

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, SQL

AI/ML: Large Language Models, NLP, Deep Learning, Transformers

Research: Experiment Design, Model Evaluation, Hyperparameter Tuning, Data Analysis

Tools: PyTorch, Hugging Face, Pandas, NumPy, Weights & Biases

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

MedAssistGPT-401M [Github](#) | [Huggingface](#) | [WandB](#)

- Pretrained a **401M-parameter transformer model from scratch** on 2M+ domain-specific documents using **PyTorch**, focusing on training stability and scalability.
- Implemented modern transformer components including **RoPE, GQA, RMSNorm, and SwiGLU**, optimizing training on **A100 GPUs** with **bf16** and **Flash Attention**.
- Optimized transformer training on **A100 GPUs** using **bf16, Flash Attention, and memory-efficient attention**, improving throughput and stability.
- Logged experiments, checkpoints, and metrics using **Weights & Biases** and **Hugging Face Hub**.

LLMs from Scratch (Research Repository) | GitHub

- Designed and implemented **large language models from first principles**, translating recent research papers into working PyTorch implementations across transformer, MoE, SSM, and dynamic-compute architectures.
- Conducted **comparative experiments** across architectural choices (RoPE, GQA, MLA, MoE, SSMs, MoD), analyzing effects on **training stability, convergence behavior, memory efficiency, and throughput**.
- Built **reproducible training and evaluation pipelines** using Flash Attention, gradient accumulation, memory-mapped datasets, and Weights & Biases for systematic experiment tracking.
- Pretrained and fine-tuned multiple models (up to **401M parameters**), documenting experimental setups, results, and insights through structured notebooks and reports.