

Kunj P. Shah

kunjcr2@gmail.com, (628)-529-6990
AI Agent Intern | LLM Developer | ML Researcher
[LinkedIn](#) | [Github](#) | [Portfolio](#) | 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	Git, Docker, vLLM, VertexAI, MongoDB, MySQL

EDUCATION

San Francisco State University	San Francisco, California
B.S. in Computer Science	
<ul style="list-style-type: none">GPA: 3.96/4.00, Dean's List	Expected Graduation 2027

EXPERIENCE

Dreamable Inc.	San Francisco, California
AI Agent Intern	June 2025 – Aug 2025
<ul style="list-style-type: none">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 <i>n8n</i>, 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
<ul style="list-style-type: none">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

Llama Finetuning on OpenHermes	GITHUB HUGGINGFACE
<ul style="list-style-type: none">Fine-tuned Meta's Llama-3.2-3B (3.2B parameters) on ~300K OpenHermes instruction-response pairs using HF Transformers, LoRA (24.3M trainable params \approx 0.75 %), and A100 GPUs; achieved ~68% reduction in training loss (1.27 \rightarrow 0.20) within 2K steps (~4.5 h) with bf16 + gradient checkpointing. Deployed inference-ready Docker image with vLLM: kunjcr2/llama-3.2-3b-vllm.	
GatorGPT	GITHUB HUGGINGFACE
<ul style="list-style-type: none">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 \rightarrow 1.50). Built and deployed inference-ready Docker image with vLLM: kunjcr2/GatorGPT2.	
theHelper - AI Research Assistant	GITHUB
<ul style="list-style-type: none">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.	

Additional projects available at: [GITHUB](#).