Kshitij Duraphe

Boston, MA | kshitijduraphe5@gmail.com | LinkedIn | Github | Portfolio | 314-886-3066

Professional Experience

Thespian Labs **Somerville, MA**

AI Engineer Nov 2025 – Present

- Managed the lifecycle of custom **Text2Motion** deep learning models from development, deployment, and monitoring, ensuring optimal performance and stability on **GCP**. Published an open-source library for MLOps: Link.
- Developed and maintained batch processing ETL pipelines for data processing of **5000+** hours of human performance data.
- Collaborated with **cross-functional teams** to build a SOTA **foundation model** for **controllable digital human performance generation** on the frontier of **human-computer interaction**.

Absentia Technologies **Boston, MA**

Founding Machine Learning Engineer Jan 2025 – Nov 2025

- Architected and deployed a production-ready SaaS video analysis platform from the ground up on AWS, establishing core AI service infrastructure and a full CI/CD pipeline with Terraform and Docker, reducing deployment cycles from days to **<2 hours**.
- Built the core infrastructure and APIs enabling a fleet of autonomous **AI agents** to **read** video streams, **understand** complex events through a reasoning engine, and **act** by flagging anomalies in real-time using Python, PyTorch, and Kafka on Kubernetes (K8S).
- Engineered the platform's high-performance inference service for computer vision models (ViM/SwinV2), achieving **<10ms p99 latency** through model quantization and optimized data loaders.
- Established a rigorous, automated **evaluation pipeline** to benchmark agent performance and API latency, reducing model regressions by 40%; improved system stability to **99.9%** by resolving a critical memory leak during on-call duties.
- Developed a multimodal Video Question-Answering (VideoQA) system for complex temporal reasoning, improving answer accuracy by **25%**, and built a synthetic data pipeline with diffusion models to reduce false positives by **15%**.

The KeelWorks Foundation **Oak Harbor, WA**

Software Engineer (Machine Learning Applications) July 2024 – Jan 2025

- Engineered a production-scale **retrieval and reasoning system** using a RAG pipeline in TypeScript/Python with LangChain, delivering a search API with **<5s p95** latency and **92% context relevance (MRR)** across 2,500+ documents on a PostgreSQL backend.
- Executed an aggressive model optimization strategy for production deployment, reducing model size by **50%** and increasing inference speed by **80%** using **8-bit GPTQ quantization** and knowledge distillation while maintaining >90% task accuracy.
- Pioneered a synthetic data generation workflow by fine-tuning **Mistral-7B**, expanding the training dataset for downstream tasks by **30%** and significantly improving model robustness; deployed all services via a Dockerized GitLab CI/CD workflow.

Space Physics Lab, Boston University **Boston, MA**

Graduate Research Assistant (ML Applications) Oct 2022 – May 2024

- Led R&D of a multimodal system to parse and reason over noisy, unstructured sensor data; Architected a high-throughput, distributed data ingestion system using Kafka, Dask, and AWS S3, slashing processing time for 3TB+ datasets from >24 hours to **<3 hours**.
- Developed a low-latency forecasting system achieving **<80ms p90** latency at **25 predictions/sec** by implementing async processing and request batching, which cut initial latency by 40%.
- Implemented **generative inpainting** and **SwinIR super-resolution** pipelines to reconstruct corrupted sensor data, improving data quality and signal-to-noise ratio for downstream predictive models by over 60%.

Projects

ArkOS (MIT) **Jun 2025 – Present**

- DevOps, documentation **(Mintlify)** and general development (frontend/backend) for ArkOS, an open source interface for a local LLM agent building utilizing long term memory for personalized requests

Halo AI (Stealth startup incubated at Columbia University) **Dec 2023 – Aug 2024**

- Developed and deployed a federated learning pipeline and ensemble of LLMs for on-device **agentic conversational assistants**, cutting inference latency by **27%** to **<1.5s**.
- Architected a scalable MLOps pipeline on AWS (**EC2/S3/Lambda, SageMaker**), automating CI/CD for federated learning models and reducing data preparation time by **70%**.
- Implemented an **evaluation pipeline** to monitor model performance, utilizing knowledge distillation and 8-bit quantization to improve inference speed by **80%** while preserving **90%** accuracy.

BeatQraft (MIT iQuHACK 2023 Hackathon) **Jan 2023**

- Built a **distributed quantum generative AI** service at the MIT iQuHACK-23 hackathon, placing **2nd** out of 1000+ teams. Link

Technical Skills

Programming & Databases:	Python, C++, TypeScript, SQL (Postgres), MongoDB, Redis
MLOps & Cloud Platforms:	Docker, AWS, GCP, Kafka, Dask, Terraform, Kubernetes, Modal, MLflow
Machine Learning & Tools:	PyTorch, GenAI, RAG, LLMs, Computer Vision, Retrieval Systems, Distributed Systems, AI Code Assistants
Core Concepts:	Software Architecture, Product Intuition, High-Performance Computing, Scalability, Agile

Education

Boston University **Boston, MA**

Master of Science with Thesis in Electrical and Computer Engineering Sep 2022 – May 2024
GPA: 3.8/4

College of Engineering Pune **Pune, India**

Bachelor of Technology in Electrical Engineering, Minor in CS Aug 2018 – June 2022
GPA: 3.83/4

Publications

1. **Optimizing Solar Panel Tilt using Machine Learning Techniques**, GPECOM 2021.
   Proposes an XGBoost-based approach to maximize energy generation from solar plants.
2. **The Platonic Universe: Do Foundation Models See the Same Sky?**, NeurIPS ML4PS 2025 - Spotlight Paper.
   Investigates if different foundation models see the same underlying astrophysical phenomena and develops custom foundation models to better learn underlying astrophysics.