Xander Davies

alexander_davies@college.harvard.edu • *github.com/xanderdavies*

HARVARD UNIVERSITY December 2023

AB in Computer Science. *GPA 3.95/4.00*

Relevant Coursework: Linear Algebra and Real Analysis I, Intro to Probability, Intro to Statistical Inference, Advanced Computer Vision, Theory of Computation, Machine Learning.

Research

REDWOOD RESEARCH Present

REMIX Co-Head, *redwoodresearch.org/remix*

- Co-led a team of 40 in a 5-week research residency on neural network interpretability.
- Oversaw tutorial curriculum development, managed TAs and 6 research leads, contributed to project selection and directed research efforts.

KRUEGER LAB *University of Cambridge* Summer 2022 – Present

Research Assistant

- Led a research project on the shared learning dynamics in grokking and deep double descent, establishing a pattern-learning framework and demonstrating model-wise grokking; resulted in a first-author paper in the NeurIPS 2022 ML Safety Workshop.
    - **Davies, X.***, Langosco, L.*, & Krueger, D. (2022). Unifying Grokking and Double Descent. In *2022 NeurIPS ML Safety Workshop.*

KREIMAN LAB *Center for Brains, Minds, and Machines* Aug 2020 – Sep 2021

Research Assistant

- Worked on a biologically inspired approach to continual learning tasks, addressing the "dead neuron problem," where activation-sparse neurons never use large groups of neurons, and the "stale momentum problem," where lingering momentum terms can lead to a higher effective learning rate; resulted in a second-author paper in review for ICLR.
    - Bricken, T., **Davies, X.**, Singh, D., Krotov, D., & Kreiman, G. (2022). Sparse Distributed Memory is a Continual Learner. *Under review (avg 6.75), ICLR 2022.*

Other

HARVARD AI SAFETY TEAM *haist.ai* January 2022 – Present

Founder, Director

- Founded and lead a research-oriented student group (HAIST) aimed at reducing risks from advanced AI system, now considered one of the strongest AI safety student groups in the world; assisted in founding and running the MIT AI Alignment (MAIA) student group.
- Set strategic direction, led top-student outreach, facilitated two cohorts of intro fellowship, and ran weekly meetings reading and discussing alignment-relevant publications.

ZETA ASSOCIATES *A Lockheed Martin Company* May 2021 – Aug 2021

Software Engineering Intern

- In collaboration with DARPA's Semantic Forensics program, worked in a small team on an internal production tool to automate the generation of synthetic media with Generative Adversarial Networks. Received full-time job offer after completion of internship.

HARVARD COLLEGE WRITING CENTER *Writing Tutor* Sept 2020 – Aug 2022

- Recommended by writing preceptor to tutor Harvard students in clarity and style.