Kunj P. Shah

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

SKILLS

Programming: Python, C++, SQL

GPU & Acceleration: CUDA (conceptual), Flash Attention, bf16, GPU Memory Optimization

ML Systems: PyTorch, cuDNN (via PyTorch), Distributed Training (data parallelism)

MLOps & Infra: Docker, vLLM, Git, Linux, Cloud GPU Environments

EXPERIENCE

AI / Machine Learning Engineer Intern, Routes Technologies, TX *Oct 2025 – Present*

- Developed, trained, and evaluated machine learning and NLP models using **Python, PyTorch, and Transformer architectures**, focusing on performance, scalability, and reliability.
- Designed and maintained **data ingestion and preprocessing pipelines** using Scrapy, BeautifulSoup, Pandas, and NumPy to collect and validate large-scale web datasets.
- Built and maintained **Python-based backend services** for ML inference, integrating REST APIs and cloud deployment workflows.
- Implemented experiment tracking, versioning, and monitoring with **Weights & Biases** to ensure reproducibility and model governance.
- Collaborated in cross-functional engineering teams to move business requirements into AI/ML solutions.

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

- Fine-tuned **LLMs (Qwen-2.5-7B)** using **PyTorch, Hugging Face Transformers, and LoRA**, optimizing accuracy under compute and memory constraints.
- Curated, cleaned, and validated datasets using **Pandas, NumPy, and Hugging Face Datasets** to support supervised and preference-based learning workflows.
- Conducted systematic experimentation and hyperparameter tuning, reducing validation loss and improving generalization; tracked results using **Weights & Biases**.
- Built and deployed **containerized ML services** on **Google Cloud Run**, enabling scalable and cloud-native inference pipelines.
- Worked closely with product and engineering stakeholders to deliver ML features aligned with real-world requirements.

PROJECTS

LLMs from Scratch (Research Repository) GitHub

- Built **GPU-accelerated training pipelines** for large language models using PyTorch, Flash Attention, and memory-efficient attention mechanisms.
- Optimized **GPU utilization and memory throughput** using bf16 precision, gradient accumulation, and memory-mapped datasets to reduce RAM overhead.
- Implemented parallel data loading and preprocessing pipelines to maximize training throughput on A100 GPUs.
- Designed modular training codebases enabling rapid experimentation with attention kernels and inference backends.

MedAssistGPT-401M Github | Huggingface | WandB

- Trained a **401M-parameter transformer model** on A100 GPUs, focusing on **training acceleration, memory efficiency, and checkpoint reliability**.
- Integrated Flash Attention, mixed precision (bf16), and automated checkpointing to support long-running GPU workloads.