Shreya

shreyapulluru9@gmail.com | (201) 644-5735 | Portfolio | LinkedIn | NJ

SUMMARY

- Versatile and detail-oriented Machine Learning Engineer with 4+ years of hands-on experience building and deploying intelligent AI systems focused on Natural Language Processing, Generative AI, and LLM-powered applications.
- Specializes in developing scalable end-to-end ML pipelines using transformers (BERT, GPT, LLaMA), LangChain, and vector databases (FAISS, Pinecone) for enterprise use cases like document intelligence, semantic search, and Q&A bots.
- Proven expertise across model fine-tuning (PEFT, LoRA), retrieval-augmented generation (RAG), and real-time model deployment using Docker, FastAPI, and cloud services (AWS, GCP, Azure).
- Known for bridging research innovation and production ML with robust MLOps practices and cross-functional delivery.

EXPERIENCE

JP Morgan & Chase, NY | Machine Learning Engineer (GenAI & NLP Focus) — Jan 2024 – Present

- Designed and deployed a Gen AI-driven NLP system for financial document classification using fine-tuned LLMs (GPT, BERT, LLaMA), on PyTorch and TensorFlow, enhancing fraud risk detection and document understanding across business workflows.
- Built an RAG-based intelligent Q&A assistant leveraging FAISS VectorDB, SentenceTransformers embeddings, with LLM inference and LangChain, enabling underwriters to retrieve compliant summaries with >85% semantic match accuracy.
- Containerized and deployed inference APIs using Docker and FastAPI on AWS SageMaker, integrating real-time LLM inference for RAG pipelines and reducing API response times by 40%.
- Led prompt tuning, LoRA-based optimization, and fine-tuning of generative models using PyTorch for use cases like clause extraction, document redlining, and semantic summarization in high-stakes financial reviews.
- Automated data pipelines for ingestion, pre-processing, and feature engineering from structured and semi-structured sources using Spark, SQL, and Python, enabling scalable batch and streaming inference with embedding generation.
- Developed Power BI dashboards for executive stakeholders to visualize LLM performance, embedding drift, and risk trends across business-critical processes.
- Created a multi-stage NLP pipeline integrating OCR, text segmentation, entity recognition, and summary generation, reducing document review time for underwriters by over 50%.

LTI Mindtree, India | Data Scientist – NLP & Predictive Analytics — Jan 2020 – Jul 2022

- Developed end-to-end text classification and intent detection pipelines using Scikit-learn, PyTorch, TensorFlow, and custom embeddings, optimizing model accuracy for real-time feedback analytics.
- Created NER and sentiment analysis models using SpaCy and Hugging Face Transformers, applied in enterprise document mining and contact center analytics with a 90%+ F1 score.
- Built NER and sentiment analysis models using SpaCy and fine-tuned BERT via Hugging Face Transformers, enabling automated extraction of key entities from unstructured documents achieving over 90% F1 score and reducing manual review time by 60%.
- Engineered batch and streaming pipelines via Azure Data Factory and PySpark, automating text preprocessing, embedding generation, and ingestion into FAISS-based VectorDBs, reducing model deployment cycle time and data prep efforts by 50%.
- Designed MLOps workflows using Docker, Flask APIs, and CI/CD pipelines for seamless LLM deployment on Azure and hybrid on-prem environments, powering RAG pipelines with sub-second inference latency and production-grade scalability.
- Developed interactive KPI dashboards in Tableau and Google Data Studio, visualizing insights from Neo4j-driven NLP outputs, entity relationships, and user engagement trends across business units for real-time decision-making.
- Collaborated with engineering teams to scale AI infrastructure by integrating Neo4j Graph DB, enabling relationship-driven analytics across customer complaint networks and optimizing resolution workflows through graph-based feature extraction.
- Built predictive models using Decision Trees, Random Forests, and SVM, enhancing customer segmentation, by conducting comprehensive analysis and statistical modeling using Python, Pandas and SPSS to identify key financial trends and anomalies.
- Designed and implemented multi-label text classification models using BERT and HAN, integrated with SentenceTransformer embeddings and FAISS VectorDB to automate and optimize support ticket triaging with over 88% routing accuracy.
- Built time series forecasting models with Prophet and LSTM hybrids on TensorFlow, used for predicting call center load and optimizing resource scheduling during peak hours.

SKILLS

Languages & Libraries: Python, SQL, R, Scala, Pandas, NumPy, Scikit-learn, TensorFlow, Keras, PyTorch, Seaborn, Matplotlib
Generative AI & NLP: BERT, GPT-3/4, LLaMA, OpenAI APIs, LangChain, RAG (Retrieval-Augmented Generation), Hugging Face Transformers, LoRA, PEFT, FAISS, Pinecone, Prompt Engineering, NER, Semantic Search
ML & Deep Learning: Linear & Logistic Regression, Decision Trees, Random Forest, SVM, KNN, Naive Bayes, Gradient Boosting, XGBoost, K-Means, Clustering, CNN, RNN, LSTM, ANN, Transformer Models (BERT, GPT, LLaMA), Time Series Forecasting,
MLOps & Deployment: FastAPI, Flask, Docker, Kubernetes, REST APIs, GitHub Actions, Azure DevOps, MLflow, CI/CD
Cloud Platforms: AWS (SageMaker, Lambda, EC2), GCP (Vertex AI, BigQuery), Azure AI Studio, Azure Data Factory
Data Engineering: Spark, PySpark, Hadoop, Airflow, SQL, Pandas, ETL Pipelines, Data Lakes
Visualization & Dashboards: Tableau, Power BI, Excel Solver, Google Data Studio
Databases & Tools: MySQL, PostgreSQL, MongoDB, Hive, SparkSQL, Neo4j, RStudio, JupyterLab
Certifications: Google Cloud Professional Data Engineer

Education

Masters in information systems | Stevens Institute of Technology, NJ, USA