Avi Schwarzschild

347·426·8421 ⋄ avis4k@gmail.com

avischwarzschild.com ⋄ Google Scholar

EMPLOYMENT

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

Arthur June 2022 - February 2023
Research Fellow

EDUCATION

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

University of Washington June 2018
MS in Applied Mathematics

Columbia University May 2017
BS in Applied Mathematics

SELECTED PAPERS

1. **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.. *arXiv preprint arXiv:2405.17399* (2024)
2. **Rethinking LLM Memorization through the Lens of Adversarial Compression**. Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C Lipton, and J Zico Kolter. *arXiv preprint arXiv:2404.15146* (2024)
3. **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. *arXiv preprint arXiv:2401.12070* (2024)
4. **TOFU: A Task of Fictitious Unlearning for LLMs**. Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. *arXiv preprint arXiv:2401.06121* (2024)
5. **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.. *arXiv preprint arXiv:2310.05914* (2023)
6. **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)
7. **Universal Guidance for Diffusion Models**. Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. *arXiv preprint arXiv:2302.07121* (2023)
8. **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. *NeurIPS* (2022)
9. **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)
10. **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)
11. **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. *NeurIPS* (2021)
12. **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)
13. **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)