

Collin Hargreaves

815-878-7034 | collin_hargreaves@berkeley.edu | github.com/cjhargreaves | cjhargreaves.me

EDUCATION

University of California, Berkeley

Bachelor of Arts, Data Science, Minor in Computer Science

Berkeley, CA

Expected December 2027

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

WORK EXPERIENCE

Amazon Web Services

Santa Clara, CA

Software Development Engineer Intern (HyperPod Dataplane Team)

May 2026 – Present

- 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

Amazon Web Services

Santa Clara, CA

Software Development Engineer Intern (HyperPod Inference Team)

May 2025 – Aug 2025

- Implemented load-balancing strategies for a Kubernetes-based LLM inference operator, reducing end-to-end latency by 11%
- Evaluated sticky-session routing across DeepSeek and Llama models from 1.5B to 70B parameters, measuring the latency benefits of KV-cache reuse with 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

RESEARCH

Lawrence Berkeley National Lab, Quantitative Ecosystem Dynamics Lab

Berkeley, CA

Undergraduate Researcher

Jan 2026 – Present

- 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

PROJECTS

Prefix-X: Prefix Cache Optimization for LLM Serving | Rust, Python, REST APIs

- Built a Rust gateway with Axum and Tokio that shares cacheable prompt prefixes across concurrent LLM requests, avoiding repeated processing of identical context

Mainstacks: Reusable Skill Library Tooling for Coding Agents | Go, CLI, TUI

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

Keel: AI Agent Observability & Spend Enforcement Platform | TypeScript, Python, Next.js, FastAPI, MongoDB

- Built a Python SDK and backend for tracing LLM requests, tracking token usage, and enforcing workflow-level spending limits

EXTRACURRICULARS

CalHacks

Berkeley, CA

Sponsorship Director

Sep 2025 – Present

- Lead sponsor outreach and partnerships for Cal Hacks 13.0 after supporting events with 2,000+ attendees and 30+ sponsors across Cal Hacks 12.0 and AI Hackathon 2026

TECHNICAL SKILLS

Languages: Python, C++, C, TypeScript, Rust, Go, Java, OCaml, SQL, Bash

Systems & Cloud: Linux, Git/GitHub, Docker, Kubernetes, Slurm, AWS (EKS, S3, Lambda, EventBridge, IAM)

Backend & Databases: Tokio, Axum, FastAPI, PostgreSQL, Redis

AI & ML: vLLM, PyTorch, SageMaker, NumPy, pandas, scikit-learn