

Loukik Naik

Software Engineer

San Francisco, CA, United States | [LinkedIn](#)

I'm a software engineer who enjoys building systems that ship and scale.

I've worked across backend, systems, and ML-adjacent infrastructure, building APIs, real-time services, and production pipelines. I like owning problems end-to-end and keeping systems simple and reliable.

I'm especially interested in:

- Backend & systems engineering
- ML / AI-powered products
- Distributed systems & infrastructure

MS in CS from UC San Diego. Always learning, building, and happy to connect.

EXPERIENCE

Software Engineer

Aug 2025 — Mar 2026

Eudia

Palo Alto, California, United States

- Owned the Mergers & Acquisitions backend service end-to-end, building and operating the core infrastructure for an agent-based AI system for large-scale legal document analysis in production.
- Partnered closely with lawyers and product managers to design and ship product features, translating legal workflows and requirements into scalable platform capabilities.
- Designed and shipped backend APIs serving as the control plane for AI agent execution, managing document ingestion, run orchestration, result persistence, and retrieval of structured and tabular outputs.
- Re-architected agent orchestration by migrating from Airflow to Temporal, enabling reliable execution of long-running, stateful agents with retries, idempotency, and deterministic behavior.
- Built backend capabilities for multiple AI agents, including flexible document analysis, configurable PII redaction, Contract Comparison, and tabular review with schema-based inference and cell-level operations, backed by structured logging, workflow tracing, telemetry, and Kubernetes-native monitoring.
- Built workflows coordinating multiple services, queues, and workers, addressing production challenges including duplicate execution, inconsistent state across retries, and downstream dependency failures.
- Designed and managed Postgres database schemas and migrations using Alembic, delivering backward-compatible schema changes to support new agent capabilities.
- Converted the service to a multi-tenant architecture, enforcing tenant-aware routing, strict isolation, and safe rollouts across live customers.
- Managed the rollout of new backend features across staging, internal, and production environments using Kubernetes, Helm, and ArgoCD, supporting multiple customer deployments.

Contributor

May 2025 — Aug 2025

OpenFilter

United States

OpenFilter is a universal abstraction for building and running computer vision workloads

Machine Learning Engineer

Jan 2025 — Aug 2025

Plainsight Technologies

San Francisco, California, United States

- Developed and deployed real-time image segmentation APIs using FastAPI for vision models including SAM and MobileSAM, enabling seamless integration into internal applications.
- Containerized services with Docker and automated infrastructure deployment via Terraform and CI/CD pipelines, ensuring scalable, consistent rollouts.
- Designed a model caching system using symlinks and config-driven scripts, reducing friction during builds and cutting cloud runtimes by avoiding redundant model downloads at runtime.
- Created a Cron-triggered retraining pipeline that detects newly labeled data and launches Vertex AI training jobs, enabling continuous model updates.
- Designed an OCR model evaluation framework utilizing edit distance, confidence scoring, and text similarity metrics to automate quality checks pre-deployment.
- Designed and implemented a robust integration testing framework to validate end-to-end performance of the model generation pipelines for segmentation, object and keypoint detection tasks.

Machine Learning Intern - Computer Vision

Jun 2024 — Dec 2024

Plainsight Technologies

San Diego, California, United States

- Implemented serverless functions on Google Cloud Functions for real-time video processing.
- Developed and orchestrated Kubeflow pipelines on GKE for video preprocessing and filter evaluation, triggered by GCS data ingestion.
- Created an environment calibration model for deriving environmental statistics from video streams.
- Automated testing and deployment using GitHub Actions, PyPi, and Docker for production readiness.
- Ensured reliability through end-to-end tests and continuous integration.

Graduate Research Assistant

May 2024 — Jun 2024

UC San Diego

AI Developer

Lab Systems (I) Pvt. Ltd.

Feb 2023 — May 2023
Mumbai, Maharashtra, India

- Developed video forensics tool mvp for law enforcement agencies to analyse CCTV footage efficiently.
- Researched different techniques of object and text recognition to improve performance of the software.
- Leveraged Python libraries like pytesseract, easyocr, yolo, mediapipe, etc for object and text recognition

Software Engineer

Hexaview Technologies Inc.

Jul 2022 — Aug 2022
Mumbai, Maharashtra, India

EDUCATION

UC San Diego

Sep 2023 — Dec 2024

Master of Science - MS in Computer Science

Specialization: Machine Learning

Relevant Coursework:

Fall 2023:

CSE 202 - Design and Analysis Of Algorithms

CSE 250A - Probability and Reasoning

CSE 258 - Recommender Systems and Web Mining

Winter 2024:

CSE210 - Software Engineering

CSE251U - Unsupervised Learning

CSE251A - ML: Learning Algorithms

Spring 2024:

CSE256 - Natural Language Processing

CSE224 - Graduate Network System

COGS209 - Data Analytics/Stats Learning

Fall 2024:

CSE291 - Advanced Data Driven Text Mining

University of Mumbai

Oct 2019 — May 2023

Bachelor's degree in Computer Engineering

Department Rank 1 - SLRTCE

Pace Junior Science College

Jan 2017 — Jan 2019

PROJECTS

Verdict - Synthetic Market Research Agent

- Built a synthetic focus-group engine that runs 200–2,000 Claude personas against a product, returns a 1 - 5 purchase-intent PMF via structured output, clusters qualitative themes, and drops an Opus 4.7 executive brief - FastAPI, Sonnet 4.6, SQLAlchemy, React + TypeScript end-to-end.
- Designed a Temporal-shaped async workflow with ContextVar-based token attribution, so concurrent simulations in one process correctly bill each Claude call back to the right sim_id without threading IDs through every callsite.
- Replaced an embedder-based SSR scorer (which clustered everything into a 3.0–3.6 band regardless of sentiment) with an LLM scoring path: Sonnet 4.6 reads each multi-turn interview, returns the full intent distribution, and grounds personas in live market data via the web_search tool.
- Shipped reusable persona libraries, A/B variant comparison with side-by-side KPI deltas, per-call cost/latency telemetry (~\$1 per 10-profile sim), and a Vite + React UI auto-deployed to verdict.loukik.dev — inspired by the PyMC × Colgate 2025 LLMs Reproduce Purchase Intent paper.

Podclipper - Long Podcasts to Reels in a Click

- Built a local pipeline that turns long-form podcasts into vertical 9:16 reels — Whisper, Claude, YOLOv8, and OpenCV chained end-to-end.
- Designed a stage-based architecture with content-hash-keyed caching, so iterating on crop or prompt logic runs in seconds instead of re-transcribing.
- Solved active-speaker tracking across hard shot cuts via per-shot x-center clustering with look-ahead crop seeding and mouth-motion → diarized-turn linking.
- Shipped an LLM-as-judge quality gate, a React + Vite landing page, and GitHub Actions auto-deploy to podclipper.loukik.dev.

DraftIn - Keyboard-Driven LinkedIn Outreach Drafter

- Built a local-first Chrome MV3 + Node pipeline that drafts personalized LinkedIn outreach from a screenshot buffer - Alt+K stacks viewports across LinkedIn profiles and JD pages, Alt+L ships them through Oracle (browser-driven ChatGPT, no API key, no SaaS), and the result lands on the clipboard via pbcopy.
- Designed it screenshot-as-source-of-truth - no DOM parsing, no LLM tokens in the browser — and used activeTab + captureVisibleTab so the same buffer works on any page (career sites, founder pages, anywhere the user scrolls).

- Solved Chrome MV3's clipboard-after-focus-loss limitation by writing through the local server with pbcopy / xclip / clip instead of fighting the service-worker clipboard restriction from a toolbar gesture.
- Tuned Oracle browser-mode flags (--browser-min-stable-ms 15s, 10-min ceiling, 4-min recheck) so mid-stream pauses during image analysis and company web-search lookups don't trip premature capture; shipped a React + Vite landing page with GitHub Actions auto-deploy to draftin.loukik.dev.

3D Vehicle Visualizer (Digital Twin + API)

- Built a browser-based 3D vehicle viewer with a FastAPI backend that controls vehicle state (speed, brake lights, doors) and syncs it with a React + Three.js frontend.
- Designed a digital-twin architecture where the backend owns state + configuration, keeping the UI purely as a renderer.
- Solved real-world GLB inconsistencies via semantic part mapping, custom hinge/axis overrides, and per-model configuration stored via API.
- Implemented production-style UX patterns: optimistic updates with rollback, debounced writes, and polling for multi-client consistency.

Synapse – Decision-Based Spaced Repetition (iOS)

- Built an offline-first iOS app that trains engineering intuition using scenario-based flashcards focused on tradeoffs instead of memorization.
- Implemented a modified SM-2 spaced repetition engine with confidence-based scheduling and per-scenario difficulty tracking.
- Designed a JSON-based import pipeline to convert LLM-generated content into structured study decks with validation and mapping to SwiftData models.
- Architected clean separation between DTOs and persistence models to handle complex object graphs and ensure scalable content ingestion.

AutoReel - Multi Agent Reel Generator

- Built AutoReel, a side project that turns a single prompt into a ready-to-share 1-minute reel.
- Set up a multi-agent pipeline (LangGraph): one agent writes the script, another generates images, another adds TTS audio, and the last stitches everything into video.
- Used OpenAI models for script, voice, and visuals, and MoviePy for editing.
- Designed the workflow to run end-to-end automatically, from idea → script → video — with no manual editing.

Prompt Based Email Classifier

- Built a smart Gmail organizer that uses a custom prompt via LangChain to classify emails through either Ollama (local models) or the Gemini API (cloud) for real-time classification into categories like Work or Finance giving users full control over privacy vs. accuracy.
- Designed a fast, responsive system using multithreaded Python (Flask) and the Gmail API to fetch and batch-process emails, with a React frontend for live updates and intuitive filtering.

Surfstore: A Distributed File Storage System

- Built a Horizontally Scalable File Storage System in Go, storing data across multiple block and metadata servers.
- This project makes use of a .proto file to generate grpc code so that the client and server stub can use different programming languages and can belong to different devices.
- Implemented Raft consensus algorithm to sync changes across multiple metadata servers.

PPG Signal Based Hypertension Prediction

- Developed robust ETL pipeline to process the MIMIC-III dataset ensuring high data quality and integrity for downstream analytics.
- Achieved a test accuracy of 74% and a sensitivity of 91% on the MIMIC-III dataset.

TritonHTTP: A Server written in Go

- Implemented a Go-based HTTP server from scratch to handle multiple virtual hosts, each serving content from distinct document roots.
- Leveraged Go's concurrency model to handle multiple client connections simultaneously using goroutines.
- The server parses the incoming requests and, based on the specifications of TritonHTTP, generates and serves 200, 400, and 404 responses.

Active Amigo - Exercise Buddy App

- Our team developed an Android app that helps users find others with similar schedules for working out or participating in physical activities.
- The app uses Firebase Firestore as the backend and includes an end-to-end CI/CD setup with GitHub Actions.

Performance Analysis of Co-ordinate Descent vs Logistic Regression

- In this analysis study I performed a brief comparison of the performance of Logistic regression and 3 different varieties of co-ordinate descent i.e. Random Co-ordinate Descent, Iterative Co-ordinate Descent and Gradient based Co-ordinate Descent on the wine data set from sk-learn.
- Although, the co-ordinate descent failed to achieve log-loss values comparable to logistic regression, they converged faster and had reasonable loss and accuracy values to be considered for practical use.

Prototype Selection Technique For 1-Nearest Neighbour

- Compared and proposed different methods for prototype selection on huge datasets to summarize the data such that 1-NN model yields optimal test results.
- Achieved identical test accuracy of 96.7% using only about 11% of MNIST dataset after prototyping.

Anime Recommendation Website

Developed a website that suggests similar animes to a given anime entered by the user.
Implemented a cosine similarity algorithm to calculate the similarity between anime vectors.

CERTIFICATIONS

SnowConvert for Developers , Snowflake	Jan 2026
Generative AI: Working with Large Language Models , LinkedIn	Sep 2025
Develop a Synthetic Monitoring Platform , Build Fellowship by Open Avenues	Sep 2024
Software Engineer Intern , HackerRank	Mar 2024
Neural Networks and Deep Learning , Coursera	Jan 2022
SQL(Intermediate) , HackerRank	Jan 2022
Codechef Snackdown , CodeChef	Dec 2021
18th Rank Code Wars , Coding Ninjas	Mar 2021
Data Structures in Python , Coursera	Jun 2020

SKILLS

Agentic Workflows · STT · Whisper · FastAPI · Swift (Programming Language) · iOS Development · Mobile Application Development · Xcode · 3D Modeling · Video Processing · Software Infrastructure · Multi-agent Systems · Airflow · Temporal · LangGraph · TTS · Text-to-Image Generation · LangChain · LLaMA · Flask · Prompt Engineering · Multithreading · Routing Protocols · Large Language Models (LLM) · Google Kubernetes Engine (GKE) · Cloud Infrastructure · Kubeflow · Google Cloud Platform (GCP) · Docker · Computer Vision · Machine Learning · Data Visualization · Remote Procedure Call (RPC) · Concurrent Programming · Computer Networking · Go · Data Synchronization · HTTP · Problem Solving · Programming · Computer Science · Data Structures · Algorithms · Natural Language Processing (NLP) · Software Project Management · Data Preparation · Data Analytics · Data Science · Statistics · Model Evaluation