Dushyant Mahajan

857-437-2831 • dushyantinboston@gmail.com • Portfolio • LinkedIn • Github
Boston, MA • Available February 2025 • Work Authorization: F1-OPT

Professional Experience

Graduate Research Assistant | Institute of Experiential AI | Boston, MA *May 2024 - Oct 2024*

- Researched and implemented a RAG pipeline for summarizing Electronic Health Records (EHRs) using **MIMIC-III dataset**, **reducing hallucinations** and improving **contextual accuracy** in generated summaries.
- Engineered and enhanced prompt-based RAG workflows using a robust question driven **text summary** framework, achieving a 30% reduction in clinician chart review time, using **locally hosted** LLM models for patient data privacy.

Data Scientist | Raga AI | Fremont, California *Jan 2024 - Jul 2024*

- Collaborated to open-source the Raga LLM Hub framework, enriched with over 50 **evaluation metrics** and critical **guardrails** for LLMs and RAG applications, enhancing response accuracy for models.
- Developed RagaAI Catalyst, an observability platform to enable **trace recording** within RAG applications; this solution streamlined deployment processes and reduced setup time by 30% for **LLM evaluation**.
- Built **"RAG Builder"**, a package featuring modular functionality that empowered users to create and optimize customizable components, resulting in a 40% reduction in development timelines for bespoke RAG pipelines.
- Pioneered **prompt engineering** for response optimization, resulting in a **60% increase** in relevant and precise responses.

Data Science Consultant | Raga AI | Bangalore, India *May 2022 - Aug 2022*

- Architected an efficient API pipeline with an interactive dashboard to visualize **clustering** patterns in **DNN embeddings**; improved the speed of data insights generation for **image classification** tasks by over 30%.
- Applied rigorous MMD and Kolmogorov-Smirnov methods to monitor **data drift** across two critical image classification projects; delivered **actionable insights** leading to the identification of four major unseen biases affecting model performance.
- Leveraged AE, VAE, Variational Auto-Encoding Gaussian Mixture Model (VAEGMM) algorithms to identify outliers in high-dimensional datasets, improving **anomaly detection** accuracy by 40%.

Software Engineer | Askim Technologies | Mumbai, India *Jan 2021 - May 2022*

- Led the development of a multimedia prescription platform, creating a **python pipeline** that processed **high volumes** of paper prescriptions, improving clarity, and reducing patient follow-up queries by **30%**.
- Developed and launched a robust full-stack application on **AWS** utilizing the **MERN stack**, implementing **multi AZ architecture** with system uptime of **99.9%** while ensuring security with **HTTPS-enabled CRUD endpoints**.
- Orchestrated automation processes that generated the most current version of **Ubuntu AMI** using **HashiCorp Packer** through **GitHub Actions** while ensuring quality checks decreased manual errors in deployments considerably over three months.
- Automated the provisioning of **AWS services** - **Route53, VPC, EC2, RDS, S3, SNS, Lambda, DynamoDB, IAM, CloudWatch** with **Pulumi IaC**.

Projects

Parallelization Techniques in Deep Learning for Weather Classification in Images |

- Leveraged custom CNN architecture and PyTorch to process weather images for classification, achieving **78% F1-score**.
- Employed PyTorch's **DP (Data Parallel)** and **DDP (Distributed Data Parallel)** for model training across 1-4 GPUs on a HPC cluster, recording a **1.41x speedup**.

Fine-Tuning Stable Diffusion for Synthetic Skin Rashes Image Generation

- Created a **Streamlit application** capable of producing high-fidelity images reflecting diverse rash characteristics across **50+ variations** in skin tone and anatomical locations, improving educational tools for dermatology professionals.
- Fine-tuned Stable Diffusion using **Dreambooth** to generate text embeddings, conditioned the **U-Net** on those embeddings to generate four realistic and detailed image variations per input query.

Publication

- Paper titled "Roux-lette at Discharge Me! Reducing EHR (Electronic Health Record) Chart Burden with a Simple, Scalable, Clinician-Driven AI Approach" published in **Association for Computational Linguistics 2024** | Link.

Technical Skills

Programming & Frameworks: Python, C, PyTorch, TensorFlow, Keras, Pandas, NumPy, Scikit-Learn, Flask, PySpark, SQL

Cloud Technologies: AWS, AWS SageMaker, S3, MLflow, Azure ML Studio, Docker, Pulumi, GitLab, GitHub Actions, Packer

Domain Expertise: AI, Machine Learning, Deep Learning, Generative AI, NLP/NLU, LLM, Hugging Face, LangChain, Transformers, MLOps, Kubernetes, XGBoost, Linear Regression, Random Forest, CNN, Unit/Regression Testing

Education

Northeastern University | Master of Science in Information Systems | Boston, MA *Sep 2022 - Dec 2024*

University of Mumbai | Bachelor of Engineering in Computer Engineering | Mumbai, India *May 2015 - Oct 2020*