Kunj P. Shah

kunjcr2@gmail.com | +1(628)-529-6990
LinkedIn | Github | Portfolio | San Francisco, CA

EDUCATION

San Francisco State University San Francisco, California
B.S. in Computer Science

- GPA: 3.96/4.00, *Dean's List* Expected Graduation 2027

EXPERIENCE

AI Agent Intern, Dreamable Inc., San Francisco, CA *May 2025 – Aug 2025*

- Contributed with the team to **finetune a Qwen-2.5-7B-param** on Q&A tasks for the product trained on **lambda** and hosted on **Cloud Run (Google Cloud Platform)**.
- Led Dataset curation, used **Low Rank Adaptation** method from **transformers** library and evaluated model to achieve ~88% accuracy using **wandb**.
- Developed an AI-powered Outreach agent using **Langchain**, **Exa.ai** along with **OpenAI API** Integration to automate messaging workflows. Currently used by 14+ interns to scale weekly outreach with minimal effort.

PROJECTS

Llama-3.2-3b Finetune on OpenHermes Github | Huggingface | Dockerhub

- **Instruct-tuned a Llama-3.2-3B** model using **huggingface transformers** and **LoRA**. Packed into inference ready container on **Docker**, and served with **vLLM** (fast inference by factor of 3).
- Used techniques like **bf16** (equivalent to Quantization) + **Gradient checkpointing** (to save models) and **Flash Attention** (to make inference ~2.5-3x faster).
- Reduced valuation loss by ~68% from **1.27 to 0.21**, evaluated and tracked at **wandb**.

Qwen-2.5-0.5B Finetune Github | Huggingface

- Tested aligning a **Qwen-2.5-0.5B** model to act more like Human using **Direct Policy Optimization** after doing supervised Instruct-tuning using LoRA, as well as using **WandB** for model tracking.
- Achieved ~66% reward accuracy while keeping loss stable at ~1.560 on about 85M tokens. Served using **vLLM**.
- Used techniques like **bf16**, **gradient checkpointing** and **tf32** (Increases GPU usability by factor of 10) calculations.

GatorGPT Github | Huggingface

- Engineered a **63M Param** model using modern techniques like **Grouped Query Attention**, **Rotary positional Encodings** and **SwiGLU** MLP layers trained on TinyStories stories dataset. Served using **vLLM**, and is available on **Huggingface** to use on one go!
- To be finetuned on University specific data and to be tailored for University students in future using techniques like DPO and Reinforcement learning after a round of Supervised finetuning.

theHelper - AI Research Assistant Github

- Engineered a **RAG** based PDF analysis tool using **PyPDF2**, **BERT** transformers, and **FAISS for semantic search**, packaged in a **Streamlit** app for real-time summarization and Q&A — reduced manual review time by **70% across 50+ academic and business documents**; actively used by peers and family for coursework and client work.

And more on Github.