

Laksh Goyal

La Jolla, CA, USA | +1 (858) 298 8940 | lakshgoyal06@gmail.com | linkedin.com/in/lakshgoyal | github.com/lgoyal6

EDUCATION

University of California San Diego

B.S in Computer Science, minor in Cognitive Science and Business-Economics

San Diego, CA

Sep 2024 – Jun 2028

Relevant Coursework: Algorithms & Data Structures; Machine Learning; Software Engineering; Operating Systems

Achievements: 5x Hackathon Winner; World Robot Olympiad National Winner

EXPERIENCE

GMI Cloud (NVIDIA Reference Platform Cloud Partner)

AI Engineer Intern

Mountain View, CA

Jun 2026 – Present

- Built a model cascade that routes easy coding tasks to cheap models and escalates hard ones, scoring 94/100 on the EvalPlus benchmark, beating every single-model baseline at 47% lower cost; published by GMI Cloud
- Replaced a manual lead-research workflow with Mint (FastAPI, PostgreSQL, React) in 7 weeks: automated identity matching, an LLM prospect classifier kept behind human review, and a 16-tool agent over the database
- Delivered 51 qualified P0/P1 leads from a 367-person triage set across two ICML event cohorts

Qualcomm Institute – Engineers For Exploration

Software Engineer, Machine Learning Systems (Part-time)

San Diego, CA

Jan 2026 – Present

- Built Python pipelines that ingest and index multi-terabyte acoustic datasets for CNN and EfficientNet training, raising usable audio 15% through automated quality checks and cutting preprocessing time 10% via throughput tuning
- Studied why classifiers trained on clean, targeted recordings degrade on noisy field audio, comparing spectral and learned features (STFT, MFCC, neural embeddings) and measuring the gap with F1 and distribution-shift metrics

Inovacity

Applied Machine Learning Intern (Part-time)

San Diego, CA

Mar 2026 – Aug 2026

- Integrated MedSAM2 as a wound-segmentation backend with click/box prompts, automatic mask cleanup, and fallbacks when a model is unavailable, reaching Dice 0.96 / IoU 0.92 against expert-labeled ground truth
- Built an LLM agent (Anthropic SDK) that drafts billing codes and healing status into a clinical note that clinicians review before saving, served through REST/FHIR endpoints covered by 69 passing integration tests

Xyntopia – Taskyon

LLM Systems Engineer Intern

San Diego, CA

Jun 2025 – Oct 2025

- Owned the Python task-orchestration framework behind a VS Code chat extension, cutting task retries 35% by resuming failed runs from their last good state instead of restarting, deployed across three internal products
- Built an end-to-end encrypted proxy for LLM and third-party API traffic, with per-token usage tracking in Supabase

PROJECTS

taut / tautq – Reliable-UDP Transport & Distributed Job Queue | C++20, epoll, SWIM

 GitHub

- Built a networking library in C++20 that stays fast on unreliable links, cutting p99 round-trip latency $\sim 6\times$ below the OS default at 5% packet loss and $\sim 12\times$ at 20%, benchmarked against TCP and ENet
- Built a distributed job queue on it with no central coordinator, sustaining 800 jobs/s and full completion at 20% packet loss
- Chaos-tested on every commit (kills, partitions, packet loss, corrupted logs) to prove no job is lost or run twice

Winnow – Real-Time LLM Prompt Compression | PyTorch, Transformers, Modal, Deepgram

 GitHub

- Implemented two recent papers with no public code, AttentionRAG and TurboQuant, cutting LLM KV-cache memory $\sim 60\times$ while holding the fp16 baseline's F1 accuracy on the LongBench benchmark
- Built a merge engine over LLMingua-2 and LongLLMLingua that compresses prompts while preserving original word order, served behind a live Deepgram voice demo with a raw vs. compressed comparison proving fidelity

Pylon – Counter-UAS Sensor Mesh | Python, scikit-learn, Next.js, deck.gl

 GitHub

- Founding engineer at Pylon, admitted to the Bow Capital incubator after winning its Defense Hackathon
- Built an RF anomaly detector (IsolationForest over radio power spectra) that learns each site's ambient baseline and flags drone emitters that signature matching misses, running on two SDR backends
- Shipped a self-organizing UDP sensor mesh that comes online in $\sim 2s$, plus a deck.gl operator dashboard with 3D terrain

SitRep – Team Project Management App | Vanilla JS, Cloudflare Pages/D1, Playwright

 GitHub

- Led an 11-engineer team through a 10-week Agile SDLC (5 sprints with retrospectives) to ship a deployed Scrum/Kanban/XP team-management app on Cloudflare Pages over a D1 database with 14 sequential migrations
- Ran pair programming and weekly role rotation across frontend, backend, DB, and QA, gating every PR behind a 5-stage CI pipeline ending in Playwright browser tests covering login, invites, and Kanban

TECHNICAL SKILLS

Languages: Python, C/C++, TypeScript, JavaScript, SQL, HTML/CSS

Frameworks & Libraries: PyTorch, Transformers, scikit-learn, NumPy, pandas, LangGraph, React, Next.js, FastAPI

Developer Tools: Git, Docker, AWS, Modal, Supabase, PostgreSQL, Cloudflare D1, GitHub Actions, CMake, libFuzzer