# SAKSHAM RAI

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

## Experience

Machine Learning Engineer | ***CreateBase*** | Python, HuggingFace, Supabase, LLMs | *May 2025 – Present*

- Building NLP pipelines for matching **10M+ messy royalty records**, boosting **metadata resolution accuracy by 65%.**
- **Fine-tuning LLMs for IP** attribution and integrated them into real-time pipelines **with sub-100ms latency.**
- Engineering vector search with Pinecone + sentence embeddings, **reducing manual rights matching by 40%.**
- Automating DSP metadata ingestion and royalty prediction workflows, **recovering $500K+ in missed payouts**.

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

- Designed conditional onboarding using Firebase user metadata and tooltip flows, routing new users to billing and campaign setup; **increased activation by 55%**
- Integrated Telnyx 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**.

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-4 & Flask**.
- Built a Retrieval-Augmented Generation (RAG) pipeline, optimizing **7+ years of historical medical data** for real-time AI-based specialist recommendations.
- Engineered backend 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%.**

Full-Stack Engineer | ***County of SD & LA*** | PostgreSQL, MySQL, NodeJS, React, PHP | *Sept 2022 – April 2025*

- 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 synchronisation** using Google Sheets API, CRON jobs, and MySQL/PHP scripts to ensure real-time updates.

## Education

**University of California San Diego | *La Jolla,* *California*** *October 2020 – June 2024*

- **Degree:** BS. Math-CS, 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.
- **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/HTM
- **Frameworks/Databases:** React, Next.js, Vue.js, Flask, MySQL, PostgreSQL, MongoDB, Firebase
- **Tools:** AWS, Postman, Docker, Render, Vercel, Git/Github