DUSHYANT MAHAJAN

(857)-437-2831 | dushyantinboston@gmail.com | LinkedIn | Portfolio | GitHub

Professional Experience

OptiMe Health | Boston, MA **Apr 2025 - Present**

AI Software Engineer

- Led development of **agentic recommendation** pipeline parsing patient's symptoms and emotions to provide insights, improving user session duration by 25%
- Deployed serverless AWS Lambda **microservices** to transform unstructured medical documents into structured data
- Implemented **HIPAA** compliance controls across backend services, including encryption, IAM access policies, and audit logging, ensuring secure handling of sensitive health data and AI workflows
- Developed a personalized educational chatbot using **RAG**, boosting user engagement metrics (course completion rates) by 30%
- Finetuned gpt-oss 20B, using PEFT technique (**LoRA**) to enhance response generation for patient-specific use cases
- Built 30+ RESTful APIs for LLM integration on AWS, leveraging **Flask** and **Redis** for scalable, event-driven architecture

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

Research Assistant

- Researched and improved a locally hosted **Llama RAG pipeline** on de-identified MIMIC-IV **EHR** dataset, combining retrieval QA, prompt controls, and context-length tuning to reduce hallucinations and improve retrieval precision
- Achieved **BERTScore F1 0.86** and reduced clinician chart review time by **30%** through targeted retrieval QA, hallucination checks, and evaluation automation. Paper Link- aclanthology.org/2024.bionlp-1.63/

Raga AI | Fremont, CA **Jan 2024 - Jul 2024**

Data Scientist

- Collaborated to **open-source** RagaAI Catalyst, an **observability** platform that aids RAG and agentic systems debugging, facilitating to identify and resolve performance bottlenecks, earning **15k+** GitHub stars
- Collaborated on **LLM evaluation** framework using **LLM-as-a-judge** techniques, PII detection guardrails, and automated evaluation pipelines to assess model quality across **50+** metrics, establishing crucial quality benchmarks for production **RAG applications** and improving LLMs response accuracy
- Led end-to-end testing of RAG and agentic workflows across multiple LLM providers, developing optimal prompting strategies and establishing performance benchmarks for production deployments

Data Science Consultant **May 2022 - Aug 2022**

- Collaborated cross-functionally to create the Raga AI platform for computer vision **drift detection**, model A/B testing using CNNs and **anomaly detection**, directly contributing to secure **$4.7 million** in seed funding
- Transformed client's image workflow with **CLIP-based semantic search**, reducing image-tagging turnaround by **70%** while maintaining high accuracy and relevance
- Delivered PoCs for automated retail checkout, driver monitoring, and satellite imaging using **synthetic data generation**, A/B tests on CNN baselines, and **outlier detection**; partnered with domain experts to validate results and hand off for productization
- Created **Docker**-based CI/CD pipelines with **ML Flow** and **DVC** for robust ML model testing and deployment

Askim Technologies | Mumbai, India **May 2020 - May 2022**

Founding Engineer

- Led **DevOps** initiatives by architecting and deploying a highly available MERN stack application on **AWS**, implementing multi-AZ infrastructure that achieved 99.9% uptime
- Automated infrastructure provisioning using **Pulumi** (IaC) for **AWS** services (VPC, EC2, RDS, Lambda), ensuring consistent, scalable, and reliable cloud environments.
- Integrated Docker containers with AWS **Lambda** functions for **serverless** processing of image and video generation pipeline
- Increased click-through rate by 15% through **A/B testing** UI/UX features utilizing **CloudWatch** logging metrics for monitoring
- Streamlined an automated AMI generation pipeline using HashiCorp Packer and **GitHub Actions**, establishing a reliable CI/CD process that reduced deployment errors and infrastructure updates

Skills

Core ML Skills: NLP, RAG, Agents, Transformers, BERT, GPT, Diffusion models, Reinforcement Learning, Computer vision, Deep Learning, MLOps, Statistics, OpenCV, Segmentation, Multimodal, GAN

Cloud & MLOps: AWS, GCP, Vertex AI, Docker, GitHub

Programming and Tools: Python, Go, PyTorch, vLLM, AutoGen, CrewAI, Pydantic, DSPy, Deepchecks, Langchain, SQL, MongoDB

Education

Northeastern University **Sep 2022 - Dec 2024**

Master of Science, Computer Software Engineering | ***GPA**: 3.67/4.00* *Boston, MA*