

# 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	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 – Aug 2025
<ul style="list-style-type: none"><li>Contributed to <b>fine-tuning a 7B-parameter open-source LLM</b> for internal document Q&amp;A tasks; handled dataset curation, low-rank adaptation (LoRA), and model evaluation, achieving <b>~88% accuracy</b> on company-specific prompts.</li><li>Developed an <b>AI-powered outreach assistant</b> using <i>n8n</i>, LangChain, and OpenAI tools to automate messaging workflows; currently used by <b>14+ interns</b> to scale weekly outreach with minimal manual effort.</li><li>Built a <b>lead generation pipeline</b> that verifies and ranks potential clients by email validity and interest score using custom agents, improving lead quality and boosting response rate by <b>~2.3×</b>.</li></ul>	
Dyna Grow Design Solution	Ahmedabad, India
Web Developer Intern	May 2024 – Jan 2025
<ul style="list-style-type: none"><li>Designed and launched a <b>responsive marketing website</b> using Node.js, Express.js, and EJS, tailored for an architecture firm's client showcase and service catalog.</li><li>Improved <b>website performance</b>, leading to a <b>2× increase in qualified client inquiries</b> within the first 2 months of deployment.</li></ul>	

## PROJECTS

Llama Finetuning on OpenHermes	<a href="#">GITHUB</a>   <a href="#">HUGGINGFACE</a>
<ul style="list-style-type: none"><li>Fine-tuned Meta's Llama-3.2-3B (3.2B parameters) on <b>~300K</b> OpenHermes instruction-response pairs using HF Transformers, LoRA (<b>24.3M trainable params <math>\approx</math> 0.75 %</b>), and A100 GPUs; achieved <b>~68%</b> reduction in training loss (<b>1.27 <math>\rightarrow</math> 0.20</b>) within <b>2K steps (~4.5 h)</b> with bf16 + gradient checkpointing. Deployed inference-ready Docker image with vLLM: <b>kunjcr2/llama-3.2-3b-openhermes</b>.</li></ul>	
Custom LLM - KsM	<a href="#">GITHUB</a>
<ul style="list-style-type: none"><li>Built a custom <b>215M-parameter</b> GPT-style language model with <b>18</b> transformer blocks and a <b>512-token</b> context window, trained on 5 novels using a self-implemented tokenizer, attention mechanism, and training loop, achieving over <b>85% accuracy</b> on internal benchmarks and <b>MAE loss of 1.8</b>.</li></ul>	
theHelper - AI Research Assistant	<a href="#">GITHUB</a>
<ul style="list-style-type: none"><li>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&amp;A — reduced manual review time by <b>70%</b> across 50+ academic and business documents; <b>actively used by peers and family</b> for coursework and client work.</li></ul>	

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