

MICAH GOLDBLUM

micah.g@columbia.edu, goldblum.github.io

EMPLOYMENT

Columbia University, Department of Electrical Engineering Assistant Professor	<i>July 2024 - Present</i>
New York University Postdoctoral Researcher (Advised by Professors Yann LeCun and Andrew Gordon Wilson)	<i>September 2021 - June 2024</i>

EDUCATION

University of Maryland Ph.D. in Mathematics	<i>September 2014 - May 2020</i>
University of Maryland B.Sc. in Mathematics	<i>September 2010 - May 2014</i>

SELECTED HONORS AND AWARDS

Blavatnik Awards for Young Scientists Finalist	<i>2023</i>
NYU Postdoctoral Research and Professional Development Support Grant	<i>2023</i>
ICML Outstanding Paper Award	<i>2022</i>
Seymour Goldberg Gold Medal - Spotlight on Graduate Student Research	<i>2016</i>
University of Maryland Dean's Fellowship	<i>2014, 2015</i>
Phi Beta Kappa	<i>2014</i>

SELECTED PREPRINTS

A Cookbook of Self-Supervised Learning Randall Balestriero, Mark Ibrahim, ... , Hamed Pirsiavash, Yann LeCun, Micah Goldblum
Commercial LLM Agents Are Already Vulnerable to Simple Yet Dangerous Attacks Ang Li, Yin Zhou, Vethavikashini Chithrara Raghuram, Tom Goldstein, Micah Goldblum
Transformers boost the performance of decision trees on tabular data across sample sizes Mayuka Jayawardhana, Samuel Dooley, Valeriia Cherepanova, Andrew Gordon Wilson, Frank Hutter, Colin White, Tom Goldstein, Micah Goldblum

PUBLICATIONS

Zebra-CoT: A Dataset for Interleaved Vision-Language Reasoning Ang Li, Charles L. Wang, Kaiyu Yue, Zikui Cai, Ollie Liu, Deqing Fu, ..., Micah Goldblum <i>International Conference on Learning Representations (ICLR) 2026</i>	<i>2026</i>
Identifying and Evaluating Inactive Heads in Pretrained LLMs Pedro Sandoval-Segura, Xijun Wang, Ashwinee Panda, Micah Goldblum , Ronen Basri, Tom Goldstein, David Jacobs <i>International Conference on Learning Representations (ICLR) 2026</i>	<i>2026</i>
Vista4D: Video Reshooting with 4D Point Clouds Kuan Heng Lin, Zhizheng Liu, ..., Bolei Zhou, Micah Goldblum , Paul Debevec, Ning Yu <i>Computer Vision and Pattern Recognition Conference (CVPR) 2026</i>	<i>2026</i>
Small Batch Size Training for Language Models: When Vanilla SGD Works, and Why Gradient Accumulation Is Wasteful Martin Marek, Sanae Lotfi, Aditya Somasundaram, Andrew Gordon Wilson, Micah Goldblum <i>Advances in Neural Information Processing Systems (NeurIPS) 2025</i>	<i>2025</i>
Gemstones: A Model Suite for Scaling Laws Sean Michael McLeish, John Kirchenbauer, David Yu Miller, Siddharth Singh, Abhinav Bhatele, Micah Goldblum , Ashwinee Panda, Tom Goldstein <i>Advances in Neural Information Processing Systems (NeurIPS) 2025</i>	<i>2025</i>
Brain-Predictive Reasoning Embedding through Residual Disentanglement Linyang He, Tianjun Zhong, Richard Antonello, Gavin Mischler, Micah Goldblum , Nima Mesgarani <i>Advances in Neural Information Processing Systems (NeurIPS) 2025</i>	<i>2025</i>

FineGRAIN: Evaluating Failure Modes of Text-to-Image Models with Vision Language Model Judges Kevin David Hayes, Micah Goldblum , Vikash Sehwal, Gowthami Somepalli, Ashwinee Panda, Tom Goldstein <i>Advances in Neural Information Processing Systems (NeurIPS) 2025</i>	2025
LiveBench: A Challenging, Contamination-Free LLM Benchmark Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, ... , Yann LeCun, Tom Goldstein, Willie Neiswanger, Micah Goldblum <i>International Conference on Learning Representations (ICLR) 2025</i>	2025
Style Outweighs Substance: Failure Modes of LLM Judges in Alignment Benchmarking Benjamin Feuer, Micah Goldblum , Teresa Datta, Sanjana Nambiar, Raz Besaleli, Samuel Dooley, Max Cembalest, John P Dickerson <i>International Conference on Learning Representations (ICLR) 2025</i>	2025
Adaptive Rentention & Correction for Continual Learning Haoran Chen, Micah Goldblum , Zuxuan Wu, Yu-Gang Jiang <i>International Conference on Learning Representations (ICLR) 2025</i>	2025
Hidden No More: Attacking and Defending Private Third-Party LLM Inference Arka Pal, Rahul Krishna Thomas, Louai Zahran, Erica Choi, Akilesh Potti, Micah Goldblum <i>International Conference on Machine Learning (ICML) 2025</i>	2025
Kolmogorov Complexity, the No Free Lunch Theorem, and the Role of Inductive Biases in Machine Learning Micah Goldblum , Marc Anton Finzi, Keefer Rowan, Andrew Gordon Wilson <i>International Conference on Machine Learning (ICML) 2024</i>	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, Tom Goldstein <i>International Conference on Machine Learning (ICML) 2024</i>	2024
Non-vacuous generalization bounds for large language models Sanae Lotfi, Marc Finzi, Yilun Kuang, Tim GJ Rudner, Micah Goldblum , Andrew Gordon Wilson <i>International Conference on Machine Learning (ICML) 2024</i>	2024
Large Language Models Must Be Taught to Know What They Dont Know Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Katherine Collins, Umang Bhatt, Adrian Weller, Micah Goldblum , Andrew Gordon Wilson <i>Advances in Neural Information Processing Systems (NeurIPS) 2024</i>	2024
TuneTables: Context Optimization for Scalable Prior-Data Fitted Networks Benjamin Feuer, Robin Tibor Schirrmeyer, Valeriia Cherepanova, Chinmay Hegde, Frank Hutter, Micah Goldblum , Niv Cohen, Colin White <i>Advances in Neural Information Processing Systems (NeurIPS) 2024</i>	2024
Unlocking Tokens as Data Points for Generalization Bounds on Larger Language Models Sanae Lotfi, Yilun Kuang, Marc Anton Finzi, Brandon Amos, Micah Goldblum , Andrew Gordon Wilson <i>Advances in Neural Information Processing Systems (NeurIPS) 2024</i>	2024
Searching for Efficient Linear Layers over a Continuous Space of Structured Matrices Andres Potapczynski, Shikai Qiu, Marc Anton Finzi, Christopher Ferri, Zixi Chen, Micah Goldblum , C. Bayan Bruss, Christopher De Sa, Andrew Gordon Wilson <i>Advances in Neural Information Processing Systems (NeurIPS) 2024</i>	2024
Compute Better Spent: Replacing Dense Layers with Structured Matrices Shikai Qiu, Andres Potapczynski, Marc Anton Finzi, Micah Goldblum , Andrew Gordon Wilson <i>International Conference on Machine Learning (ICML) 2024</i>	2024
Measuring Style Similarity in Diffusion Models Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta,	2024

Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, Tom Goldstein
European Conference on Computer Vision (ECCV) 2024

On the Reliability of Watermarks for Large Language Models 2024
John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu,
Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha,
Micah Goldblum, Tom Goldstein
International Conference on Learning Representations (ICLR) 2024

Universal guidance for diffusion models 2024
Arpit Bansal, Hong-Min Chu, Avi Schwarzschild,
Soumyadip Sengupta, **Micah Goldblum**,
Jonas Geiping, Tom Goldstein
International Conference on Learning Representations (ICLR) 2024

NEFTune: Noisy Embeddings Improve Instruction Finetuning 2024
Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer,
Hong-Min Chu, Gowthami Somepalli, Brian R Bartoldson,
Bhavya Kaikhura, Avi Schwarzschild, Aniruddha Saha,
Micah Goldblum, Jonas Geiping, Tom Goldstein
International Conference on Learning Representations (ICLR) 2024

Battle of the Backbones: A Large-Scale Comparison of Pretrained Models across Computer Vision Tasks 2023
Micah Goldblum, Hossein Sourì, Renkun Ni, Manli Shu, Viraj Uday Prabhu,
Gowthami Somepalli, Prithvijit Chattopadhyay, Adrien Bardes, Mark Ibrahim,
Judy Hoffman, Rama Chellappa, Andrew Gordon Wilson, Tom Goldstein
Advances in Neural Information Processing Systems (NeurIPS) 2023

Rethinking Bias Mitigation: Fairer Architectures Make for Fairer Face Recognition 2023
Samuel Dooley, Rhea Sukthanker, John P Dickerson, Colin White, Frank Hutter, **Micah Goldblum**
Advances in Neural Information Processing Systems (NeurIPS) 2023

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

Simplifying Neural Network Training Under Class Imbalance 2023
Ravid Shwartz-Ziv*, **Micah Goldblum***, Yucen Lily Li, C. Bayan Bruss, Andrew Gordon Wilson
Advances in Neural Information Processing Systems (NeurIPS) 2023

A Performance-Driven Benchmark for Feature Selection in Tabular Deep Learning 2023
Valeriia Cherepanova, Gowthami Somepalli, Jonas Geiping, C. Bayan Bruss,
Andrew Gordon Wilson, Tom Goldstein, **Micah Goldblum**
Advances in Neural Information Processing Systems (NeurIPS) 2023

Gradient-based optimization is not necessary for generalization in neural networks 2023
Ping-yeh Chiang, Renkun Ni, David Yu Miller, Arpit Bansal, Jonas Geiping,
Micah Goldblum, Tom Goldstein
International Conference on Learning Representations (ICLR) 2023

Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise 2023
Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang,
Micah Goldblum, Jonas Geiping, Tom Goldstein
Advances in Neural Information Processing Systems (NeurIPS) 2023

Why Diffusion Models Memorize and How to Mitigate Copying 2023
Gowthami Somepalli, Vasu Singla, **Micah Goldblum**,
Jonas Geiping, Tom Goldstein
Advances in Neural Information Processing Systems (NeurIPS) 2023

Hard Prompts Made Easy: Gradient-Based Discrete Optimization for Prompt Tuning and Discovery 2023
Yuxin Wen, Neel Jain, John Kirchenbauer, **Micah Goldblum**, Jonas Geiping, Tom Goldstein
Advances in Neural Information Processing Systems (NeurIPS) 2023

- When Do Neural Nets Outperform Boosted Trees on Tabular Data?** 2023
Duncan C. McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, **Micah Goldblum**, Colin White
Advances in Neural Information Processing Systems (NeurIPS) 2023
- What Can We Learn from Unlearnable Datasets?** 2023
Pedro Sandoval-Segura, Vasu Singla, Jonas Geiping, **Micah Goldblum**, Tom Goldstein
Advances in Neural Information Processing Systems (NeurIPS) 2023
- Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models** 2023
Gowthami Somepalli, Vasu Singla, **Micah Goldblum**, Jonas Geiping, Tom Goldstein
Computer Vision and Pattern Recognition Conference (CVPR) 2023
- The Lie Derivative for Measuring Learned Equivariance** 2023
Nate Gruver, Marc Anton Finzi, **Micah Goldblum**, Andrew Gordon Wilson
International Conference on Learning Representations (ICLR) 2023
- Exploring and Exploiting Decision Boundary Dynamics for Adversarial Robustness** 2023
Yuancheng Xu, Yanchao Sun, **Micah Goldblum**, Tom Goldstein, Furong Huang
International Conference on Learning Representations (ICLR) 2023
- Canary in a Coalmine: Better Membership Inference with Ensembled Adversarial Queries** 2023
Yuxin Wen, Arpit Bansal, Hamid Kazemi, Eitan Borgnia, **Micah Goldblum**, Jonas Geiping, Tom Goldstein
International Conference on Learning Representations (ICLR) 2023
- Panning for Gold in Federated Learning: Targeted Text Extraction under Arbitrarily Large-Scale Aggregation** 2023
Hong-Min Chu, Jonas Geiping, Liam Fowl, **Micah Goldblum**, Tom Goldstein
International Conference on Learning Representations (ICLR) 2023
- How Much Data Are Augmentations Worth? An Investigation into Scaling Laws, Invariance, and Implicit Regularization** 2023
Jonas Geiping, **Micah Goldblum**, Gowthami Somepalli, Ravid Shwartz-Ziv, Tom Goldstein, Andrew Gordon Wilson
International Conference on Learning Representations (ICLR) 2023
- Decepticons: Corrupted Transformers Breach Privacy in Federated Learning for Language Models** 2023
Liam Fowl, Jonas Geiping, Steven Reich, Yuxin Wen, Wojciech Czaja, **Micah Goldblum**, Tom Goldstein
International Conference on Learning Representations (ICLR) 2023
- A Deep Dive into Dataset Imbalance and Bias in Face Identification** 2023
Valeriia Cherepanova, Steven Reich, Samuel Dooley, Hossein Souri, **Micah Goldblum**, Tom Goldstein
Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES)
- Bayesian Model Selection, the Marginal Likelihood, and Generalization** 2022
Sanae Lotfi, Pavel Izmailov, Gregory Benton, **Micah Goldblum**, Andrew Gordon Wilson
International Conference on Machine Learning (ICML) 2022 (Outstanding Paper Award)
- Pre-Train Your Loss: Easy Bayesian Transfer Learning with Informative Priors** 2022
Ravid Shwartz-Ziv*, **Micah Goldblum***, Hossein Souri, Sanyam Kapoor, Chen Zhu, Yann LeCun, Andrew Gordon Wilson
Advances in Neural Information Processing Systems (NeurIPS) 2022
- Scalable Algorithm Synthesis with Recurrent Networks: Extrapolation without Overthinking** 2022
Arpit Bansal, Avi Schwarzschild, Eitan Borgnia, Zeyad Emam, Furong Huang, **Micah Goldblum**, Tom Goldstein
Advances in Neural Information Processing Systems (NeurIPS) 2022
- PAC-Bayes Compression Bounds So Tight That They Can Explain Generalization** 2022
Sanae Lotfi, Sanyam Kapoor, Marc Anton Finzi, Andres Potapczynski,

- Micah Goldblum**, Andrew Gordon Wilson
Advances in Neural Information Processing Systems (NeurIPS) 2022
- Where do Models go Wrong? Parameter-Space Saliency Maps for Explainability** 2022
Roman Levin, Manli Shu, Eitan Borgnia, Furong Huang, **Micah Goldblum**, Tom Goldstein
Advances in Neural Information Processing Systems (NeurIPS) 2022
- Sleeper Agent: Scalable Hidden Trigger Backdoors for Neural Networks Trained from Scratch** 2022
Hossein Souri, Liam H Fowl, Rama Chellappa, **Micah Goldblum**, Tom Goldstein
Advances in Neural Information Processing Systems (NeurIPS) 2022
- Chroma-VAE: Mitigating Shortcut Learning with Generative Classifiers** 2022
Wanqian Yang, Polina Kirichenko, **Micah Goldblum**, Andrew Gordon Wilson
Advances in Neural Information Processing Systems (NeurIPS) 2022
- Autoregressive Perturbations for Data Poisoning** 2022
Pedro Sandoval-Segura, Vasu Singla, Jonas Geiping,
Micah Goldblum, Tom Goldstein, David W. Jacobs
Advances in Neural Information Processing Systems (NeurIPS) 2022
- Fishing for User Data in Large-Batch Federated Learning via Gradient Magnification** 2022
Yuxin Wen, Jonas Geiping, Liam Fowl, **Micah Goldblum**, Tom Goldstein
International Conference on Machine Learning (ICML) 2022
- Plug-In Inversion: Model-Agnostic Inversion for Vision with Data Augmentations** 2022
Amin Ghiasi, Hamid Kazemi, Steven Reich, Chen Zhu, **Micah Goldblum**, Tom Goldstein
International Conference on Machine Learning (ICML) 2022
- Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses** 2022
Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild,
Dawn Song, Aleksander Madry, Bo Li, Tom Goldstein
IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 2022
- Can You Learn the Same Model Twice? Investigating Reproducibility and Double Descent from the Decision Boundary Perspective** 2022
Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk,
Micah Goldblum, Tom Goldstein
Conference on Computer Vision and Pattern Recognition (CVPR) 2022
- Contrastive Learning is Just Meta-Learning** 2022
Renkun Ni, Manli Shu, Hossein Souri, **Micah Goldblum**, Tom Goldstein
International Conference on Learning Representations (ICLR) 2022
- Stochastic Training is Not Necessary for Generalization** 2022
Jonas Geiping, **Micah Goldblum**, Phil Pope, Michael Moeller, Tom Goldstein
International Conference on Learning Representations (ICLR) 2022
- Robbing the Fed: Directly Obtaining Private Data in Federated Learning with Modified Models** 2022
Liam H Fowl, Jonas Geiping, Wojciech Czaja, **Micah Goldblum**, Tom Goldstein
International Conference on Learning Representations (ICLR) 2022
- The Uncanny Similarity of Recurrence and Depth** 2022
Avi Schwarzschild, Arjun Gupta, Amin Ghiasi, **Micah Goldblum**, Tom Goldstein
International Conference on Learning Representations (ICLR) 2022
- Towards Transferable Adversarial Attacks on Vision Transformers** 2022
Zhipeng Wei, Jingjing Chen, **Micah Goldblum**, Zuxuan Wu, Tom Goldstein, Yu-Gang Jiang
AAAI Conference on Artificial Intelligence (AAAI) 2022
- Adversarial Examples Make Strong Poisons** 2021
Liam Fowl*, **Micah Goldblum***, Ping-yeh Chiang, Jonas Geiping, Wojtek Czaja, Tom Goldstein
Advances in Neural Information Processing Systems (NeurIPS) 2021
- Can You Learn an Algorithm? Generalizing from Easy to Hard Problems with Recurrent Networks** 2021

Avi Schwarzschild, Eitan Borgnia, Arjun Gupta, Furong Huang, Uzi Vishkin,
Micah Goldblum, Tom Goldstein

Advances in Neural Information Processing Systems (NeurIPS) 2021

**Prepare for the Worst: Generalizing across Domain Shifts with Adversarial
Batch Normalization**

2021

Manli Shu, Zuxuan Wu, **Micah Goldblum**, Tom Goldstein

Advances in Neural Information Processing Systems (NeurIPS) 2021

Adversarial Attacks on Machine Learning Systems for High-Frequency Trading

2021

Micah Goldblum, Avi Schwarzschild, Naftali Cohen, Tucker Balch, Ankit B. Patel,
Tom Goldstein

ACM International Conference on AI in Finance (ICAIF) 2021

**Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and
Data Poisoning Attacks**

2021

Avi Schwarzschild*, **Micah Goldblum***, Arjun Gupta, John P Dickerson, Tom Goldstein

International Conference on Machine Learning (ICML) 2021

Data Augmentation for Meta-Learning

2021

Renkun Ni, **Micah Goldblum**, Amr Sharaf, Kezhi Kong, Tom Goldstein

International Conference on Machine Learning (ICML) 2021

The Intrinsic Dimension of Images and Its Impact on Learning

2021

Phil Pope, Chen Zhu, Ahmed Abdelkader, **Micah Goldblum**, Tom Goldstein

International Conference on Learning Representations (ICLR) 2021

**LowKey: Leveraging Adversarial Attacks to Protect Social Media Users
from Facial Recognition**

2021

Valeriia Cherepanova, **Micah Goldblum**, Harrison Foley, Shiyuan Duan,
John P Dickerson, Gavin Taylor, Tom Goldstein

International Conference on Learning Representations (ICLR) 2021

**Strong Data Augmentation Sanitizes Poisoning and Backdoor Attacks Without an
Accuracy Tradeoff**

2021

Eitan Borgnia, Valeriia Cherepanova, Liam Fowl, Amin Ghiasi, Jonas Geiping,
Micah Goldblum, Tom Goldstein, Arjun Gupta

International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2021

Robust Few-Shot Learning: A Meta-Learning Approach

2020

Micah Goldblum, Liam Fowl, Tom Goldstein

Advances in Neural Information Processing Systems (NeurIPS) 2020

Unraveling Meta-Learning: Understanding Feature Representations for Few-Shot Tasks

2020

Micah Goldblum, Steven Reich, Liam Fowl, Renkun Ni, Valeriia Cherepanova, Tom Goldstein

International Conference on Machine Learning (ICML) 2020.

Truth or backpropaganda? An empirical investigation of deep learning theory

2020

Micah Goldblum, Jonas Geiping, Avi Schwarzschild, Michael Moeller, Tom Goldstein

International Conference on Learning Representations (ICLR) 2020.

WITCHcraft: Efficient PGD attacks with random step size

2020

Ping-Yeh Chiang, Jonas Geiping, **Micah Goldblum**, Tom Goldstein, Renkun Ni,
Steven Reich, Ali Shafahi

International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2020.

Adversarially Robust Distillation

2020

Micah Goldblum, Liam Fowl, Soheil Feizi, Tom Goldstein

Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34.

SELECT INVITED TALKS

Jump Trading, AI Symposium Keynote Speaker

2025

Columbia University, Distinguished AI Lectures

2025

Princeton University, Algorithms and Machine Learning Seminar

2025

Columbia University, Statistics Seminar

2025

Two Sigma

2025

Columbia University, Conference on AI Agents for Work	2025
Cornell University, Department of Computer Science	2024
University of Pennsylvania, Department of Computer Science	2024
Georgia Institute of Technology, Department of Interactive Computing	2024
Northeastern University, Department of Computer Science	2024
University of Edinburgh, School of Informatics	2024
University of North Carolina Chapel Hill, Department of Computer Science	2024
University of British Columbia, Department of Computer Science	2024
Arthur Panel on AI and Science	2024
Chalmers AI Research Center Workshop for Structured Learning	2023
UCLA + MPI MIS Math Machine Learning Seminar	2023
Vanderbilt Machine Learning Seminar	2023
CISPA Helmholtz Center for Information Security	2023
University of California Los Angeles, Department of Mathematics	2023
ML Collective	2023
École Polytechnique Fédérale de Lausanne (EPFL)	2022
C3.ai Digital Transformation Institute: Open Problems in Robustness	2020

MEDIA COVERAGE

How close is AI to human-level intelligence? <i>Nature</i>	2024
LiveBench is an open LLM benchmark that uses contamination-free test data and objective scoring <i>Venture Beat</i>	2024
A New AI Detection Tool May Have Solved False Positives for Student Writing <i>Business Insider</i>	2024
AI Spam Threatens the Internet – AI Can Also Protect It <i>IEEE Spectrum</i>	2024
Researchers develop new LiveBench benchmark for measuring AI models response accuracy <i>Silicon Angle</i>	2024
Artists Are Slipping Anti-AI Poison into Their Art. Here’s How It Works <i>Scientific American</i>	2024
Binoculars is the most powerful AI text detector with over 90% accuracy <i>The Decoder</i>	2024
New AI detection tool measures how “surprising” word choices are <i>Freethink</i>	2024
AI and the future of work <i>Daily Maverick</i>	2023
Image-generating AI can copy and paste from training data, raising IP concerns <i>TechCrunch</i>	2022
How I Lost Control Over My Own Face <i>Der Spiegel Magazine</i>	2021
Cómo evitar que los sistemas de reconocimiento facial descifren las fotos de tus redes <i>El País</i>	2021
LowKey cool: This web app will tweak your photos to flummox facial-recognition systems, apparently <i>The Register</i>	2021

COMMUNITY SERVICE

- Chair of the organizing committee for the NeurIPS 2020 Workshop on Dataset Curation and Security.
- Organizer of the NeurIPS 2023 Workshop on Backdoors in Deep Learning,
- Organizer of the NeurIPS 2024 Workshop on Red Teaming GenAI: What Can We Learn from Adversaries?

- Organizer of the NeurIPS 2024 Workshop on Scientific Methods for Understanding Neural Networks: Discovering, Validating, and Falsifying Theories of Deep Learning with Experiments.
- Chair of the organizing committee for the ICLR 2025 workshop on Building Trust in LLMs and LLM Applications: From Guardrails to Explainability to Regulation
- Served as an Area Chair or Reviewer for conferences and journals including NeurIPS, ICML, ICLR, CVPR, and TPAMI.
- Member of the steering committee for Columbia University Foundations of Data Science Center