Avneesh Muralitharan

amuralitharan+jobsearch@ucsd.edu — (408) 332-9324 — github.com/amuralitharan

Education

University of California, San Diego **San Diego, CA**
Program: *Master of Science in Computer Science, Artificial Intelligence* *March 2026*
GPA: 3.54/4.0
- **Relevant Coursework:** Parallel Programming, Algorithm Design and Analysis, Web Mining and Recommender Systems, Deep Learning, Operating Systems, Web Mining and Recommender Systems, Probabilistic Learning, Data Mining and Predictive Analytics

University of California, Santa Cruz **Santa Cruz, CA**
Major: *Bachelor of Science in Computer Science* *June 2024*
GPA: 3.65/4.0
- **Honors/Awards:** Cum Laude, Dean's List, College Scholar's Program
- **Relevant Coursework:** Computer Architecture, Advanced Linear Algebra, Discrete Math, Calculus I/II/III, Probability Theory

Work Experience

Syneris **Sunnyvale, CA**
SWE Intern *June 2025 - September 2025*
Manages regulatory compliance for medtech companies by integrating quality management, product lifecycle, and manufacturing into a single platform, using AI to automate compliance-related paperwork and processes.
- Writing a RAG document retrieval pipeline for retrieving assays, regulatory documents, etc.

SMoL Lab - Professor Kristan Vaccaro **San Diego, CA**
Research Assistant *March 2025 - Present*
Social computing lab focused on content moderation. Writing a scraper that scrapes all actions any user performs on Bluesky (posts, reposts, likes, follows, blocks).
- Loads 800+ user-events (posts, reposts, likes, follows, blocks) a second from the Bluesky firehouse, a stream of all Bluesky activity, into a Clickhouse (SQL) Database. After running for $<$ 1 day stores 10 million like events, half a million posts, and half a million follow events.
- Can scrape user-activity from a historical point in time (eg. can start scraping from a week ago instead of only being real time)
- Clickhouse OLAP allows for identification of bot accounts in real time through looking at periodicity of user-events

Dell Technologies **San Jose, CA**
SWE Intern (Co-op) *January 2024 - June 2024*
American technology company that develops, sells, repairs, and supports computers and related products and services.
- Developed an assistant designed to recover from Dell PowerProtect system upgrade failures by using Llama3 and RAG over 10,000 server manuals.
- Wrote document-aware chunking script to parse and embed manuals based on markdown structure. This prevented partial workflow retrievals - a common issue where RAG systems only retrieve fragments of workflows or bug fixes rather than complete procedures.
- Brought down system downtime on Dell PowerProtect systems from upgrade failure from 3 days to 1 hour in test environments.

Neuromorphic Computing Group - Jason Eshraghian **Santa Cruz, CA**
Research Assistant *Jan 2024 - March 2024*
Research lab dedicated to neuromorphic hardware design and Spiking Neural Networks (SNNs).
- Worked on training a SNN on the Moving MNIST dataset, which involves 2 digits moving in a 64 x 64 frame. Goal was to predict the final 10 frames of a sequence based on the 10 initial frames using the cross-entropy loss metric.

Autonomous Vehicles Lab - Jim Whitehead **Santa Cruz, CA**
Research Assistant *September 2022 - December 2022*
Research project devoted to creating procedurally generated road networks for testing autonomous vehicles.
- Used Blender Python API to generate road networks and place objects like trees and procedurally generated buildings.
- Exported in the OpenDrive format, enabling import into the Carla simulation environment for autonomous vehicle testing.

Alveo Technologies **Alameda, CA**
SWE Intern [Co-op + Internship] *February 2021 - August 2021*
Startup providing rapid, portable and cloud-based coronavirus tests.
- Developed a React/Electron/Flask app to validate COVID-19 testing kits at the end of the manufacturing process.
- Designed a React/Electron/Flask app to run chemical assays remotely and plot results using plotly.
- COVID-19 test validation and remote assay software are currently deployed and actively used in the manufacturing of test kits.

Projects

High-Performance Matrix Multiplication Optimization *October 2024*
Implemented and optimized a BLAS-like matrix multiplication library targeting ARM architecture, achieving performance comparable to OpenBLAS for square matrices.
- Developed a multi-level cache blocking scheme targeting L1, L2, and L3 cache hierarchies.

- Optimized matrix operations using ARM SVE vector instructions and loop unrolling, achieving an 18+ GFLOPS performance peak across various matrix sizes (32 to 2049)
- Analyzed performance using hardware performance counters and cache profiling tools, conducting detailed studies of cache behavior and architectural impacts
- Implemented vectorized microkernel using SVE instructions to process entire matrix rows in single vector registers

High-Performance GPU Matrix Multiplication *October 2024 - November 2024*

Implemented and optimized a CUDA-based matrix multiplication kernel for NVIDIA's Turing GPU architecture.

- Achieved up to 1,690 GFlops/sec performance (54x speedup over CPU implementation) using warp-tiled matrix multiplication
- Implemented cache tiling, outer product algorithms, thread tiling, and coalesced memory access patterns
- Optimized register usage and thread block dimensions to maximize SM occupancy on the T4 GPU
- Used NVIDIA's NSight Compute (NCU) and NVPROF profiling tools to identify memory bottlenecks and kernel performance issues
- Analyzed performance across matrix sizes from 256 to 4096 using roofline modeling and comparative benchmarks

Parallel 2D Wave Simulation with MPI *December 2024*

Implemented a high-performance MPI-based simulation of 2D wave equations than ran on UCSD's Expanse Supercomputer

- Achieved over 7,300 Mpts/sec (66 GFlops/sec) performance when running on 256 cores using optimized process geometry.
- Implemented non-blocking MPI communication to hide memory latency, overlapping interior point calculations with boundary data exchange.
- Conducted extensive strong and weak scaling studies to determine optimal process geometry, achieving near-linear speedup.

AlphaZero Mancala Bot *March 2025 - Present*

Developed an AI to play Mancala using a combination of deep learning and tree search techniques.

- Implemented a feed-forward neural network with a Monte Carlo Tree Search (MCTS) policy gradient
- Trained the neural network through self-play and reinforcement learning using Proximal Policy Optimization (PPO)
- Designed and executed evaluation frameworks to test agent performance against baseline bots, random agents, and human players

Large-Scale Recommendation System for RateBeer *December 2024*

Developed a recommendation system for a dataset of 2.8 million beer reviews.

- Engineered a scalable hybrid recommendation architecture combining matrix factorization with feature-based models, achieving RMSE of 0.1360 (15.0% improvement over global baseline of 0.1599)
- Quantified feature importance through ablation studies: beer popularity (47.30%), user-item interactions (35.78%), user activity (13.33%), and style preferences (2.30%)

SlugEvents - Campus Event Aggregator *Spring 2023*

Developed a React-based web application that aggregates and displays college campus events from social media accounts.

- Built a full-stack React application that scrapes event data from Instagram accounts of different clubs
- Integrated Firebase Firestore database to store and retrieve event information with real-time updates
- Implemented a Python script with OpenAI API integration to automatically identify events from Instagram posts
- Created an interactive map with Google Maps API to display event locations on campus with custom markers and filtering

Contrastive Learning and Fine-Tuning Techniques for BERT *March 2025*

Fine tuned BERT on MASSIVE dataset (Multilingual Amazon Scenario Set for Intent and Slot Labeling).

- Worked with a comprehensive NLU dataset containing 1M+ annotated examples across 51 languages for intent classification (identifying the user's purpose across 18 categories like "get_weather" or "play_music") and slot annotation tasks (extracting specific information like locations, dates, or entities needed to fulfill the intent)
- Implemented Layer-wise Learning Rate Decay (LLRD) to fine tune BERT model with decay factor 0.9, reducing loss from 533.98 to 106.77 while improving accuracy from 2.7% to 92.0%
- Compared SimCLR (using dual dropout rates of 0.3 and 0.5) and SupCon approaches for contrastive learning, with SupCon achieving 90.82% accuracy
- Experimented with various LoRA configurations (r=4,8,16,32), reducing trainable parameters by 99.5% (from 109M to 600K at r=16) while maintaining 84.2% accuracy

Skills

Programming Languages: Fluent in C++, Python, Java, Javascript, HTML/CSS, RISC-V Assembly
Web and Cloud Technologies: Experienced in full-stack development with MERN (MongoDB, Express.js, React.js, Node.js) and application deployment via Firebase, AWS, GCP, and Docker containerization
Data Science and Machine Learning: Experience with SNNTorch, Pytorch, and Tensorflow. Proficient in ARM and CUDA optimization.