Collin Hargreaves

815-878-7034 | collin_hargreaves@berkeley.edu | linkedin | github | hargreaves.sh

Education

University of California, Berkeley — Berkeley, CA
Bachelor of Arts in Data Science, Minor in Computer Science — *Expected Dec. 2027*

- **Relevant Coursework**: Data Structures & Algorithms, Machine Learning, Programming Languages & Compilers, Embedded Systems, Database Systems, Networking, Probability Theory

Experience

Machine Learning Engineer Intern — Sep. 2026 – Present
Git.M-Labs — *Remote*

- Adding SGLang as a second backend target to an autonomous inference optimization engine, lowering engine-agnostic optimization plans to both vLLM and SGLang

Software Development Engineer Intern (HyperPod Dataplane Team) — May 2026 – Aug. 2026
Amazon Web Services — *Santa Clara, CA*

- Built Go and TypeScript components for a package distribution system that updates HyperPod agents on active Slurm and EKS clusters without rebuilding nodes or interrupting long-running training jobs
- Implemented secure cross-account package access using IAM roles, S3 bucket policies, and VPC endpoints
- Extended the HyperPod host agent to configure package sources at runtime, including support for SigV4-authenticated S3 repositories

Undergraduate Researcher — Jan. 2026 – Aug. 2026
Lawrence Berkeley National Lab, Quantitative Ecosystem Dynamics Lab — *Berkeley, CA*

- Co-authored a paper accepted to the ICML 2026 AI for Science Workshop (arXiv:2606.31976) on identifying bias in forestry remote-sensing data
- Built a LangGraph system combining expert decision trees with vision-language models, outperforming the project's supervised-learning baselines

Software Development Engineer Intern (HyperPod Inference Team) — May 2025 – Aug. 2025
Amazon Web Services — *Santa Clara, CA*

- Implemented prefix-aware load balancing for a Kubernetes-based LLM inference operator, reducing end-to-end latency by 11%
- Benchmarked routing strategies across DeepSeek and Llama models from 1.5B to 70B parameters, measuring the impact of KV-cache reuse and prefix caching in vLLM
- Built load tests to validate the operator under production-scale traffic for cluster creation and model serving, presenting results before product launch

Projects

Prefix-X | *Rust, Python, Prefix Caching*

- Built an async Rust gateway that injects byte-identical shared prompt prefixes and pins same-topic requests to one server, so the upstream prefix cache hits across users

Keel | *TypeScript, Next.js, FastAPI, MongoDB*

- Shipped a live web app with a Next.js dashboard and FastAPI backend for tracing LLM requests, tracking token usage, and enforcing workflow-level spending limits

Mainstacks | *Go, CLI, TUI*

- Built a Go CLI and terminal interface for extracting, indexing, and searching reusable code patterns for coding agents

Leadership

Sponsorship Director — Sep. 2025 – Present
CalHacks — *Berkeley, CA*

- Lead sponsor outreach and partnerships for Cal Hacks 13.0 after supporting events with 2,000+ attendees and 30+ sponsors including Anthropic, Meta, and SpaceX

Technical Skills

Languages: Python, C/C++, Go, Rust, TypeScript, Java, OCaml, SQL, Bash
Systems & Cloud: Linux, Git/GitHub, Docker, Kubernetes, Slurm, AWS (EKS, S3, Lambda, EventBridge, IAM, CDK)
Backend & Databases: FastAPI, PostgreSQL, Redis, Prometheus
Web: React, Next.js, Supabase, Playwright
AI & ML: vLLM, SGLang, PyTorch, NVML, NumPy, pandas, scikit-learn