




Sumer Kohli

 /in/sumerkohli |  sumer@cs.stanford.edu |  @firebolt55439

 Saratoga, CA 95070 |  (408) 621-6422

EDUCATION

Stanford University

GPA: 4.06

M.S., Computer Science (Distributed Systems & Machine Learning)

Sep 2022 – June 2024

- **Distinctions:** Guaranteed Course Assistantship (full financial support, awarded to the top admitted students).
- **Relevant Coursework:** CS349D (Cloud Computing); CS149 (Parallel Computing); CS217 (Hardware Acceleration for ML); CS140E/240LX (Adv. OS Development); CS144 (Networking); CS244 (Programming Languages); CS243 (Adv. Compilers)

University of California, Berkeley

GPA: 4.00

B.S., Electrical Engineering & Computer Sciences

Aug 2018 – May 2022

- **Distinctions:** Highest Honors; 2021 Outstanding GSI Award; CalHacks Prize Winner (4x); HackMIT Prize Winner (2x)
- **Relevant Coursework:** CS170 (Algorithms); CS162 (OS & Systems); CS189 (Machine Learning); CS161 (Security); CS186 (Databases); CS61C (Computer Architecture); CS61B (Data Structures); EECS127 (Optimization Models); EECS126 (Probability)

INDUSTRY EXPERIENCE

Databricks Inc.

San Francisco, CA

Software Engineer Intern

Jun - Sep 2023

- Improved Unity Catalog view performance by 2 to 15× on shared clusters by introducing multi-level caching of dependency DAGs.
- Designed and deployed a new batch endpoint to linearly reduce table credentialing latency and database load for views, computing batch size and view-table dependent pairing (heuristically, as problem is NP-complete) to optimize entity cache performance.

Roblox Corp.

San Mateo, CA

Software Engineer Intern

Jun - Aug 2022

- Delivered a Rust and C# implementation of a bulk update operation for the core storage layer that can scale to 100K+ QPS and provides a linear latency reduction with respect to batch size, unblocking migration onto our next-gen storage infrastructure.
- Created an accessible and feature-complete Rust template for gRPC backend services, reducing Rust onboarding time to minutes.

Citadel LLC

New York, NY

Software Engineer Intern

Jun - Aug 2021

- Engineered a Kafka trade pipeline in Java for regulatory reporting that parses, transforms, and transports up to 6B trades/day.
- Built a Java library and accompanying write-behind cache to replay misprocessed Kafka messages, critical for error handling.
- Rigorously tested pipelines and replay library due to zero industry error tolerance for missing trades, and deployed to production.

University of California, Berkeley

Berkeley, CA

Teaching Assistant for EECS16A (Fall '19, '20), 16B (Spring '20), and CS61B (Spring '21 to Spring '22) Aug 2019 - May 2022

- Led development of group matching software that has been used during COVID semesters by classes totaling over 5,000 students. I co-authored and published a conference research paper on its efficacy. Won the 2021 Outstanding GSI Award.
- Jointly led the core infrastructure team to support 1,000+ students and 40+ staff, designing and writing software as needed.
- Taught discussion sections, labs, and office hours, and was rated markedly above (4.81/5) the course staff average (4.64/5).

Microsoft Inc.

Sunnyvale, CA

Software Engineer Intern

Jun - Aug 2020

- Designed, developed, and deployed a new customer-facing Azure Communications service using C#/ASP.NET with my team, and a fully-featured UI using React/TypeScript (further details under NDA). Won the 2020 Garage Team Hero award.
- Implemented a C# backend for automatic ML-based captioning for the Windows Photo app with 300M+ yearly users.

Lawrence Livermore National Laboratory

Livermore, CA

Computational Scholar Intern

Jun - Aug 2019

- Researched and developed a Python-based key-escrow server on AWS and Docker to enable Full Disk Encryption (FDE) on the Lab's 3,500+ Macs, greatly improving operational security in response to escalating state-sponsored cyberattacks.
- Integrated and documented a REST API to enable authenticated access to user, machine, and recovery key metadata.

Nutanix Inc.

San Jose, CA

Software Engineer Intern

Jun - Aug 2015, Jun - Aug 2017

- Developed a performant Python and MySQL backend to process and store over 1 million product telemetry data points a day.
- Built a fully-featured web interface to tabulate and visualize gigabytes of product telemetry in near real-time for 2,000+ customers.
- Implemented reliable logging of core processes in C++, preventing potential catastrophic data loss during cluster imaging.

RESEARCH

Berkeley RISELab (now Sky Lab)

Research Assistant – Advisor: Prof. Randy Katz

Berkeley, CA
Aug 2020 - present

- Researching runtime optimizations for distributed serverless applications with the aim of running optimized code from the very start ("hot-start"), thereby greatly reducing latency and cost of serverless workloads.
- Designed and iterated upon software to execute research strategies, evaluate findings, and visualize results; pushed over 12,000 lines of code over course of research, illustrating rapid pace of work.
- Published initial findings in HotOS '21; have a paper in review for EuroSys '24 to present findings on the potential of intelligent checkpoint scheduling to provide secular improvements for serverless applications employing a checkpoint-restore mechanism.

Group Matching Research

Undergraduate Teaching Assistant, later Research Assistant – Advisor: Prof. Gireeja Ranade


Berkeley, CA
Jun 2020 - May 2022

- Led development of group matching software that has been used during by classes totaling over 5,000 students at UC Berkeley. Won Berkeley's 2021 Outstanding GSI Award and Extraordinary Teaching in Extraordinary Times Award for this work.
- Designed an ML approach that synthesizes simulated annealing, and feature engineering to efficiently form groups while optimizing for predicted student satisfaction over a very large search space.
- Published a paper detailing our initial matcher design and analysis of study group quality in SIGCSE '23 as a co-first author.

Berkeley RISELab

Research Assistant – Advisor: Prof. Alvin Cheung

Berkeley, CA
Oct 2021 - Mar 2023

- Researching how to apply verified lifting, a breakthrough in program analysis, to improve and generalize program synthesis.
- Building a framework called *Metalift*  to leverage these techniques for a variety of domain-specific languages, offering low-to-zero-cost language abstractions by synthesizing performant and formally correct equivalent programs.
- Published a paper describing the design of Metalift and its ability to provably-correctly transpile languages in ECOOP '23.

PUBLICATIONS

Building Code Transpilers for Domain-Specific Languages Using Program Synthesis

Jul 2023

Sahil Bhatia, **Sumer Kohli**, Sanjit A. Seshia, and Alvin Cheung. 2023. Building Code Transpilers for Domain-Specific Languages Using Program Synthesis. In *European Conference on Object-Oriented Programming (ECOOP '23)*.

Inclusive Study Group Formation At Scale

Mar 2023

Sumer Kohli*, Neelesh Ramachandran*, Ana Tudor*, Gloria Tumushabe*, Olivia Hsu, and Gireeja Ranade. 2023. Inclusive Study Group Formation At Scale. In *ACM Technical Symposium on Computer Science Education (SIGCSE '23)*. <https://doi.org/10.1145/3545945.3569885>

From Warm to Hot Starts: Leveraging Runtimes for the Serverless Era

Feb 2021

Joao Carreira, **Sumer Kohli**, Rodrigo Bruno, and Pedro Fonseca. 2021. From Warm to Hot Starts: Leveraging Runtimes for the Serverless Era. In *Workshop on Hot Topics in Operating Systems (HotOS 21)*. <https://doi.org/10.1145/3458336.3465305>

PROJECTS

DreamRL

Jan 2019

A CNN-VAE → MDN-RNN → CMA-ES TensorFlow model co-developed with my Cal Launchpad team; achieved state-of-the-art performance on the CarRacing-v0 OpenAI gym.

$\frac{d}{dx}$ it!

Oct 2018

A web application that instantaneously takes partial derivatives of arbitrary mathematical functions; supports extremely complex input, and computes everything client-side using Scheme and JavaScript (far faster than WolframAlpha!).

Cobra Compiler

Sep 2017

An LLVM-based optimizing compiler for my very own object-oriented, Turing-complete programming language; code generation is fully C/C++ ABI-compatible and supports multithreading and custom operators.

Sumer's OS

Dec 2016

A Unix-like mini-operating system with scheduling and multitasking support written in C; able to run code in user land and provides a (mostly) POSIX-compliant environment as well as a basic libc runtime.

SCE Chess

May 2016

A blazingly-fast C++ chess engine rated around 2045 ELO, or able to beat over 98% of chess players; employs alpha-beta pruning and a bitboard representation to evaluate nearly 2 million positions per second.

SKILLS

Languages	C/C++, Python, Java, JavaScript/Node.js, Go, Rust, Scala, TypeScript, Objective-C, C#, Shell, OCaml, R
Technologies	AWS, GCP, Heroku; Docker, Kubernetes, Spark; MongoDB, MySQL, PostgreSQL; React, AngularJS, Vue.js
AI/ML	TensorFlow, PyTorch, Keras; LASSO, ANOVA, Kalman Filter; CNN, R-CNN, GNN, Transformer, BERT