

AARYA SRINIVASAN

San Jose, CA | 669-204-6525 | aaryasrinivasan@berkeley.edu | www.linkedin.com/in/aarya-srinivasan

EDUCATION

University of California, Berkeley

Graduating May 2028

BS Computer Science | BA Data Science

- **Relevant Coursework:** Data Structures, Structure of Computer Programs, Linear Algebra & DEs, Principles and Techniques of Data Science, Data Science Foundations, Introduction to Artificial Intelligence

PROFESSIONAL EXPERIENCE

NextXI | AI Cricket Coaching Platform

June 2026 – Present

AI/ML Engineer

San Francisco, CA

- Trained custom **YOLO detectors** alongside a pose estimation model to track body keypoints, then wrote an OpenCV post processing pipeline using Kalman filtering to stabilize output into clean joint trajectories and derive coaching metrics
- Shipped a production (Next.js, TypeScript, Prisma) scouting platform with role-based access and co-located server actions connected to the computer vision service via a **Postgres-as-a-queue worker** with idempotent job handling
- Served multiple cricket academies with **250+ players and 10+ coaches** actively running video analysis on the platform

University of California, Berkeley

Jan – May 2026

Data Science Research Assistant

Berkeley, CA

- Processed and interpreted raw 10-K data to construct comprehensive corporate restructuring and reorganization datasets
- Implemented data validation protocols to resolve noise, ensuring integrity for downstream econometric and ML modeling
- Programmed a **pipeline using Hugging Face and LLMs** to automate the scraping and processing of SEC filing data

AshaDx | Maternal Health Diagnostics

Jan – May 2026

Software Engineering Intern

San Jose, CA

- Researched ML architectures and strategies to develop an AI pipeline to automate de novo antibody design for clinicians
- Designed and built the company's database in Supabase (PostgreSQL), structuring schemas to consolidate contact data
- Automated site contact points with Make webhooks, routing form submissions to emails and eliminating manual lead triage

Sync Visual Media

June – Aug 2025

Software Engineering Intern

Phoenix, AZ

- Developed full-stack web applications for client platforms using React and Python-based APIs, optimizing user flows, UI/UX, and backend performance to increase visitor-to-customer **conversion rates by ~15%** per client
- Engineered **Search Engine Optimization** rendering pipelines (metadata tagging, structured schema, image indexing) to optimize 100+ digital assets for clients, boosting search ranking visibility across Google Image and web results

PROJECTS

Sid | Deterministic LLM Inference Engine

July 2026 – Present

Python, PyTorch, CUDA, Triton, FlashAttention-3, SGLang

- Built a vLLM-style serving engine with a continuous-batching scheduler, paged KV cache with prefix caching and a Triton scatter kernel, tensor parallelism over NCCL, and CUDA-graph capture with FlashAttention-3
- Designed a deterministic decoding mode that replays tokens through fixed-shape windowed verification with rollback, guaranteeing batch-invariant outputs without batch-invariant kernels
- Added a sensitivity router to send flagged requests to the deterministic path, keeping full throughput for standard traffic

LLMBDA | Inference Optimization Tool

Nov 2025 – June 2026

Python, PyTorch, Hugging Face, LLMingua, FastAPI, TypeScript, React

- Designed a novel inference-optimization architecture that **cuts costs by over 80%** with no measurable Ragas quality loss
- Engineered a joint scoring RL policy formulated as an offline contextual bandit that selects the optimal (compression × refinement × target model) configuration per request under a quality constraint to choose best path at scale
- Fine-tuned an LLMingua token classifier for task-aware prompt compression, retaining **99% accuracy on CRAG**

Coca-Cola Consumer Sentiment Forecasting (Contract)

Jan – May 2024

Python, SERP API, VADER, SQL, TextBlob, Pandas

- Designed a web-scraping pipeline to extract and preprocess consumer sentiment data from SERP API, Twitter, and Reddit
- Developed sentiment analysis models using **VADER NLP** and LDA Topic Modeling, optimizing forecasting **by 43%**
- Engineered and integrated sentiment-based forecasting that quantified consumer sentiment and surface market trends, using AWS SQS, PostgreSQL, and streaming queues to power real-time supply chain and marketing decisions for Coca-Cola

SKILLS & AWARDS

Languages: Python (asyncio, FastAPI), C/C++, SQL, R, Java, JavaScript/TypeScript (Node, React, Next.js, Prisma)

AI/ML: PyTorch, Transformers, Hugging Face, CUDA, Triton, FlashAttention, vLLM/SGLang, RAG, YOLO/OpenCV, scikit-learn

Infra & Tools: Docker, AWS (SQS), PostgreSQL/Supabase, Git, Linux

Awards: Debate *Tournament of Champions Winner*, National Debate Tournament **Champion**, National Economics Challenge **4th**