supervisor and hierarchical architectures, enabling autonomous task decomposition and execution that decreased operational costs by $180K annually
- Deployed 15+ scalable **ML models on AWS infrastructure** (EC2, S3, SageMaker) with Docker/Kubernetes orchestration, implementing CI/CD pipelines via GitHub Actions and MLflow that reduced deployment cycles to 3 days and improved model versioning efficiency by 70%
- Established comprehensive **model monitoring and alerting systems** using AWS CloudWatch and custom Python frameworks to track model drift, data quality, and performance degradation, maintaining 99.5% uptime for critical AI services and catching 12 production issues before user impact
- Developed production **deep learning models** using PyTorch and TensorFlow for NLP and computer vision applications, optimizing inference latency from 500ms to 180ms through quantization, pruning, and TensorRT acceleration, serving 2M+ daily users with 92% accuracy
- **Optimized LLM performance** and cost efficiency through prompt engineering, fine-tuning strategies, and contextual compression techniques, reducing OpenAI API costs by 35% ($45K annually) while maintaining response quality and implementing automated evaluation frameworks (RAGAS, LangSmith)
- Collaborated with product, engineering, and data teams to translate business requirements into AI solutions, delivering 8 production models in 18 months including fraud detection (89% precision), customer segmentation (32% ROI lift), and intelligent document processing systems; mentored 3 junior engineers on agentic AI patterns and AWS best practices

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

- Built 20+ production **predictive models** (XGBoost, Random Forest, Neural Networks) for customer churn forecasting and risk assessment, achieving 85% prediction accuracy that enabled proactive retention strategies, reducing churn by 15% and saving $2.3M in annual revenue
- Designed and executed **30+ A/B tests and statistical experiments** using Python (SciPy, statsmodels) to validate product hypotheses and marketing strategies, delivering actionable insights that improved conversion rates by 18% and drove $4.1M in incremental revenue across digital channels
- Engineered **customer segmentation models** using advanced clustering algorithms (K-means, DBSCAN, hierarchical clustering) and RFM analysis, enabling personalized marketing campaigns across 8 customer segments that increased marketing ROI by 32% and reduced acquisition costs by 24%.
- Developed interactive business intelligence dashboards using Tableau and Python (Matplotlib, Plotly) to visualize KPIs, customer behavior patterns, and model performance metrics, empowering 50+ stakeholders to make data-driven decisions that contributed to 22% YoY revenue growth

SKILLS

Languages: Python (Advanced), 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, 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

GenAI & LLM Technologies: OpenAI API, LangChain, LangGraph, LlamaIndex, Prompt Engineering, RAG, Vector Search, Fine-tuning LLMs, Multi-Agent Systems
MLOps & Cloud: AWS (SageMaker, Lambda, S3, EC2, Bedrock, CloudWatch), Azure (OpenAI, Azure ML, Databricks), Vertex AI, Docker, Kubernetes, MLflow, CI/CD, Git, GitHub Actions, Terraform
Data Engineering & ETL: PySpark, Spark, Hadoop Ecosystem, Databricks (Delta Lake), Apache Airflow, Kafka, REST API Pipelines
Databases: PostgreSQL, MySQL, MongoDB, Redis, Elasticsearch, Pinecone, ChromaDB, Weaviate, 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