Rohan Mehta

rohan.mehta@email.com +1 123 456 7890 rohanmehta rohanmehta rohanmehta.dev

Machine Learning Engineer with 4+ years of experience designing and deploying scalable deep learning pipelines. Proven track record of improving model performance, optimizing inference latency, and driving impactful AI solutions in cloud environments.

Technical Skills

Languages: Python, C++, SQL
ML/DL: PyTorch, TensorFlow, JAX, scikit-learn, Hugging Face
MLOps: Docker, Kubernetes, MLflow, Airflow, Ray
Cloud: AWS SageMaker, GCP Vertex AI

Experience

Senior Machine Learning Engineer Jan 2023 – Present
TechNova AI *San Francisco, CA*

- Architected a distributed training pipeline using Ray and PyTorch, reducing LLM fine-tuning time by 40% and cutting AWS compute costs by $50k annually.
- Developed and deployed a real-time recommendation engine with GCP Vertex AI, serving over 1M requests per day with a P99 inference latency of under 45ms.
- Improved the core ranking model's AUC from 0.78 to 0.85 by incorporating cross-attention mechanisms and sparse user behavior features.

Machine Learning Engineer Jun 2020 – Dec 2022
DataSphere Solutions *New York, NY*

- Built and maintained an MLOps platform using Kubernetes, Docker, and MLflow, enabling CI/CD for 15+ machine learning models across 3 teams.
- Designed a computer vision defect detection system utilizing a custom ResNet architecture in TensorFlow, achieving a 94% F1-score and automating 80% of manual QA processes.
- Optimized BERT-based NLP models using model quantization and ONNX Runtime, reducing memory footprint by 3x without significant accuracy drop.

Selected Projects

Distributed LLM Serving Engine | *C++, Python, PyTorch, gRPC* 2023

- Implemented continuous batching and PagedAttention algorithms from scratch, increasing token throughput by 2.5x compared to standard Hugging Face Transformers.

Multimodal Semantic Search | *Python, Hugging Face, Qdrant, FastAPI* 2022

- Engineered a scalable vector search API combining CLIP embeddings and a Qdrant vector database, enabling sub-second retrieval over a 10M+ image/text dataset.

Publications & Preprints

Mehta, R., Doe, J., & Smith, A. "Efficient Cross-Attention Mechanisms for Sparse Vector Retrieval." *NeurIPS 2023*

Doe, J., & **Mehta, R.** "Scaling Multi-Agent Reinforcement Learning with Ray." *arXiv:2401.12345 (2024)*

Education

MSc in Machine Learning Sep 2018 – May 2020
Stanford University *Stanford, CA*

BSc in Computer Science Sep 2014 – May 2018
University of California, Berkeley *Berkeley, CA*

Achievements

- **Kaggle Competition Master:** Ranked in the top 1% globally out of 150k+ data scientists.
- **1st Place, NeurIPS 2022 Challenge:** Won the robust computer vision challenge against 200+ international teams.