**Kunj P. Shah**

kunjcr2@gmail.com, +1(628)-529-6990

[LinkedIn](http://www.linkedin.com/in/kunjcr2) | [Github](https://github.com/kunjcr2) | [Portfolio](https://kunjcr2.github.io/) | 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

Dreamable Inc. San Francisco, California

*AI Agent Intern June 2025 – Aug 2025*

* Contributed with the team to **Finetune a Qwen-2.5-7B-param** on Q&A tasks for the product. Handled Dataset curation, used **LoRA** from **transformers** library and evaluated model to achieve ~88% accuracy using **wandb**.
* Developed an AI-powered Outreach agent using **LangGraph**, **Exa.ai** along with **OpenAI API** Integration to automate messaging workflows. **Currently used by 14+ interns** to scale weekly outreach with minimal effort.

Dyna Grow Design Solution Ahmedabad, India

*Web Developer Intern May 2024 – Jan 2025*

* Designed and launched a responsive marketing website using **ReactJS, TailwindCSS, Node**, **Express**, and **MongoDB**, tailored for an architecture firm’s client showcase and service catalog.

PROJECTS

Llama Finetuning on OpenHermes [GITHUB](https://github.com/kunjcr2/llama-3.2-3b-openhermes) | [HUGGINGFACE](https://huggingface.co/kunjcr2/llama3-3b-lora-openhermes) | [DOCKERHUB](https://hub.docker.com/r/kunjcr2/llama-3.2-3b-vllm)

* **Finetuned Llama-3.2-3B** model using **SFTtrainer**, **LoRA**, **PyTorch**, deployed inference-ready container with **Docker**, and served using **vLLM**. Trained on ~300K OpenHermes instruction-response pairs, achieving ~68% reduction in validation loss within 4.5 hours of A100.
* Used techniques like **bf16** + **Gradient checkpointing** and **Flash Attention** to make inference faster.

Qwen-2.5-0.5B Finetune [GITHUB](https://github.com/kunjcr2/how-llms-are-made/blob/main/docs/ml-and-dl/Reasoning%20Models/RLHF/DirectPreferenceOptimization.py) | [HUGGINGFACE](https://huggingface.co/kunjcr2/qwen2.5-0.5b-sft-dpo) | [DOCKERHUB](https://hub.docker.com/r/kunjcr2/llama-3.2-3b-vllm)

* **Finetuned Llama-3.2-3B** model using **SFTtrainer**, **LoRA**, **PyTorch**, deployed inference-ready container with **Docker**, and served using **vLLM**. Trained on ~300K OpenHermes instruction-response pairs, achieving ~68% reduction in validation loss within 4.5 hours of A100.
* Used techniques like **bf16** + **Gradient checkpointing** and **Flash Attention** to make inference faster.

GatorGPT [GITHUB](https://github.com/kunjcr2/GatorGPT) | [HUGGINGFACE](https://huggingface.co/kunjcr2/GatorGPT2)

* Pretrained a 63M–parameter Grouped Query Attention model with Flash Attention + Rotary Positional Encoding on ~1.5M stories (~350M tokens) using A100 GPUs (bf16 + tf32); achieved ~99% reduction in eval loss (246 → 1.50). Built and deployed inference-ready Docker image with vLLM: kunjcr2/GatorGPT2.

theHelper - AI Research Assistant [GITHUB](https://github.com/kunjcr2/AIResearchAssistant)

* Engineered a 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.

Additional projects available at: [GITHUB](https://github.com/kunjcr2).

SKILLS

AI/ML Large Language Models, Transformers, Retrieval-Augmented Generation (RAG), LoRA, PyTorch, TensorFlow, LangChain, LangFlow, n8n, OpenCV, Machine Learning, Deep Learning, Natural Language Processing (NLP), Weights and Biases

Web & Database Node.js, Express.js, React.js, Flask, Tailwindcss

Database and Tools Git, Docker, vLLM, VertexAI, MongoDB, MySQL