

Abhishek Desai

(408) 981 4724 | abhijaydesai@gmail.com | GitHub | LinkedIn | [X](#) | [Personal Site](#)

EDUCATION

University of Southern California

Jan 2026 – Present

Candidate for Masters of Science in Computer Science – General Track, 3.9 GPA

Pennsylvania State University

Aug 2021 – May 2025

Bachelors of Science in Computer Science, 3.8 GPA

SKILLS

Technical Languages/Frameworks: C, Java, Python, JavaScript, Verilog, MIPS, React, SQL, PyTorch, Tensorflow, OpenCV, NumPy

Coursework: Deep Learning, Large Language Models, Natural Language Processing, Computer Vision, Operating Systems, AI, ML

EXPERIENCE

Findwell

Jul 2026 – Present

Founder / Solo Engineer

A messaging-first archive for social posts: text a link and Findwell transcribes, and structures the video; text for the post and it returns it

- Built hybrid search in Postgres (BM25 + pgvector + reciprocal rank fusion, LLM query rewrite, Cohere Rerank v4), raising MRR@3 from 0.69 to 0.99 on a hand-labeled 46-query golden set
- Designed a tiered enrichment pipeline in which each tier is independently retryable: Whisper transcripts, LLM tagging and categories, and structured extraction from video keyframes that records the source of every field
- Built the backend (phone-number auth with Twilio Verify, private resume links, account deletion and export), an iMessage agent on Photon that saves and answers searches by text, a React Native app now pending App Store review, and the marketing site

GBOX.ai

Dec 2025 – April 2026

Software Engineer

San Francisco, CA

- Built a model-agnostic agent automation and benchmarking framework with support for multiple cross-environment adapters
- Contributed to cross-platform agent evaluation environments (mobile, web, desktop), including task definitions, orchestration, and tool integration
- Designed a Windows agent environment with vision-based grounding to translate UI observations into executable actions for OS-level benchmarking within a unified pipeline

Infineon

Jun 2025 – Dec 2025

Embedded Machine Learning Intern

San Jose, CA

- Built and optimized a Wi-Fi-based presence-sensing pipeline on PSoC MCU hardware by leveraging channel state information (CSI) features to detect subtle human motion
- Designed and trained a radar-based gesture recognition system on embedded hardware, extracting micro-motion features (range–Doppler, phase changes) and experimenting with time-series ML models to enable reliable, low-latency hand and body gesture detection
- Built task-driven embedded applications with FreeRTOS, leveraging semaphores, queues, and schedulers for efficient real-time task management and inter-task communication, resulting in improved system responsiveness and reliability

Synaptics

May 2024 – Aug 2024

Computer Vision/Deep Learning Intern

San Jose, CA

- Created a cloud-based automated edge AI model deployment pipeline using zero-shot models, improving deployment efficiency and reducing manual intervention
- Developed an automated cloud-based human presence detection pipeline using zero-shot models and a teacher-student network for edge deployment, enhancing detection accuracy and deployment speed
- Conducted in-depth research on state-of-the-art zero-shot models including GroundingDINO, YOLO-World, and Detr, contributing to the selection of optimal models for deployment
- Generated object detection annotations and trained a smaller model optimized for Synaptics chips, resulting in improved model performance and efficiency on proprietary hardware
- Leveraged Facebook's SAM to create diverse datasets by overlaying segmentations on random backgrounds and correcting false negatives from Grounding DINO, enhancing dataset quality and model accuracy

Penn State Computational Biology Lab

Aug 2023 – Dec 2023

Research Assistant

University Park, PA

- Collaborated with Professor Stefan Canzar to optimize sequence alignment between long-read and short-read RNA barcodes, enhancing the accuracy and efficiency of genetic analysis
- Implemented Levenshtein automaton to form string groupings based within certain threshold edit distance which ultimately can reduce runtime complexity from $O(MN)$ to linear time
- Utilized an automaton to form 'buckets' of similar strings, avoiding computationally expensive full pairwise string comparisons and enhancing processing efficiency
- Used an AVL Tree to implement a least sensitive bucketing algorithm with (3, 5) sensitivity by enumerating strings

PROJECTS

CLI Based Agentic Video Editor | LangGraph, Gemini, OpenAI, Twelve Labs

- Developed agentic workflows using LangGraph and LLMs to automate repetitive editing tasks such as timeline assembly, clip trimming and scene detection
- Integrated video understanding models for semantic video search, enabling editors to query and arrange footage using natural language

Agent Gary (Origin Weekend x Replit 2026 Finalist) | Twilio, Deepgram, Jev, MCP, Playwright, Supabase

- Led a four-person team as PM and lead engineer through a two-day sprint to build a phone-based voice agent that lets older adults order food, book rides and home services, and pay bills by calling a regular number; a CUA completes each task (no API access)
- Designed the family safety loop: a policy layer holds unfamiliar payees, new destinations, and scam-pattern requests, texts the caregiver for a YES/NO, and reports the outcome into the live call

