Abhi Lad

New York, NY | (551)-344-7494 | abhinlad@gmail.com | Linkedin.com/abhilad1009

EDUCATION

Columbia University New York, NY
M.S. in Computer Science, **GPA: 3.87/4.0** Dec 2023

- Courses: Artificial Intelligence, Machine Learning, Natural Language Processing, High Performance ML

Pandit Deendayal Energy University Gandhinagar, India
B.Tech. in Computer Engineering, **GPA: 3.96/4.0** Jul 2020

- Courses: Data Structures, Algorithms, Database Systems, Big Data and Cloud Computing, Software Engineering

TECHNICAL SKILLS

Programming: Python, Javascript, C++, C, Java, HTML, CSS, SQL, NoSQL, PHP, LaTeX
Frameworks: PyTorch, TensorFlow, Keras, Scikit-learn, OpenCV, NLTK, spaCy, FastAPI, Pandas, Numpy, React
Developer Tools: AWS, GCP, Docker, Kubernetes, MLFlow, WandB, MongoDB, Git, PostgreSQL, REST, Linux, Jira

PROFESSIONAL EXPERIENCE

Aktus AI New York, NY
Machine Learning Engineer Feb 2024 – Present

- Fine-tuned Llama3, Mixtral using QLoRA and DPO/RLHF on VertexAI, used SentenceBERT embeddings, improving RAGAS metrics by 23%, retrieval recall by 12%
- Reduced latency of FastAPI app by ~2x using Nvidia NIM, and multi-threading Langchain RAG, VectorDB, PostgreSQL API calls
- Devised GenAI solutions for Fortune 500 firms, creating AI Search Agents, multimodal Document AI, LLM Evaluation and A/B testing platform

Skylark Labs Inc San Francisco, CA (Remote)
Machine Learning Intern Jun 2023 – Aug 2023

- Prototyped few-shot self learning Docker pipeline with YOLOv8, Swin transformer achieving 100+ detections/sec
- Optimized active learning based label refining with BIRCH Clustering, improving query resolution time by 50%

Origin Health AI Bangalore, India
Machine Learning Engineer Jun 2021 – Jul 2022

- Led team of 5, built deep learning based AI diagnosis tool with 93% agreement rate for study presented at FMF’22
- Accelerated experiments with ETL redesign improving loading times by 20%, segmentation performance by 23%, automating clinical record deidentification and ultrasound image processing
- Scaled prototyping, analyzed 20+ Pytorch AI models, 4M+ images by deploying multimodal Python framework on AWS GPU infrastructure (EC2, CUDA, S3)
- Managed model registry, hyperparameter sweeps on 30+ Segmentation (Segformer, UNet, DeepLab, GAN), Classification (ViT, ResNet, Inception), Object Detection (YOLO, DETR) neural network architectures

Kontiki Vision Labs Bangalore, India
Full-Stack Engineer Feb 2021 – May 2021

- Achieved 20% increase in fps by translating Javascript codebase to TensorFlow Blazepose pose detection model
- Developed online multiplayer Node.js application with body tracking (Google MediaPipe) for 50+ exercise routines

Summer Research School Gandhinagar, India
Co-founder & Researcher Jun 2020 – Jun 2021

- Published 6 research articles in ACM, IEEE, MICCAI, mentored 5 students, managed projects and resources
- Secured 1st place in GWC’21 (ICCV) with 0.715 ADA, annotated lane-classification dataset with 300k images

PROJECTS

GPT-2 Fine-Tuning and Distillation | GitHub

- Fine-tuned and distilled GPT-2 reducing model size by 33%, MSE by 29% , served HF model on AWS Sagemaker

NLP driven Product Design

- Utilized Sentiment Analysis, Topic Modelling, Keyword Analysis, Data Visualization to ideate top 3 hypotheses

Storyboard – AI assisted Social Media Platform | GitHub

- Integrated OpenAI ChatGPT and DALL-E for automated text summarization, language translation, art generation
- Deployed AWS serverless app using Lambda, DynamoDB, S3, Elasticsearch, CodePipeline and CloudFormation