SUTIKSHAN B
AI/ML Engineer & Gen AI Engineer

- Generative AI / ML / NLP Engineer with experience supporting the design, fine-tuning, and deployment of AI solutions using Large Language Models (LLMs) for conversational and intelligent systems.
- Hands-on experience working with LLMs and Transformers using Azure OpenAI, Hugging Face Transformers, LangChain, LangGraph, and Azure AI Search for natural language understanding, semantic retrieval, and AI-assisted user interactions.
- Experience assisting in LLM fine-tuning, evaluation, summarization, Q&A, semantic extraction, and prompt optimization for domain-specific applications.
- Strong exposure to prompt engineering strategies to improve natural language-based interactions and grounded responses in AI systems.
- Skilled in preprocessing and cleaning large-scale text and document datasets to support ML and NLP model training and evaluation.
- Experience deploying and integrating AI/ML services using FastAPI, Docker, and Azure Functions, collaborating with senior engineers on production APIs.
- Developed vector search and retrieval pipelines, document comparison workflows, and agentic AI workflows using LangChain and LangGraph.
- Strong background in Databricks (PySpark, SQL, Delta Lake) for data ingestion, feature engineering, and preprocessing of structured and unstructured data.
- Solid foundation in machine learning fundamentals, including model training, evaluation, and cross-functional collaboration in fast-paced environments.

EDUCATION, TRAINING, CERTIFICATION

MS in Artificial Intelligence,New Jersey Institute of Technology. 2025

TECHNICAL ENVIRONMENT

Machine Learning & Deep Learning:	PyTorch, TensorFlow, XGBoost, Transformers (LLaMA, GPT, BERT), Keras, NLTK, spaCy
Gen AI & LLM:	RAG, HuggingFace Transformers, Lang Chain, GPT-3/4, LangChain, LangGraph, Prompt Engineering, Agentic AI, Azure OpenAI, Azure AI Search,CrewAI,MCP(Model Context Protocol)
MLOps & Model Deployment:	MLflow (experiment tracking, model registry), Docker, Azure Functions, FastAPI, Vertex AI (training, deployment, pipelines),Grafana
Big Data & Distributed Systems:	Apache Spark, Databricks
Programming & Frameworks:	Python, SQL, NumPy, Pandas,Scikit-Learn, Pydantic, Fast API, Golang
Tools:	Databricks, Jupyter Notebook
LLM & NLP:	Named-entity Recognition, Text Classification, Intent Detection, Summarization, Semantic Search, Vector Search, Text Analytics, Text Preprocessing

EXPERIENCE

Humana — Kentucky, USA (Remote) **Jan'26 to Till date**
AI/ML Engineer

Responsibilities:

- Designed and deployed RAG-based Generative AI solutions for internal healthcare and insurance documentation, including policy summarization, Q&A, semantic extraction, and document comparison.
- Built secure and compliant LLM pipelines using Azure OpenAI, LangChain, LangGraph, and Azure AI Search, supporting enterprise healthcare use cases.
- Implemented hybrid retrieval (vector + keyword search) with optimized chunking, metadata enrichment, caching, and grounding to improve response accuracy and traceability.
- Developed agentic AI workflows using LangGraph and MCP to automate multi-step internal processes.
- Created and maintained versioned prompt templates, improving response consistency, explainability, and task-specific accuracy.
- Built Python-based PDF ingestion and parsing pipelines to extract structured and unstructured healthcare content for vector indexing.
- Developed Databricks (PySpark, Delta Lake) pipelines for preprocessing large-scale structured and unstructured datasets used in retrieval and analytics.
- Supported LLM fine-tuning and evaluation on Azure ML to improve summarization and question-answering quality for healthcare documents.
- Deployed LLM-backed APIs using FastAPI and Docker, collaborating with platform teams on Azure-based production deployments.
- Applied LLMOps best practices, including experiment tracking, evaluation metrics, monitoring, and performance validation.
- Collaborated with senior engineers, data scientists, and compliance teams to deliver scalable and production-ready GenAI solutions.

New Jersey Institute of Technology **June'25 to Dec'25**
Assistant AI Mentor -Newark, NJ

Responsibilities:

- Mentored 60+ students in CS 116 through dedicated office hours and code reviews, guiding them in C++ manual memory management, pointer arithmetic, dynamic allocation, and memory leak prevention.
- Standardized course grading rubrics to systematically evaluate code quality, object-oriented programming (OOP) principles, multi-level inheritance, dynamic polymorphism, and generic template logic.
- Guided students in debugging and troubleshooting complex data structure implementations, including self-balancing Red-Black Trees, Min/Max Heaps, and adjacency-list Graph representations.
- Validated student-reported runtime execution against theoretical Big-O time and space complexity models, conducting empirical profiling to verify asymptotic performance bounds across varying input sizes.
- Assisted students in designing and refining Breadth-First Search (BFS) pathfinding algorithms, ensuring optimal state tracking, efficient queue utilization, and proper handling of cyclic graph structures.

AT&T — Hyderabad, India **Aug'22 to June'24**
AI&ML Engineer

Responsibilities:

- Engineered hybrid ML and Generative AI solutions—combining XGBoost/Random Forest with Large Language Models (LLMs like GPT-4, Llama 3)—for predictive customer behavior modeling, automated risk scoring, and generative financial synthesis.
- Constructed scalable data and embedding pipelines using Python, PySpark, and Vector Databases (Pinecone, Chroma, Milvus) to ingest and process millions of unstructured transaction and customer interaction records.
- Designed Retrieval-Augmented Generation (RAG) architectures and automated feature engineering workflows to create centralized vector data stores for enterprise LLM search and agentic applications.
- Deployed fine-tuned LLM and ML inference services on Azure OpenAI and Azure Machine Learning, exposing REST APIs via LangChain and LlamaIndex for real-time integration into banking platforms.
- Implemented LLM evaluation and monitoring frameworks (Ragas, DeepEval) to track precision, recall, F1-score, hallucination rates, semantic drift, and prompt security compliance.
- Developed AI-driven dashboards and conversational interfaces using Streamlit, PowerBI, and Matplotlib, leveraging Text-to-SQL and natural language generation for automated executive reporting.
- Spearheaded Responsible AI governance initiatives alongside analytics and compliance teams, enforcing guardrails for data privacy, PII masking, enterprise security, and financial regulatory standards.
- Enhanced financial decision-support workflows and automated document synthesis, boosting analytics delivery speed and reducing manual review turnaround times across core banking use cases.