Sanjay Mukhyala

Foster City, CA 94404 | (650) 279-8516 | smukhyala@gmail.com | https://sanjay-mukhyala.com/

EDUCATION

University of California, Berkeley (Expected May 2028) | B.A. Data Science, Applied Mathematics | **GPA: 3.88**
Relevant Coursework: Data Structures and Algorithms, Machine Learning, Abstract Linear Algebra, Discrete Math, Linear Programming, Deep Learning, Model Optimization, Artificial Intelligence, Database Systems, Numerical Analysis

PROFESSIONAL EXPERIENCE

Fireworks AI | San Mateo, CA — May 2027 – August 2027
Incoming Applied Machine Learning Engineer Intern

- Joining the Applied ML team to enable new model stacks and optimize inference performance across production workloads

Berkeley Statistical Machine Learning Group | Berkeley, CA — March 2026 – Present
Undergraduate Researcher advised by Ian Waudby-Smith and Keegan Harris

- Built an AI agent evaluation framework using sequential testing and anytime-valid e-processes for fixed inference budgets
- Developed an automated early-stop controller, avoiding 71% of planned evaluations (170 runs) in a live benchmark trial
- Engineered scalable benchmarking pipelines across 4 web applications with adaptive allocation and inference cost tracking
- Analyzed 1,000+ prompt-evaluation settings, quantifying execution noise, selection bias, and agent performance reliability

Prologis | San Francisco, CA — May 2026 – August 2026
Machine Learning Intern

- Built enterprise RAG platform integrating Dataiku, Snowflake, and OpenAI API workflows for citation-backed querying
- Developed hybrid retrieval pipelines over financial data and 1,000+ unstructured documents across two retrieval engines
- Built a production API managing retrieval engines and custom actions, exposing unified search to downstream applications
- Consolidated fragmented enterprise knowledge into a scalable retrieval system, reducing information retrieval time by 85%

Zendesk | San Francisco, CA — February 2026 – May 2026
AI Engineer (Contract)

- Built and benchmarked ITSM agent architectures (OODA, F-CoT) for tool-using workflows under uncertain task conditions
- Designed evaluation frameworks across 100+ tickets measuring reasoning, hallucinations, tool call, and action accuracy
- Implemented evidence-based validation layers to detect unsupported reasoning and unsafe tool calls before agent execution
- Improved agent reliability to 95% and reduced hallucination and unsafe-tool-use failures by $\tilde{2}7\%$ across evaluated tasks

Breefly | Berkeley, CA — August 2025 – February 2026
Co-Founder & Head of Product

- Developed a multi-stage LLM analysis with sub 2-second transcript retrieval from long-form videos through indexing
- Reduced repeat-access latency ~95% by implementing Redis caching, Celery task orchestration, and WebSocket streams
- Raised $150K pre-seed investment ($3M valuation) from Character Capital after 1.5K+ total users and 5M+ impressions
- Containerized and deployed multiple backend services with Docker on AWS, scaling Celery workers and production APIs

PROJECTS

World-Paried Demonstrations (Winner at Berkeley Robotics Hackathon) | Berkeley, CA — August 2026

- Developed a world-paired learning framework combining egocentric data with exocentric world context for robot learning
- Increased held-out task-labeling accuracy by 6.2% by introducing exo-ego pairing on Assembly101 and WorldContext data
- Found ceiling-camera view provided no gain over ego-only labeling, showing exocentric benefit depends on viewpoint

MIRN | Berkeley, CA — July 2026 – August 2026

- Built a counterfactual crowd simulator measuring perturbation by replaying identical scenes with and without a robot
- Evaluated disturbance metrics against causal effects, exposing a 0.334 meter false signal when true influence was zero
- Designed paired simulations with shared seeds and initial conditions, with 1,400+ tests validating metric reproducibility

Cortex | Berkeley, CA — July 2026 – August 2026

- Designed an AI memory graph with 93% accurate retrieval with conflict resolution from multiple chat-agent histories
- Implemented J-Space (Anthropic) workspace distillation for durable memory retrieval via MCP and semantic search
- Built conflict-aware memory consolidation across multi-agent histories, preserving durable context despite contradiction

SKILLS & INTERESTS

Fellowship and Membership: .406 Ventures, Afore Capital, Entrepreneurs@Berkeley, AI Entrepreneurs at Berkeley
Skills: Python, Java, SQL, Git, Docker, Linux, PyTorch, TensorFlow, scikit-learn, NumPy, Pandas, FastAPI, PostgreSQL, pgvector, Redis, Celery, Snowflake, AWS, OpenAI API, MCP, REST APIs, RAG, embeddings, semantic search, LLM agents
Interests: Poker (A10s), Acoustic Guitar, Running (1:44 Half Marathon), Basketball (Pistons), Soccer (Chelsea), Math