**Kunj P. Shah**

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

AI Agent Intern | LLM Developer | ML Researcher

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

**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)

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

**Database and Tools** MongoDB, MySQL, Git, Docker, VertexAI, Microsoft Azure

**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 – Present*

* Contributed to **fine-tuning a 7B-parameter open-source LLM** for internal document Q&A tasks; handled dataset curation, low-rank adaptation (LoRA), and model evaluation, achieving **~88% accuracy** on company-specific prompts.
* Developed an **AI-powered outreach assistant** using *n8n*, LangChain, and OpenAI tools to automate messaging workflows; currently used by **14+ interns** to scale weekly outreach with minimal manual effort.
* Built a **lead generation pipeline** that verifies and ranks potential clients by email validity and interest score using custom agents, improving lead quality and boosting response rate by **~2.3×**.

**Dyna Grow Design Solution Ahmedabad, India**

*Web Developer Intern May 2024 – Jan 2025*

* Designed and launched a **responsive marketing website** using Node.js, Express.js, and EJS, tailored for an architecture firm’s client showcase and service catalog.
* Improved **website performance**, leading to a **2× increase in qualified client inquiries** within the first 2 months of deployment.

**PROJECTS**

**Custom LLM - KsM** [github](https://github.com/kunjcr2/how-llms-are-made)

* Built a custom **215M-parameter** GPT-style language model with **18** transformer blocks and a **512**-token context window, trained on 5 novels using a self-implemented tokenizer, attention mechanism, and training loop, achieving over **85% accuracy** on internal benchmarks and **MAE loss of 1.8**.

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

* Engineered a PDF analysis tool using PyPDF2, BERT/BART 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.

**LLM Domain Adaptation POC** [github](https://github.com/kunjcr2/how-llms-are-made/tree/main/My%20Models/Finetuned_flan_t5) | [huggingface](https://huggingface.co/kunjcr2/stackoverflow-flan-finetune)

* Fine-tuned Google’s Flan-T5-Base (**220M-parameter**) on **15 K StackOverflow** Q&A pairs over **300 steps** using PyTorch Lightning, Hugging Face Transformers & LoRA adapters—reduced **eval loss** by **~30 %** (**2.36** → **1.66**) and sustained **17.7 samples/sec** throughput.

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