Krish Arora

Berkeley, CA — karora27@berkeley.edu — linkedin.com/in/karora27 — (949) 350–2698

Education

University of California, Berkeley **Dec 2027**

Bachelor of Science in Data Science (Applied Math & Modeling Emphasis)

Relevant Coursework: Data Structures, Algorithms, Discrete Math, Machine Learning Algorithms, OS

Awards: AWS QBS Hackathon: 1st Place Overall, PwC Case Competition: 1st Place, Bio-Rad STEM Scholar

Technical Skills

Languages: Python, Java, C++, Go, TypeScript, JavaScript, React, SQL, Perl, Rust

Frameworks/Tools: Dagger, AWS (Lambda, S3, ESM, DynamoDB, Kinesis, SQS, SNS), TensorFlow, LangChain, Redis

Libraries: NumPy, Pandas, Seaborn, Scikit-learn, Keras, XGBoost, Matplotlib, NLTK, Gensim, TensorLy, Hugging Face

Experience

Zipline **San Francisco, CA**

Software Engineer Intern *Sep 2026 – Present*

- Developing flight simulations in C++/Rust along with a web tool to validate failsafe behaviors through Monte Carlo runs

Amazon Web Services (AWS) **Santa Clara, CA**

Software Development Engineer Intern *Jun 2026 – Aug 2026*

- Surfaced silent permission and preference changes for 300k+ users by designing an event-driven notification system on DynamoDB Streams and AWS Lambda; decoupled from write path so zero latency was added to live user-facing requests
- Designed a concurrency-safe, idempotent notification pipeline using deterministic SHA-256 notification IDs, preventing duplicate alerts across Lambda retries and parallel DynamoDB Stream shard processing without adding persistent state
- Reduced unnecessary Lambda invocations across the entire table's change stream by designing an ESM filter (SK prefix and scope condition) that drops irrelevant writes before the worker is ever invoked, with no producer code on write path

International Business Machines Corporation (IBM) **San Jose, CA**

Software Engineer Intern *Feb 2026 – May 2026*

- Cut incorrect tool invocations by 60% across a multi-agent Db2 assistant by designing a hierarchical LangGraph routing architecture that classified user intent and orchestrated 8 specialized agents across performance and support workflows
- Lowered unresolved documentation queries by 40% across 74 indexed sources by implementing hybrid vector search with Redis, HuggingFace all-MiniLM-L6-v2 embeddings, and TF-IDF cosine re-ranking to improve source matching precision

Amazon Web Services (AWS) **Sunnyvale, CA**

Software Development Engineer Intern *Jun 2025 – Aug 2025*

- Expanded the file upload pipeline to support mixed-media files. Reduced alert latency by 70% through an asynchronous polling architecture utilizing S3-triggered Lambda workers, API rate limiting, and real-time UI progress updates
- Architected the backend transcription workflow across S3, Lambda, and DynamoDB, using Step Functions to add retry/catch state handling and DynamoDB status tracking to monitor and recover failed audio/video transcription jobs
- Automated structured insight routing into downstream workflows for CSV/Excel and Google Sheets uploads by designing and deploying a document ingestion system integrated with Amazon Bedrock, eliminating manual processing overhead

Projects

Tensor Decomposition for Deep Neural Network Compression | *Lead Researcher*

- Achieved 87% deep neural network parameter compression with <0.015 average loss across VGG-16, ResNet-50, and custom CNNs by designing tensor-train decomposition algorithms for layer shape optimization
- Reduced compute efficiency by 50% for edge devices through recursive tensor reshaping and gradient-based optimization; first authored research paper and presented technical poster to 100+ Ph.D. candidates at UCSB

Settle | *Google x Qualcomm Developer Hackathon*

- Eliminated cloud dependency for document/contract analysis by deploying an on-device LLM pipeline using LiteRT-LM with Gemma 4 E2B and Hexagon NPU, achieving low-latency inference with GPU fallback for cross-device reliability
- Enabled real-time contract clause detection by designing a pixel-coordinate mapping system from bitmap captures to text blocks, with an LLM-based risk heuristic gating summary generation and multi-turn Q&A

Aurabase | *aura-base.com*

- Built a 250+ user platform allowing friend groups to log events and compete on a dynamic "aura" leaderboard
- Developed client-side dynamic pages with ReactJS, Tailwind CSS, Express.js, and RESTful APIs via Axios
- Implemented secure user authentication with 4 Supabase PostgreSQL databases and server-side validation