

Alan Sun

awsun@cmu.edu • alansun17904.github.io

EDUCATION

Carnegie Mellon University

Expected Dec. 2025

- Master of Science in **Computer Science** GPA 4.0/4.0
- **Relevant Coursework:** Machine Learning, Convex Optimization, Intermediate Statistics, Advanced Statistical Theory I, Large Language Model Systems

Dartmouth College

June 2024

- Bachelor of Arts, **Computer Science, Mathematics**, *magna cum laude*, high honors GPA 3.9/4.0
- **Honors Thesis:** Achieving Domain-Independent Certified Robustness via *Knowledge Continuity*
- **Relevant Coursework:** Machine Learning, Information Theory, Probability (Honors), Real Analysis (Honors), Measure Theory, Probability and Statistical Inference, Computer Vision, Data-Driven Uncertainty Quantification, Algorithms, Randomized Algorithms

PUBLICATIONS

1. Circuit Stability Characterizes Language Model Generalization

A. Sun

In Review @ ACL (2025)

2. Algorithmic Phase Transitions in Large Language Models: A Mechanistic Case Study of Arithmetic

A. Sun, E. Sun, W. Shepard

2nd Workshop on Attributing Model Behavior at Scale at NeurIPS (2024) [\[pdf\]](#)

3. Achieving Domain-Independent Certified Robustness via Knowledge Continuity

A. Sun, C. Ma, K. Ge, S. Vosoughi

NeurIPS (2024) [\[pdf\]](#)

John G. Kemeny Computing Prize for Innovation (2024)

Neukom Prize for Outstanding Undergraduate Research—First Prize (2024)

4. On the Exploration of LM-Based Soft Modular Robot Design

W. Ma, L. Zhao, C.Y. She, Y. Jiang, A. Sun, B. Zhu, D. Balkcom, S. Vosoughi

In Review @ IEEE RoboSoft (2025) [\[pdf\]](#)

5. Deciphering Stereotypes in Pre-Trained Language Models

W. Ma, H. Scheible, B. Wang, G. Veeramachaneni, P. Chowdhary, A. Sun, A. Koulogeorge, L. Wang, S. Vosoughi

EMNLP (2023) [\[pdf\]](#)

Oral Presentation

6. ThanosNet: A Novel Trash Classification Method Using Metadata

A. Sun and H. Xiao

IEEE Big Data (2020) [\[pdf\]](#)

HONORS AND AWARDS

- [National Science Foundation Graduate Research Fellowship—NSF GRFP \(2025\)](#).
- [John G. Kemeny Computing Prize for Innovation \(2024\)](#). Intended to encourage novel uses of computing by undergraduate Dartmouth students. Rewards students who produce original, creative, well-designed, and well-implemented computer programs.
- [Neukom Prize for Outstanding Undergraduate Research—First Prize \(2024\)](#). Recognizes outstanding graduate/undergraduate research in computational sciences at Dartmouth.

- **Francis L. Town Prize for Achievement in Computer Science (2023).** Presented annually to one exceptional student in computer science at Dartmouth.
- **James O. Freedman Presidential Scholar (2023).** Provides funding for undergraduate students to work as research assistants with Dartmouth faculty.
- **Goldwater Scholarship Program Nominee (2023).** One of five students nominated to represent Dartmouth in the national Barry Goldwater Scholar selection.
- **Dartmouth College Second Honors Group (2023; 2022).** Awarded annually to top 15% of all undergraduates.
- **JHU/APL Achievement Award for Technical Excellence (2022).** Given to interns who make meaningful technical contributions to their projects, produce work of exception quality.
- **Bronze Medal in Options Trading at UChicago Trading Competition (2022).** Created a real-time algorithm which makes markets for options sensitive to catastrophic events.
- **Silver Medal for Kaggle Toxic Comment Classification Challenge (2020).** Achieved a top 4% finish, 72nd out of 1621 teams. Created language models to identify multilingual toxic comments using only English training set.

RESEARCH EXPERIENCE

Bridging AI and Neuroscience Group (Max-Planck Institute for Software Systems) June 2024 – Present

Advisor: Mariya Toneva

- **Mechanistic Interpretability and Brain Alignment.** Language models have been shown to be aligned with brain activity, we seek to uncover the mechanisms (subcircuits) that cause this alignment and analyze the downstream abilities of these mechanisms.

Mind, Machines, Society Group (Dartmouth College)

Sept. 2022 – June 2024

Advisors: Soroush Vosoughi, Chiyu Ma

- **Domain-Independent Certified Robustness via Knowledge Continuity (March 2023 – June 2024)**
Formulated a new metric, *knowledge continuity*, which when minimized provably minimizes adversarial robustness while not limiting the expressiveness of the hypothesis class. Constructively demonstrating that robustness and accuracy are not at odds.
- **Automated Soft Modular Robot Design (March 2023 – June 2023)**
Created language models trained using policy-gradient methods to automatically assemble soft-lattice robots based on user-specified task and robot deployment environment. This work was done in collaboration with the Dartmouth Robotics Lab.

Fu Lab (Dartmouth College)

Aug. 2023 – Jan. 2024

Advisor: Feng Fu

- **Mechanistic Interpretation with Large Language Models (Aug. 2023 – June 2024)**
Developing novel methods to mechanistically interpret neural networks trained to play iterated prisoner's dilemma. The goal is to generate (end-to-end) natural language explanations both at the neuronal and network level using large language models.
- **Information Bottleneck Theory to Explain Adversarial Attacks (March 2022 – June 2022)**
Used information bottleneck theory and the information plane to characterize and explain adversarial robustness of neural networks based on their activation functions. *Awarded Francis L. Town Prize.*

Applied Physics Lab (Johns Hopkins University), Threat Analytics Group

June 2022 – Sept. 2022

Advisors: Sarah Prata, Alex Memory

- Built and designed graph neural networks to detect and predict trends of toxic posts and comments throughout Reddit communities.
- Worked with agency responsible for generating the data on collecting techniques to avoid implicit biases.
- Performed exploratory data analysis on online forum data and proposed novel, graph-based metrics for quantifying post-comment relationships.

TEACHING

MATH20 Probability, Dartmouth College	
Grader	2024
MATH63 Real Analysis, Dartmouth College	
Grader	2024
CS31 Algorithms, Dartmouth College	
Teaching Assistant	2023
MATH11 Multivariate Calculus, Dartmouth College	
Peer Tutor	2022
CS56 Digital Electronics, Dartmouth College	
Teaching Assistant	2022

MENTORSHIP

Ethan Sun	
BA, Dartmouth College	2024-2025
Warren Shepard	
BA, Dartmouth College	2024-2025
Ava Carlson	
BA, Dartmouth College	
Women in Science Project Intern	2024
Kenneth Ge	
BS, Carnegie Mellon University	2023-2024
Chikwanda Chisha	
BA, Dartmouth College	
E.E Just Summer Research Intern	2023

SERVICE

ICLR , Reviewer	2025
MATH-AI @ NeurIPS , Reviewer	2024
ATTRIB @ NeurIPS , Reviewer	2024