Zachary S. Siegel

51 Vassar St, Cambridge, MA 02139

zss@mit.edu
www.zacharysiegel.org
Github
Google Scholar

EDUCATION

Massachusetts Institute of Technology, Cambridge, Massachusetts, USA — Sep 2025 – Present

PhD in Electrical Engineering and Computer Science
Advisors: Leslie P. Kaelbling, Tomás Lozano-Pérez, and Joshua B. Tenenbaum

Princeton University, Princeton, New Jersey, USA — Sep 2021 – May 2025

B.S.E. in Computer Science, Minor in Philosophy
Honors: Summa Cum Laude, Tau Beta Pi, Sigma Xi, Outstanding Computer Science Independent Work Prize

PUBLICATIONS

[1] **Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation**
Sayash Kapoor*, Benedikt Stroebl*, Peter Kirgis, Nitya Nadgir, **Zachary S. Siegel**, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, Franck Ndzomga, Dheeraj Oruganty, Sophie Luskin, Kangheng Liu, Botao Yu, Amit Arora, Dongyoon Hahm, Harsh Trivedi, Huan Sun, Juyong Lee, Tengjun Jin, Yifan Mai, Yifei Zhou, Yuxuan Zhu, Rishi Bommasani, Daniel Kang, Dawn Song, Peter Henderson, Yu Su, Percy Liang, Arvind Narayanan
Preprint.

[2] **Characterizing the Implicit Bias of Regularized SGD in Rank Minimization**
Tomer Galanti, **Zachary S. Siegel**, Aparna Gupte, Tomaso Poggio
Conference on Parsimony and Learning (CPAL 2025)

[3] **AI Agents That Matter**
Sayash Kapoor*, Benedikt Stroebl*, **Zachary S. Siegel**, Nitya Nadgir, Arvind Narayanan
Transactions on Machine Learning Research (2025)

[4] **CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark**
Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, Arvind Narayanan
Transactions on Machine Learning Research (2025)

[5] **BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive Retrieval**
Hongjin Su*, Howard Yen*, Mengzhou Xia*, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, **Zachary S. Siegel**, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik, Danqi Chen, Tao Yu
The Thirteenth International Conference on Learning Representations (ICLR 2025)

[6] **Language Guided Operator Learning for Goal Inference**
Zachary S. Siegel, Jiayuan Mao, Nishanth Kumar, Tianmin Shu, Jacob Andreas
Workshop on Learning Effective Abstractions for Planning @ CORL (2024)

[7] **Learning Grounded Action Abstractions from Language**
Lionel Wong*, Jiayuan Mao*, Pratyusha Sharma*, **Zachary S. Siegel**, Jiahai Feng, Noa Korneev, Joshua B Tenenbaum, Jacob Andreas
The Twelfth International Conference on Learning Representations (ICLR 2024)

[8] **Superimposing height-controllable and animated flood surfaces into street-level photographs for risk communication**
Zachary S. Siegel, Scott A Kulp
Weather and Climate Extremes, Volume 32 (2021)

RESEARCH EXPERIENCE

Senior Thesis — May 2024 – May 2025
Department of Computer Science, Princeton University

Modeling Open-Ended Goal Inference. I worked with Professors Tom Griffiths and Jacob Andreas to model how people infer the goals of others in open-ended goal spaces. We hypothesize that people use a learned transition model of the environment to assist with predicting goals, which we will validate by creating a new domain, running human experiments, and building a computational model. I am also the first author of a paper showing how to learn operators both symbolically and with language models for goal prediction, which was published at the Learning Effective Abstractions for Planning Workshop @ CORL 2024.

Undergraduate Researcher Nov 2023 – May 2025
Center for Information Technology Policy, Princeton University

Building and Evaluating Agent Benchmarks. I worked with Professor Arvind Narayanan to evaluate and build agent benchmarks for automating aspects of scientific research. I am the first author of CORE-Bench, published at TMLR, which evaluates how agents can computationally reproduce existing scientific papers, a co-author of AI Agents that Matter, published at TMLR, which argues many existing agent evaluation benchmarks are methodologically flawed.

Visiting Undergraduate Researcher May 2023 – Sep 2023
Language & Intelligence Group, MIT

Learning Grounded Abstractions from Language. I worked with Professors Jacob Andreas and Josh Tenenbaum to use large language models (LLMs) to learn operators for task and motion planning systems (TAMP). I implemented policy learning approaches for low-level motion planning in the Alfred domain and integrated the pipeline of using LLMs for TAMP. I am a co-author on the paper, which was published at ICLR 2024.

Undergraduate Researcher Apr 2024 – May 2024
Department of Computer Science, Princeton University

Building Retrieval Benchmarks. I worked with Professor Danqi Chen to build a retrieval benchmark for tasks where semantic similarity is not sufficient for matching queries to documents. I helped select a domain and developed a pipeline, from the Art of Problem Solving wiki, to scrape questions for use in the retrieval benchmark. I am a co-author of the paper currently accepted at ICLR.

Junior Independent Work Sep 2023 – May 2024
Department of Computer Science, Princeton University

Training LLMs on Podcasts. I worked with Professor Danqi Chen to use podcasts for improving the conversational abilities of LLMs. I found podcast data sources, developed methods to transcribe hundreds of thousands of hours of audio in parallel, fine-tuned LLaMA-2-7B, and compared performance to base models. Additionally, I investigated the quality of podcast audio as a training source.

Visiting Undergraduate Researcher May 2022 – Aug 2022
Center for Brains, Minds, and Machines, MIT

Investigating the Low-Rank Bias of Neural Networks. I worked with Professors Tomer Galanti and Tomaso Poggio to investigate why the ranks of weight matrices of neural networks are minimized during stochastic gradient descent. I helped develop a theoretical bound on the rank and ran extensive experiments with different network architectures to investigate the rank and singular values of weight matrices. I am a co-author on the paper published at the M3L Workshop @ NeurIPS 2024.

Summer Intern May 2019 – Jun 2021
Climate Central

Visualizing Flood Surges for Risk Communication. I worked with Dr. Scott Kulp to build a system that superimposes water surfaces on street view images to communicate the risk of flood surges to vulnerable communities. From LIDAR depth maps and RGB images, the method uses depth completion techniques to generate a 3D model of the environment, and then superimposes a flood surface using Blender. The system is now being deployed across the country to communicate risk. I am the first author on the paper published in the Weather and Climate Extremes journal.

SELECTED COURSES

▪ ORF 309: Probability and Stochastic Systems	Fall 2022
▪ PSY 255: Cognitive Psychology	Fall 2022
▪ COS 484: Natural Language Processing	Spring 2023
▪ ORF 307: Optimization	Spring 2023
▪ COS 597Q: AI Safety and Alignment	Fall 2023
▪ COS 345: Robotics	Fall 2024
▪ COS 435: Reinforcement Learning	Spring 2025

TEACHING EXPERIENCE

▪ COS 226 (Algorithms) Precept Assistant	Spring 2022
▪ MAT 203 (Multivariable Calculus) Course Assistant	Fall 2022

- MAT 204 (Linear Algebra) Course Assistant Spring 2023
- COS 484 (Natural Language Processing) Course Assistant and Grader Spring 2024
- COS 426 (Computer Graphics) Course Assistant Fall 2024
- COS 484 (Natural Language Processing) Course Assistant and Grader Spring 2025

[*CV compiled on 2025-11-10*]