Kunj P. Shah

Machine Learning Engineering Intern | ML Systems, NLP, Model Training & Deployment
kunjcr2@gmail.com | +1 (628)-529-6990
LinkedIn | Github | Portfolio | San Francisco, CA

EDUCATION

San Jose State University — B.S. Computer Science | GPA: 3.96/4.00 | Dean's List
Expected Graduation: May 2027

EXPERIENCE

ML Intern, Routes Technologies, Remote, TX *Oct 2025 – Present*

- Developed, trained, and deployed ML models using PyTorch and Transformers, with experiment tracking and model governance via Weights & Biases, ensuring reproducibility and monitoring.
- Built scalable data ingestion pipelines (Scrapy, BeautifulSoup) to collect, clean, and validate large-scale web data under ethical and compliance-aware constraints.
- Designed and deployed REST-based ML endpoints on Azure ML Studio, supporting secure model inference and lifecycle management.
- Collaborated cross-functionally to translate business requirements into data-driven ML solutions.

AI Engineer Intern, Dreamable Inc., San Francisco, CA *May 2025 - Aug 2025*

- Fine-tuned large-scale transformer models (Qwen-2.5-7B) using PyTorch, Hugging Face, and LoRA, optimizing for accuracy, cost-efficiency, and memory constraints.
- Led dataset curation, cleaning, and validation using Pandas, NumPy, and Hugging Face Datasets to ensure high-quality training data.
- Evaluated models using systematic experimentation and hyperparameter tuning, reducing validation loss and improving generalization; tracked results via Weights & Biases.
- Built and deployed ML services on Google Cloud Run, supporting scalable and reproducible inference pipelines.
- Collaborated with engineers and product stakeholders to align ML solutions with real-world business requirements.

PROJECTS

MedAssistGPT-401M Github | Huggingface | WandB

- Pretrained a 401M-parameter transformer model from scratch using PyTorch on 2M+ domain-specific documents, focusing on large-scale data preprocessing, training stability, and evaluation.
- Designed efficient ML pipelines using memory-mapped datasets and automated preprocessing to support reproducibility and scalability.
- Applied modern transformer techniques (RoPE, GQA, RMSNorm) and optimized training on A100 GPUs with bf16 and Flash Attention.
- Logged experiments, metrics, and checkpoints using W&B and Hugging Face Hub for transparent model tracking.

Qwen-2.5-0.5B Finetune Github | Huggingface | Dockerhub

- Fine-tuned transformer models using supervised and preference-based learning (DPO) for instruction-following and text understanding tasks.
- Implemented efficient training strategies (LoRA, gradient checkpointing, bf16) to reduce compute, memory costs.
- Built a production-ready inference pipeline using vLLM and Docker, enabling scalable and low-latency deployment.
- Emphasized experiment tracking, versioning, and reproducibility across training and deployment stages.

SKILLS

Machine Learning: Supervised & Unsupervised Learning, Deep Learning, Model Evaluation
Natural Language Processing: Transformers, Fine-tuning, Text Classification
Programming: Python
Libraries: PyTorch, Hugging Face, Scikit-learn, NumPy, Pandas
MLOps & Cloud: Docker, Azure ML, GCP, Weights & Biases