Hao Wang

314 Main St, Cambridge MA, 02142
✉ hao_wang@mit.edu & hao-wang@redhat.com 🌐 haowang94.github.io

Professional Experience

MIT-IBM Watson AI Lab **Boston, MA**
Senior Principal Research Scientist *Sept. 2022 – present*

- Senior Principal Research Scientist and Founding Member, Red Hat AI Innovation (Dec. 2024 – present)
- Research Scientist, IBM Research (Sept. 2022 – Nov. 2024)

Harvard University **Cambridge, MA**
Postdoctoral Fellow *July 2022 – Sept. 2022*

IBM Thomas J. Watson Research Center **Yorktown Heights, NY**
Research Intern *May 2019 – Aug. 2019*

Education

Harvard University **Cambridge, MA**
Ph.D. in Applied Mathematics *Sept. 2016 – May 2022*
Advisor: Flavio P. Calmon Committee: Demba Ba, Na Li, Salil Vadhan
Thesis: Information Theory for Trustworthy Machine Learning

Harvard University **Cambridge, MA**
M.S. in Applied Mathematics *Sept. 2016 – May 2019*

University of Science and Technology of China (USTC) **Hefei, China**
B.Sc. in Mathematics and Applied Mathematics *Sept. 2012 – July 2016*

Research Interests

Areas: Information Theory, Statistical Learning Theory, Optimization, Uncertainty Quantification
Topics: Efficient LLMs, AI Safety, Data Privacy, Personalization, Trustworthy ML, Inference-time Scaling

Awards, Honors, and Scholarships

NeurIPS Outstanding Reviewer Award (top 8% of reviewers)	*2021*
Harvard Certificate of Distinction in Teaching	*2021*
Winning Outreach Video at the ISIT Student Video Exposition (Video)	*2020*
ICML Travel Award	*2019*
Harvard Certificate of Distinction in Teaching	*2018*
The 35th Guo Moruo Scholarship (highest honor for USTC students)	*2015*
China National Scholarship	*2014*

Publications

Journal Publications

K. Alim*, **H. Wang***, O. Gulati, A. Srivastava, and N. Azizan, "Differentially Private Synthetic Data Generation for Relational Databases," *under review*, 2025. ***Equal contribution**.

H. Wang, R. Gao, and F. P. Calmon, "Generalization Bounds for Noisy Iterative Algorithms Using Properties of Additive Noise Channels," ***Journal of Machine Learning Research***, 2023.

H. Wang, H. Hsu, M. Diaz, and F. P. Calmon, "To Split or Not to Split: The Impact of Disparate Treatment in Classification," ***IEEE Transactions on Information Theory***, 2021.

H. Wang, L. Vo, F. P. Calmon, M. Médard, K. R. Duffy, and M. Varia, "Privacy with Estimation Guarantees," ***IEEE Transactions on Information Theory***, 2019.

M. Diaz*, **H. Wang***, F. P. Calmon, and L. Sankar, "On the Robustness of Information-Theoretic Privacy Measures and Mechanisms," ***IEEE Transactions on Information Theory***, 2019. ***Equal contribution**.

H. Wang, "Information Theory for Trustworthy Machine Learning," ***PhD Thesis***, 2022.

Peer-Reviewed Conference Proceedings

H. Wang, L. Han, K. Xu, and A. Srivastava, "SQuat: Subspace-orthogonal KV Cache Quantization," in *Conference on Language Modeling (**COLM**)*, 2025.

Y.-J. Park, K. Greenewald, K. Alim, **H. Wang**, and N. Azizan, "Know What You Don't Know: Uncertainty Calibration of Process Reward Models," in *Advances in Neural Information Processing Systems (**NeurIPS**)*, 2025.

M. Eyceoz, N. S. Nayak, **H. Wang**, L. Han, and A. Srivastava, "Hopscotch: Discovering and Skipping Redundancies in Language Models," in *Empirical Methods in Natural Language Processing (**EMNLP**) Findings, Short Paper*, 2025.

I. Ngong, S. Kadhe, **H. Wang**, K. Murugesan, J. D. Weisz, A. Dhurandhar, and K. N. Ramamurthy, "Protecting Users from Themselves: Safeguarding Contextual Privacy in Interactions with Conversational Agents," in *Annual Meeting of the Association for Computational Linguistics (**ACL**) Findings*, 2025.

A. Pareja, N. Nayak, **H. Wang**, K. Killamsetty, S. Sudalairaj, W. Zhao, S. Han, A. Bhandwaldar, G. Xu, K. Xu, L. Han, L. Inglis, and A. Srivastava, "Unveiling the Secret Recipe: A Guide For Supervised Fine-Tuning Small LLMs," in *International Conference on Learning Representations (**ICLR**)*, 2025.

K. Greenewald, Y. Yu, **H. Wang**, and K. Xu, "Privacy without Noisy Gradients: Slicing Mechanism for Generative Model Training," in *Advances in Neural Information Processing Systems (**NeurIPS**)*, 2024.

Y.-J. Park, **H. Wang**, S. Ardeshir, and N. Azizan, "Quantifying Representation Reliability in Self-Supervised Learning Models," in *Conference on Uncertainty in Artificial Intelligence (**UAI**)*, 2024.

H. Wang, S. Sudalairaj, J. Henning, K. Greenewald, and A. Srivastava, "Post-processing Private Synthetic Data for Improving Utility on Selected Measures," in *Advances in Neural Information Processing Systems (**NeurIPS**)*, 2023.

R. Feng, F. P. Calmon, and **H. Wang**, "Adapting Fairness Interventions to Missing Values," in *Advances in Neural Information Processing Systems (**NeurIPS**)*, 2023.

Y. Chen, W. Huang, **H. Wang**, C. Loh, A. Srivastava, L. M. Nguyen, and T.-W. Weng, "Analyzing Generalization of Neural Networks through Loss Path Kernels," in *Advances in Neural Information Processing Systems (**NeurIPS**)*, 2023.

H. Wang, L. He, R. Gao, and F. P. Calmon, "Aleatoric and Epistemic Discrimination: Fundamental Limits of Fairness Interventions," in *Advances in Neural Information Processing Systems (**NeurIPS**)*, 2023. **Spotlight**.

W. Alghamdi, H. Hsu, H. Jeong, **H. Wang**, P. W. Michalak, S. Asoodeh, and F. P. Calmon, "Beyond Adult and COMPAS: Fair Multi-Class Prediction via Information Projection," in *Advances in Neural Information Processing Systems (**NeurIPS**)*, 2022. **Oral presentation**.

H. Jeong, **H. Wang**, and F. P. Calmon, "Fairness without Imputation: A Decision Tree Approach for Fair Prediction with Missing Values," in *AAAI Conference on Artificial Intelligence (**AAAI**)*, 2022. **Oral presentation**.

H. Wang, Y. Huang, R. Gao, and F. P. Calmon, "Analyzing the Generalization Capability of SGLD Using Properties of Gaussian Channels," in *Advances in Neural Information Processing Systems (**NeurIPS**)*, 2021.

H. Wang, H. Hsu, M. Diaz, and F. P. Calmon, "The Impact of Split Classifiers on Group Fairness," in *IEEE International Symposium on Information Theory (**ISIT**)*, 2021.

W. Alghamdi, S. Asoodeh, **H. Wang**, F. P. Calmon, D. Wei, and K. N. Ramamurthy, "Model Projection: Theory and Applications to Fair Machine Learning," in *IEEE International Symposium on Information Theory (**ISIT**)*, 2020.

H. Wang, B. Ustun, and F. P. Calmon, "Repairing without Retraining: Avoiding Disparate Impact with Counterfactual Distributions," in *International Conference on Machine Learning (**ICML**)*, 2019.

H. Wang, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon, "An Information-Theoretic View of Generalization via Wasserstein Distance," in *IEEE International Symposium on Information Theory (**ISIT**)*, 2019.

H. Wang, M. Diaz, F. P. Calmon, and L. Sankar, "The Utility Cost of Robust Privacy Guarantees," in *IEEE International Symposium on Information Theory (**ISIT**)*, 2018.

H. Wang, B. Ustun, and F. P. Calmon, "On the Direction of Discrimination: An Information-Theoretic Analysis of Disparate Impact in Machine Learning," in *IEEE International Symposium on Information Theory (**ISIT**)*, 2018.

H. Wang and F. P. Calmon, "An Estimation-Theoretic View of Privacy," in *Annual Allerton Conference on Communication, Control, and Computing*, 2017.

Workshop Papers

I. Ngong, S. Kadhe, **H. Wang**, K. Murugesan, J. D. Weisz, A. Dhurandhar, and K. N. Ramamurthy, "Protecting users from themselves: Safeguarding contextual privacy in interactions with conversational agents," in *NeurIPS Workshop on Socially Responsible Language Modelling Research (SoLaR)*, 2024.

K. Alim*, **H. Wang***, O. Gulati, A. Srivastava, and N. Azizan, "Adapting differentially private synthetic data to relational databases," in *Theory and Practice of Differential Privacy Workshop (TPDP) and ISIT Workshop on Information-Theoretic Methods for Trustworthy Machine Learning*, 2024. ***Equal contribution**.

Y.-J. Park, **H. Wang**, S. Ardeshir, and N. Azizan, "Representation reliability for foundation models," in *Robotics Systems and Science (RSS) Safe Autonomy Workshop*, 2023. **Spotlight presentation**.

H. Jeong, **H. Wang**, and F. P. Calmon, "Fairness without Imputation: A Decision Tree Approach for Fair Prediction with Missing Values," in *Symposium on Foundations of Responsible Computing (FORC)*, 2022.

W. Alghamdi, H. Hsu, H. Jeong, **H. Wang**, P. W. Michalak, S. Asoodeh, and F. P. Calmon, "Beyond Adult and COMPAS: Fairness in Multi-Class Prediction," in *ICML Workshop on Responsible Decision Making in Dynamic Environments*, 2022.

H. Wang, H. Hsu, M. Diaz, and F. P. Calmon, "To Split or Not to Split: The Impact of Disparate Treatment in Classification," in *Symposium on Foundations of Responsible Computing (FORC)*, 2020.

H. Wang, B. Ustun, and F. P. Calmon, "Avoiding Disparate Impact with Counterfactual Distributions," in *NeurIPS Workshop on Ethical, Social and Governance Issues in AI*, 2018.

Pre-Prints and Technical Reports

G. Xu, K. Xu, S. Sudalairaj, **H. Wang**, and A. Srivastava, "Dr. sow: Density ratio of strong-over-weak llms for reducing the cost of human annotation in preference tuning," *arXiv preprint arXiv:2411.02481*, 2024.

H. Sun, N. Azizan, A. Srivastava, and **H. Wang**, "Private synthetic data meets ensemble learning," *arXiv preprint arXiv:2310.09729*, 2023.

News on My Research

My ISIT'18 paper, entitled "an information-theoretic analysis of disparate impact in ML", has been featured in an article "Just data: How algorithms go bad – and how they can be saved" at the Harvard GSAS alumni magazine.

My NeurIPS'23 paper on "differential privacy for synthetic data generation" has been featured in an article "An AI model trained on data that looks real but won't leak personal information" at the IBM Research blog.

My UAI'24 paper on "quantifying uncertainty of self-supervised learning models" has been featured in an article "How to assess a general-purpose AI model's reliability before it's deployed" at the MIT News.

My NeurIPS'25 paper on "LLM reasoning and inference-time scaling" has been featured in an article "A smarter way for large language models to think about hard problems" at the MIT News.

Research Funding

MIT-IBM Watson AI Lab grant: Quantifying Reliability/Uncertainty of Foundation Models (IBM PI). MIT PI: Navid Azizan, IBM Co-PI: Akash Srivastava, Kristjan Greenewald. 2023–2026.

MIT-IBM Watson AI Lab grant: Private Synthetic Data Generation: From Theoretical Foundations to Financial Applications (IBM PI). MIT PI: Navid Azizan, IBM Co-PI: Akash Srivastava. 2024–2025.

Mentorship

- Lisa Vo (Harvard College), 2017 – 2019
  Project: Privacy with estimation guarantees (published a paper in IEEE Trans. Inf. Theory).
- Winston Michalak (Harvard College), 2019 – 2020
  Project: Using ADMM for solving model projection (published a paper in NeurIPS 2022).
- Raymond Feng (Harvard College), 2022 – 2023
  Project: Investigating algorithmic discrimination in the presence of missing values (published a first-author paper in NeurIPS 2023, wrote a senior thesis).
- Luxi He (Harvard College), 2022 – 2023
  Project: Analyzing fairness-accuracy trade-offs in classification (published a paper in NeurIPS 2023, wrote a senior thesis).
- Haoyuan Sun (MIT PhD, summer intern at MIT-IBM), 2023
  Project: Private synthetic data meets ensemble learning.
- Ivoline Ngong (University of Vermont PhD, summer intern at IBM Research), 2024
  Project: Contextual privacy in interactions with conversational agents (presented a paper at NeurIPS 2024 SoLaR workshop, published a first-author paper in ACL (findings) 2025)
- Kaveh Alim (MIT PhD, summer intern at MIT-IBM), 2025
  Project: Differential private synthetic relational database & uncertainty quantification of foundation models.
- Shuhang Lin (Rutgers University PhD, summer intern at RedHat AI), 2025
  Project: KV cache compression & agent memory.

Teaching Experience

ES 250: Information Theory – Graduate Level Course *Fall 2022 and Fall 2024*
Engineering and Applied Sciences | Harvard University
Guest Lecturer
Gave a guest lecture that introduced a set of useful information-theoretic tools for trustworthy machine learning.

ES 201: Decision Theory – Graduate Level Course *Spring 2021*
Engineering and Applied Sciences | Harvard University ***Rating: 4.9/5.0***
Section Leader
Improved section (recitation) notes to include advanced topics. Led weekly sections that extended the lectures. Held weekly office hours to address individual questions and guided 20+ students through their final projects.

ES 156: Signals and Systems – Undergraduate Level Course *Spring 2018*
Engineering and Applied Sciences | Harvard University ***Rating: 4.8/5.0***
Section Leader
Created new section notes that complemented the lectures. Led weekly sections and office hours. Contributed to the design and grading of the midterm and final exams. Prepared new assignments, graded, and gave feedback to 20+ undergraduate students.

Selected Presentations

Random Sample AI Series at Red Hat *2025*

Conference on Neural Information Processing Systems (poster)	*2023*
MIT-IBM Watson AI Lab	*2022*
Yale Institute for Network Science (YINS)	*2022*
MIT Institute for Data, Systems, and Society (IDSS)	*2022*
Neural Information Processing Systems (virtual)	*2021*
IEEE International Symposium on Information Theory (virtual)	*2021*
Symposium on Foundations of Responsible Computing (virtual)	*2020*
IBM AI Systems Day	*2019*
International Conference on Machine Learning	*2019*
IEEE International Symposium on Information Theory	*2019*
North American School of Information Theory (poster)	*2019*
Annual New England Machine Learning Day (poster)	*2018*
Annual Allerton Conference on Communication, Control, and Computing	*2017*

Professional Service

Area Chair

- ACM Fairness, Accountability, and Transparency Conference (FAccT)
- International Conference on Learning Representations (ICLR)

Conference Reviewer

- IEEE International Symposium on Information Theory (ISIT)
- Neural Information Processing Systems (NeurIPS)
- International Conference on Machine Learning (ICML)
- International Conference on Artificial Intelligence and Statistics (AISTATS)
- IEEE Information Theory Workshop (ITW)
- The Web Conference (TheWebConf)

Journal Reviewer

- IEEE Transactions on Information Theory (T-IT)
- IEEE Transactions on Information Forensics & Security (T-IFS)
- IEEE Transactions on Automatic Control (TAC)
- IEEE Journal on Selected Areas in Information Theory (JSAIT)
- IEEE Journal of Selected Topics in Signal Processing (JSTSP)

Workshop Program Committee

- NeurIPS 2020 Workshop on Fair AI in Finance (FAIF)