Shaun Mendes

New York, NY | (551) 260-5614 | linkedin.com/in/shaun-mendes | smendes2901@gmail.com

TECHNICAL SKILLS

Languages: Python, R, Java, Scala, C++, C, Bash, Javascript, HTML, CSS, SQL, Prompt Engineering

Machine Learning & Deep Learning: Pandas, NumPy, PyTorch, TensorFlow, Keras, JAX, PyTorch Lightning, Hugging Face, Speech-to-Text, Audio Classification, Scikit-learn, NLTK, Spacy, PySpark, A/B Testing, Data Mining, Crew AI, Autogen

Models: GPT Series, LLama Series, BERT, RoBERTa, DeBERTa, Sentence Transformers, CNN, RNN, LSTM, Wav2Vec, Whisper

Cloud & Deployment: AWS, Azure, GCP, Flask, FastAPI, Streamlit, React, Linux, MySQL, Postgres; NoSQL – Redis, Cassandra, Elasticsearch; VectorDBs: LlamaIndex, FAISS; Docker, Kubernetes, MLflow, Kubeflow, AirFlow, Jenkins, Git, Gitlab CI/CD, Prometheus

PROFESSIONAL EXPERIENCE

Data Scientist | HERE Technologies (Chicago, IL) May 2024 – August 2024

- Led a successful Proof of Concept (POC) to assess the effectiveness and interpretability of **prompt-engineered** versus **fine-tuned Large Language Models (LLMs)** in extracting multilingual geospatial data to enhance the extraction efficiency of place attributes from text.
- Engineered an end-to-end **AI chatbot** with LLMs and **Retrieval Augmented Generation (RAG)** on AWS, using FastAPI and ReactJS. This reduced the onboarding time of data scientists/engineers by 40% and improved information extraction by 30%.
- Containerized and deployed **LLMs (Llama-3, OpenChat)** using Docker and **AWS Cloudformation** to detect, extract and format Hours of Operation from place website text, optimizing inference efficiency and achieving 85% extraction accuracy.

Senior Data Scientist | HERE Technologies (Mumbai, India) April 2021 – August 2023

- Expanded HERE Maps **global coverage by 17% and saved over $2.5M** by leveraging web-crawled data to generate 10M high-quality place records. Utilized ML, DL and LLMs to extract key place attributes, including name, category, address, and hours of operation.
- Identified place websites with an accuracy of 92.5% by creating labeled data using heuristics and unsupervised models (**K-Means, DBSCAN**) for clustering. This data was employed to train a **Mixture of Experts** models (**Random Forest, SVM**) for classification.
- Extracted street addresses, place names and hours of operation from 9 countries by supervised finetuning foundational models such as **T5**, **GPTJ** and **DeBERTa** on **Named Entity Recognition (NER)** and **Semantic Re-Ranking** achieving an overall accuracy of 94.3% .
- Enhanced classification of places across 400+ categories in 6 languages by adapting **Transformer(BERT, DeBERTa, XLNet)** models to unique regional nuances and improving classification metrics by 7% to 0.88 over previous benchmarks.
- Leveraged **GPT-3.5/4** for dataset annotation, accelerating training of Small Language Models and cutting labeling time by 40%.
- Built a scalable MLOps pipeline on AWS that **crawled 1M URLs daily** and achieved **25x cost reduction** in generating GPS data by using optimized GPU/Sagemaker cloud instances and model compression techniques such as Quantization and Knowledge Distillation.

Machine Learning Engineer | Fractal Analytics (Mumbai, India) August 2017 – April 2021

- Detected fraud activity in client's critical products saving over $200K annually by developing scalable Extract Transform Load (ETL) solution for processing terabytes of clickstream data with **PySpark** deployed on **AWS** EMR, utilizing Jenkins and Oozie.
- Optimized end-to-end data audit, dashboarding and product mapping processes for client sales data by engineering heuristics and machine learning models optimizing delivery timelines by 60% and **saving over $500K** in operational costs.
- Pretrained and fine-tuned a **Speech-to-Text** model(Wav2Vec2) on a dataset comprising 16,000 hours of Indian English audio and improved transcription accuracy by 27%. Integrated speaker identification and text-to-speech capabilities.
- Expedited Consumer Packaged Goods (CPGs) client acquisition by parallelizing the training of gradient boosting regression models **(XGBoost and LightGBM)** for demand forecasting, **reducing turnaround time by 75%.**

OPEN-SOURCE PROJECTS

- Developed a GPT-4o / Llama3 / Claude chatbot leveraging **LangChain, Ollama, Streamlit**, and **Elasticsearch** with RAG to provide real-time, Responsible AI-driven assistance to university students, faculty, and prospective applicants.[code]
- Worked on a robust Generative AI hybrid recommendation and question-answering pipeline by fine-tuning Llama-2 with **QLoRA**, **Direct Preference Optimization (DPO)** and using a Mixture of Experts (MoE) approach alongside RAG.[code]
- Modeled **collaborative filtering-based** recommender systems for personalized Instacart recommendations, comparing performance with TF-IDF, Singular Value Decomposition(SVD), and Bayesian Personalized Ranking(BPR) methods. [code]

EDUCATION

Stevens Institute Of Technology Hoboken, New Jersey
Master of Science in Machine Learning, GPA: 4.0/4.0 Sept 2023 - Dec 2024

Mumbai University Mumbai, India
Bachelor of Science in Electronics and Telecommunications Engineering Jul 2013 - Jun 2017