Avi Schwarzschild

schwarzschild@cmu.edu ◇ avischwarzschild.com ◇ Google Scholar

EMPLOYMENT

Carnegie Mellon University 2023 – Present
Postdoctoral Researcher (Advised by Zico Kolter)

Arthur 2022 – 2023
Research Fellow

EDUCATION

University of Maryland 2018 – 2023
PhD in Applied Mathematics and Scientific Computation (Advised by Tom Goldstein)

University of Washington 2017 – 2018
MS in Applied Mathematics

Columbia University 2013 – 2017
BS in Applied Mathematics

SELECTED AWARDS

The First Workshop on Data Contamination (CONDA@ACL) Best Paper Award 2024
University of Maryland Invention of the Year 2024
Three Minute Thesis Award 2023
Backdoors in Deep Learning Workshop @ NeurIPS 2023 Best Paper Award 2023
Aziz Osborn Gold Medal in Teaching Excellence 2020
University of Maryland Dean's Fellowship 2018, 2019

PUBLICATIONS

Rethinking LLM Memorization through the Lens of Adversarial Compression.
Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C Lipton, and J Zico Kolter.
Neural Information Processing Systems (NeurIPS) (2024)

Transformers Can Do Arithmetic with the Right Embeddings.
Sean McLeish, Arpit Bansal, Alex Stein, Neel Jain, John Kirchenbauer, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, Jonas Geiping, Avi Schwarzschild, et al..
Neural Information Processing Systems (NeurIPS) (2024)

TOFU: A Task of Fictitious Unlearning for LLMs.
Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter.
Conference on Language Modeling (COLM) (2024)

Forcing Diffuse Distributions out of Language Models.
Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, Zico Kolter, and Daphne Ippolito.
Conference on Language Modeling (COLM) (2024)

Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text.
Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein.
International Conference on Machine Learning (ICML) (2024)

Universal Guidance for Diffusion Models.
Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein.
International Conference on Learning Representations (ICLR) (2024)

NEFTune: Noisy Embeddings Improve Instruction Finetuning.
Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, et al..
International Conference on Learning Representation (ICLR) (2024)

Reckoning with the Disagreement Problem: Explanation Consensus as a Training Objective.
Avi Schwarzschild, Max Cembalest, Karthik Rao, Keegan Hines, and John Dickerson.
Artificial Intelligence, Ethics, and Society (AIES) (2023)

Transfer Learning with Deep Tabular Models.
Roman Levin, Valeriia Cherepanova, Avi Schwarzschild, Arpit Bansal, C Bayan Bruss, Tom Goldstein, Andrew Gordon Wilson, and Micah Goldblum.
International Conference on Learning Representations (ICLR) (2023)

Effective Backdoor Mitigation Depends on the Pre-training Objective.
Sahil Verma, Gantavya Bhatt, Soumye Singhal, Arnav Mohanty Das, Chirag Shah, John P Dickerson, and Jeff Bilmes.
NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly (2023)

End-to-end Algorithm Synthesis with Recurrent Networks: Logical Extrapolation Without Overthinking.
Arpit Bansal, Avi Schwarzschild, Eitan Borgnia, Zeyad Emam, Furong Huang, Micah Goldblum, and Tom Goldstein.
Neural Information Processing Systems (NeurIPS) (2022)

The Uncanny Similarity of Recurrence and Depth.
Avi Schwarzschild, Arjun Gupta, Amin Ghiasi, Micah Goldblum, and Tom Goldstein.
International Conference on Learning Representations (ICLR) (2022)

Can You Learn an Algorithm? Generalizing from Easy to Hard Problems with Recurrent Networks.
Avi Schwarzschild, Eitan Borgnia, Arjun Gupta, Furong Huang, Uzi Vishkin, Micah Goldblum, and Tom Goldstein.
Neural Information Processing Systems (NeurIPS) (2021)

Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks.
Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein.
International Conference on Machine Learning (ICML) (2021)

Adversarial Attacks on Machine Learning Systems for High-Frequency Trading.
Micah Goldblum, Avi Schwarzschild, Ankit B Patel, and Tom Goldstein.
ACM International Conference on AI in Finance (ICAIF) (2021)

Headless Horseman: Adversarial Attacks on Transfer Learning Models.
Ahmed Abdelkader, Michael J Curry, Liam Fowl, Tom Goldstein, Avi Schwarzschild, Manli Shu, Christoph Studer, and Chen Zhu.
ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020)

Truth or Backpropaganda? An Empirical Investigation of Deep Learning Theory.
Micah Goldblum, Jonas Geiping, Avi Schwarzschild, Michael Moeller, and Tom Goldstein.
International Conference on Learning Representations (ICLR) (2019)

SELECTED PREPRINTS

Prompt Recovery for Image Generation Models: A Comparative Study of Discrete Optimizers.
Joshua Nathaniel Williams, Avi Schwarzschild, and J Zico Kolter.
arXiv preprint arXiv:2408.06502 (2024)

Benchmarking ChatGPT on Algorithmic Reasoning.
Sean McLeish, Avi Schwarzschild, and Tom Goldstein.
arXiv preprint arXiv:2404.03441 (2024)

The CLRS-Text Algorithmic Reasoning Language Benchmark.
Larisa Markeeva, Sean McLeish, Borja Ibarz, Wilfried Bounsi, Olga Kozlova, Alex Vitvitskyi, Charles Blundell, Tom Goldstein, Avi Schwarzschild, and Petar Veličković.
arXiv preprint arXiv:2406.04229 (2024)

Neural Auctions Compromise Bidder Information.
Alex Stein, Avi Schwarzschild, Michael Curry, Tom Goldstein, and John Dickerson.
arXiv preprint arXiv:2303.00116 (2023)

A Cookbook of Self-Supervised Learning.
Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al..
arXiv preprint arXiv:2304.12210 (2023)

Baseline Defenses for Adversarial Attacks Against Aligned Language Models.
Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein.
arXiv preprint arXiv:2309.00614 (2023)

Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses.
Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein.
IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

MetaBalance: High-Performance Neural Networks for Class-Imbalanced Data.
Arpit Bansal, Micah Goldblum, Valeriia Cherepanova, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein.
arXiv preprint arXiv:2106.09643 (2021)

SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training.
Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein.
arXiv preprint arXiv:2106.01342 (2021)

Datasets for Studying Generalization from Easy to Hard Examples.
Avi Schwarzschild, Eitan Borgnia, Arjun Gupta, Arpit Bansal, Zeyad Emam, Furong Huang, Micah Goldblum, and Tom Goldstein.
arXiv preprint arXiv:2108.06011 (2021)

INVITED TALKS

Google Research Privacy Seminar October 2024
Allen Institute for AI MOSAIC Group August 2024
Max Planck Institute for Intelligent Systems in Tübingen July 2024
Deep Learning ONR MURI Meeting September 2021