Dushyant Mahajan

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

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*

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) from **MIMIC-III dataset**, **reducing hallucinations**, and improving chunk **retrieval accuracy** utilizing prompt and context-length optimization.
- Engineered RAG workflows with robust QA framework, and **zero-shot prompting** approach on locally hosted **Llama models**, reducing clinician chart review times by 30% and maintaining patient data privacy.

Data Scientist Co-op | 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 Langchain and LlamaIndex RAG applications; this solution streamlined deployment processes and reduced setup time by 30% for **LLM evaluation**.
- Built **"RAG Builder"**, a package featuring modular functionality managing document loaders, embedding models, vector databases, KV stores and reranking, resulting in a 40% reduction in development timelines for end-to-end bespoke RAG pipelines.
- Pioneered **prompt engineering** for response optimization, resulting in a **60% increase** in relevant and precise responses.
- Finetuned quantized Qwen2-7B model, using Peft technique (LoRA), on text moderation dataset for improved response evaluations.

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

- Architected efficient Fast API pipeline with an interactive dashboard to visualize **clustering** patterns in high-dimensional datasets.
- 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.
- Created optimized Docker images and testing suites for CI/CD pipelines with MLflow and DVC(Data Version Control) integration.
- Applied CLIP model embeddings to perform text-based image search leading to 70% reduction in clients image tags processing.

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

- 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**.
- Increased click-through rate by 15% through **A/B testing** UI features utilizing CloudWatch logging metrics for real-time monitoring.
- 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

Career Performance Coach for Streamlining Job Search with Intelligent Agents

- Implemented an agentic workflow with specialized agents leveraging **AutoGen** to deliver tailored job recommendations and skill-building plans based on user profile and job preferences, enhancing user job search success.
- Optimized agent efficiency by implementing token optimization, **dynamic batching**, and **query caching**, reducing LLM operational costs by 30% and improving system responsiveness and scalability.

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 minimizing FID score.

Publications

Roux-lette at Discharge Me! Reducing EHR (Electronic Health Record) Chart Burden using LLMs - Accepted in ACL 2024.
Link: Proposes a new approach that mimics human clinical workflow, to summarize insights from EHR data.

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