Shreya

shreyapulluru9@gmail.com | (201) 644-5735 | 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 generative AI-driven NLP system for financial document classification using fine-tuned LLMs (GPT, BERT, LLaMA), enhancing fraud risk detection and document understanding across 10+ business workflows.
- Built an RAG-based intelligent Q&A assistant that combined FAISS vector search with LLM inference via LangChain and SentenceTransformers, enabling underwriters to retrieve compliant summaries with >85% semantic match accuracy.
- Containerized and deployed inference APIs using Docker and FastAPI on AWS SageMaker, facilitating real-time predictions and reducing API response times by 40%.
- Led prompt tuning, LoRA-based model optimization, and evaluation of generative models for use cases like clause extraction, document redlining, and semantic summarization.
- Automated data pipelines for ingestion, pre-processing, and feature engineering from structured and semi-structured sources using Spark, SQL, and Python, supporting batch and stream inference across millions of records.
- Developed visualization dashboards in Power BI for executive stakeholders to track model performance, data drift metrics, and risk trends.
- Created a multi-stage NLP pipeline combining OCR, text segmentation, entity extraction, and summarization, 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, 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.
- Engineered batch and streaming pipelines via Azure Data Factory and PySpark, reducing data prep time by 50% and increasing model update cadence.
- Designed MLOps pipelines with Docker, Flask, and CI/CD for secure, low-latency model deployment into production across Azure Cloud and on-prem platforms.
- Built customized KPI dashboards using Tableau and Google Data Studio, enabling product managers and business heads to monitor real-time sentiment and customer segmentation insights.
- Collaborated with engineering teams to scale AI model infrastructure, manage data lakes, and maintain reproducibility across environments with versioned data and models.
- Developed predictive models using Decision Trees, Random Forests, and SVM, enhancing customer segmentation.
- Conducted comprehensive Exploratory Data Analysis (EDA) and statistical modeling using Python, Pandas, R, and SPSS to identify key financial trends and anomalies.
- Designed and implemented multi-label text classification models using BERT and hierarchical attention networks (HAN), improving customer intent recognition in ticketing systems.
- Built time series forecasting models using Prophet and LSTM hybrids to predict call center traffic and guide model inference resource allocation 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, Hierarchical Clustering, Collaborative Filtering, 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

---

Masters in Information Systems | Stevens Institute of Technology, NJ, USA
Bachelors in Information Technology | G.Narayanamma Institute of Technology, India