Saahir Dhanani

saahirdhanani@gmail.com | (409) 365-3693 | saahird.com | linkedin.com/in/saahird | github.com/saaahir

Experience

Machine Learning Engineer, **NCR Voyix** July 2024 – Present

- Redesigned time series forecasting system for sales and item demand across hundreds of restaurant locations, improving model accuracy by 30% and reducing training time by ∼70x
- Led migration from Kubeflow to Databricks, building a scalable ML platform that cut infrastructure costs by 80% and enabled faster experimentation and deployment
- Built production-grade ML pipelines for data ingestion, feature engineering, training, and daily inference using Spark Structured Streaming, Databricks Workflows, and BigQuery, processing multi-terabyte datasets
- Implemented automated model monitoring and retraining pipelines using MLflow Model Registry with drift detection, enabling continuous deployment of top-performing models without manual intervention
- Engineered event-driven data infrastructure using GCP (Cloud Functions, Workflows, Eventarc) and Terraform to automate data streaming from BigQuery to Databricks
- Developed GPT-4 powered text-to-SQL analytics interface with programmatic few-shot prompting and dynamic schema injection, enabling store managers to query sales data without SQL; integrated with React, BigQuery, and Plotly

Research Assistant, **Data to Insights Lab @ Georgia Institute of Technology** Aug 2023 – May 2024

- Co-authored and published 'Demonstration of VCR: A Tabular Data Slicing Approach to Understanding Object Detection Model Performance' in VLDB 2024
- Developed tabular slicing approach for object detection model debugging, combining vision foundation models with frequent itemset mining to automatically discover and semantically label failure modes at scale
- Engineered multimodal vision pipeline integrating Meta's SAM for segmentation and DINOv2 + OpenAI CLIP for embeddings to extract thousands of visual concepts, enabling semantic search and interpretation across 100,000+ images

Software Engineering Intern, **NCR Corporation** May 2022 – Aug 2023

- Built high-performance MVPs using Firestore, MongoDB Realm, and Flutter to migrate on-premise restaurant management features to the cloud, improving scalability and reducing operational costs
- Implemented offline synchronization and data persistence using Ditto, enabling reliable functionality for 10,000+ global restaurant locations even during network outages
- Awarded "Best Overall Project" out of 200+ interns for innovation, scalability, and business impact

Projects

ShorthandAI June 2025 – Present

- Built an end-to-end natural language to Apple Shortcuts generation tool using a multi-stage agentic LLM pipeline with Retrieval Augmented Generation (RAG), vector embeddings, and iterative compilation validation with error correction
- Constructed a large-scale dataset by scraping and reverse-engineering Apple Shortcuts collected from the web, extracting triplets (natural language intent, Domain Specific Language (DSL) code, plist/XML) to enable model fine-tuning with LoRA
- Implemented context-aware code generation by embedding DSL documentation using OpenAI text-embedding models to enable semantic search and information retrieval of relevant code examples

Transformer Architecture from Scratch Nov 2025

- Implemented complete Transformer encoder-decoder architecture from "Attention Is All You Need" paper in PyTorch from first principles, including custom multi-head attention, positional encoding, and layer normalization
- Trained models on sequence-to-sequence tasks including string manipulation and machine translation with custom character-level tokenizer and end-to-end training pipeline

Education

Georgia Institute of Technology Aug 2023 – May 2024

- M.S. in Computer Science | Concentrating in Machine Learning | *GPA: 4.0*

Georgia Institute of Technology Aug 2020 – May 2023

- B.S. in Computer Science | Concentrating in Intelligence and Systems & Architecture | *GPA: 3.97*
- *Relevant Coursework*: Data Structures & Algorithms, Artificial Intelligence, Machine Learning, Deep Learning, Algorithms Design, Processor Design, High Performance Computer Architecture, Compilers, Systems & Networks, Operating Systems

Skills

Languages: Python, C/C++, SQL, Java, JavaScript/TypeScript

ML & Data: PyTorch, TensorFlow, scikit-learn, Transformers, LangChain, Optuna, Spark, PySpark, BigQuery, Delta Lake, Pandas

MLOps & Infrastructure: MLflow, Databricks, Kubeflow, Docker, Kubernetes, Terraform, AWS, GCP, Git, GitHub