

SAKSHAM RAI

(408) 513-5080 | raisaksham2001@gmail.com | Portfolio -  | LinkedIn -  | GitHub - 

Education

University of California San Diego | La Jolla, California

October 2020 – June 2024

- **Degree:** BS. Mathematics-Computer Science, Double Minor – Cognitive Science (ML Focus) and Entrepreneurship
- **Academic Achievements:** Departmental Provost Honors (x6), Sophomores Honors Program (x1), Blackstone Launchpad.
- **Associations:** CSForEach, Engineers for Exploration, UCSD Speech and Debate, The Basement Entrepreneurs Program and UCSD Cricket Team.
- **Relevant Coursework:** Web-Client Technologies, Software Engineering, Algorithms and Data Structures, Object-Oriented Design, Computation, Systems Programming, Unsupervised Machine Learning, AI Algorithms, Product Innovation and Venture Finance
- **Coursework GPA:** 3.7

Skills

Languages: Python, JavaScript, Java, C, Typescript, PHP, CSS/HTML

Frameworks/Databases: React, Next.js, Vue.js, Flask, MySQL, PostgreSQL, MongoDB, Firebase

Tools: AWS, Postman, Docker, Render, Vercel, Git/Github

Experience

Backend Software Engineer | Gabriel AI | AWS, Firebase, Python, Telnix, Typescript | *April 2025 – Present*

- Implemented Firebase Authentication with session persistence and fixed routing bugs on signup/login flows, **reducing user drop-offs by 40% on mobile.**
- Integrated Telnix messaging and call webhooks; **resolved backend mismatches in campaign replies and standardised analytics timestamps** using server-side UTC handling.
- Deployed backend services via AWS Lambda and EC2; enabled CloudWatch logging and IAM-based billing access **to monitor \$400+ monthly AWS costs.**

Full-Stack Engineer | County of SD & LA | PostgreSQL, MySQL, NodeJS, React, PHP | *Sept 2024 – Present*

- Engineered a real-time drug retrieval portal, **parsing & ingesting 5,000 + county records** into PostgreSQL, **reducing query latency by 20%:** [Link](#)
- Engineered real-time opioid distribution portals for **13,000+ citizens via 21 vending machines**, integrating Google Maps API for accurate location services: [Link](#)
- Automated inventory tracking and backend synchronization using Google Sheets API, CRON jobs, and MySQL/PHP scripts to **ensure real-time updates.**

AI Software Engineer Intern | AfyaChat | Flask, Python, FastAPI, Twilio, GPT 3.5, Llama Index | *March 2024 – April 2024*

- Built an AI-powered eConsultation system, **reducing specialist response times by 50%** using GPT-3.5 & Flask.
- Built a Retrieval-Augmented Generation (RAG) pipeline, optimizing **7+ years of historical medical data** for real-time AI-based specialist recommendations.
- Engineered microservices architecture using FastAPI, ensuring **scalable AI model execution with low-latency inference.**
- Integrated real-time physician notifications via Twilio APIs, **reducing average patient response time by 35%.**

Impact-Driven Projects

Talk-to-Abel | Next.js, Firebase, WebSockets, ONNX, Whisper, VAD | [Link](#)

August 2024

- Built a **low-latency real-time AI voice agent** handling **6,000+ sessions** via WebSockets, Whisper + VAD, and GPT-4 streaming; deployed on Dockerized Next.js with Firebase auth/session handling.

GymScout | MongoDB, Redux, Redis, Typescript, React, Node.js | [Link](#)

Feb 2025

- Built a gym discovery app using geospatial queries, Redis caching, and MongoDB; **reduced API latency by**

30% via real-time data pipelines and Redux-backed state.