Hongyi Wang

Assistant Professor

Piscataway, NJ
hw689@cs.rutgers.edu
hwang595.github.io

Positions

9/2025- **Assistant Professor** *CS Department*, Rutgers University.

8/2024- **Head of Infrastructure** *GenBio AI.*
Led the development of enterprise-scale DevOps, inference, and fine-tuning platforms for the company, serving both internal and external customers.

02/2023-7/2024 **Senior Research Scientist** *Carnegie Mellon University.*
Hosted by Eric P. Xing

09/2021-01/2023 **Postdoctoral Fellow** *Carnegie Mellon University.*
Hosted by Eric P. Xing

Research interests

I am interested in co-designing systems and algorithms for efficient and scalable large-scale machine learning. I am particularly interested in applying my research in the development and deployment of foundation models.

Education

2016–2021 **Ph.D. in Computer Science** *University of Wisconsin–Madison.*
Advisor: Dimitris Papailiopoulos

2016–2019 **M.S. in Computer Science** *University of Wisconsin–Madison.*

2012–2016 **B.S. in Electrical Enigneering** *Hangzhou Dianzi University.*

Publications

* stands for the joint first author. Here is my Google Scholar Profile.

[1] Haonian Ji, Shi Qiu, Siyang Xin, Siwei Han, Zhaorun Chen, Dake Zhang, **Hongyi Wang**, and Huaxiu Yao. From eduvisbench to eduvisagent: A benchmark and multi-agent framework for reasoning-driven pedagogical visualization. In *ICLR*, 2026.

[2] Zhengzhong Liu, Liping Tang, Linghao Jin, Haonan Li, Nikhil Ranjan, Desai Fan, Shaurya Rohatgi, Richard Fan, Omkar Pangarkar, Huijuan Wang, et al. K2-v2: A 360-open, reasoning-enhanced llm. *arXiv preprint arXiv:2512.06201*, 2025.

[3] Wenhao Zheng, Yixiao Chen, Weitong Zhang, Souvik Kundu, Yun Li, Zhengzhong Liu, Eric P Xing, **Hongyi Wang**, and Huaxiu Yao. Citer: Collaborative inference for efficient large language model decoding with token-level routing. *COLM*, 2025.

[4] Zhengzhong Liu, Bowen Tan, Hongyi Wang, Willie Neiswanger, Tianhua Tao, Haonan Li, Fajri Koto, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. Llm360 k2: Scaling up 360-open-source large language models. *arXiv e-prints*, pages arXiv–2501, 2025.

[5] Caleb N Ellington, Ning Sun, Nicholas Ho, Tianhua Tao, Sazan Mahbub, Dian Li, Yonghao Zhuang, Hongyi Wang, Le Song, and Eric P Xing. Accurate and general dna representations emerge from genome foundation models at scale. *bioRxiv*, pages 2024–12, 2024.

[6] Nicholas Ho, Caleb N Ellington, Jinyu Hou, Sohan Addagudi, Shentong Mo, Tianhua Tao, Dian

Li, Yonghao Zhuang, Hongyi Wang, Xingyi Cheng, et al. Scaling dense representations for single cell with transcriptome-scale context. *bioRxiv*, pages 2024–11, 2024.

[7] Ning Sun, Shuxian Zou, Tianhua Tao, Sazan Mahbub, Dian Li, Yonghao Zhuang, Hongyi Wang, Xingyi Cheng, Le Song, and Eric P Xing. Mixture of experts enable efficient and effective protein understanding and design. *bioRxiv*, pages 2024–11, 2024.

[8] Shuxian Zou, Tianhua Tao, Sazan Mahbub, Caleb N Ellington, Robin Algayres, Dian Li, Yonghao Zhuang, Hongyi Wang, Le Song, and Eric P Xing. A large-scale foundation model for rna function and structure prediction. *bioRxiv*, pages 2024–11, 2024.

[9] Xinyu Zhao, Guoheng Sun, Ruisi Cai, Yukun Zhou, Pingzhi Li, Peihao Wang, Bowen Tan, Yexiao He, Li Chen, Yi Liang, Beidi Chen, Binhang Yuan, **Hongyi Wang**, Ang Li, Zhangyang Wang, and Tianlong Chen. Model-glue: Democratized LLM scaling for a large model zoo in the wild. In *NeurIPS*, 2024.

[10] Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, **Hongyi Wang**, Lingjuan Lyu, and Ang Li. FLoRA: Federated fine-tuning large language models with heterogeneous low-rank adaptations. In *NeurIPS*, 2024.

[11] Tianhua Tao, Junbo Li, Bowen Tan, **Hongyi Wang**, William Marshall, Bhargav M Kanakiya, Joel Hestness, Natalia Vassilieva, Zhiqiang Shen, Eric P. Xing, and Zhengzhong Liu. Crystal: Illuminating LLM abilities on language and code. In *COLM*, 2024.

[12] Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, **Hongyi Wang**, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Timothy Baldwin, and Eric P. Xing. LLM360: Towards fully transparent open-source LLMs. In *COLM*, 2024.

[13] Samuel Horvath, Stefanos Laskaridis, Shashank Rajput, and **Hongyi Wang**. Maestro: Uncovering low-rank structures via trainable decomposition. *ICML*, 2024.

[14] Song Bian, Dacheng Li, **Hongyi Wang**, Eric Xing, and Shivaram Venkataraman. Does compressing activations help model parallel training? *MLSys*, 2024.

[15] **Hongyi Wang**, Felipe Maia Polo, Yuekai Sun, Souvik Kundu, Eric Xing, and Mikhail Yurochkin. Fusing models with complementary expertise. *ICLR*, 2024.

[16] Junbo Li, Ang Li, Chong Tian, Qirong Ho, Eric Xing, and **Hongyi Wang**. Fednar: Federated optimization with normalized annealing regularization. *NeurIPS*, 2023.

[17] **Hongyi Wang**, Saurabh Agarwal, Pongsakorn U-chupala, Yoshiki Tanaka, Eric Xing, and Dimitris Papailiopoulos. Cuttlefish: Low-rank model training without all the tuning. *MLSys*, 2023.

[18] Dacheng Li*, Rulin Shao*, **Hongyi Wang***, Han Guo, Eric Xing, and Hao Zhang. Mpcformer: fast, performant and private transformer inference with mpc. *ICLR (**Spotlight**)*, 2023.

[19] Han Guo, Philip Greengard, **Hongyi Wang**, Andrew Gelman, Eric Xing, and Yoon Kim. Federated learning as variational inference: A scalable expectation propagation approach. *ICLR*, 2023.

[20] Kai Zhang, Yu Wang, **Hongyi Wang**, Lifu Huang, Carl Yang, and Lichao Sun. Efficient federated learning on knowledge graphs via privacy-preserving relation embedding aggregation. *Findings of EMNLP*, 2022.

[21] Kartik Sreenivasan, Jy-yong Sohn, Liu Yang, Matthew Grinde, Alliot Nagle, **Hongyi Wang**, Eric Xing, Kangwook Lee, and Dimitris Papailiopoulos. Rare gems: Finding lottery tickets at initialization. *NeurIPS*, 2022.

[22] Dacheng Li, **Hongyi Wang**, Eric Xing, and Hao Zhang. Amp: Automatically finding model parallel strategies with heterogeneity awareness. *NeurIPS*, 2022.

[23] Saurabh Agarwal, **Hongyi Wang**, Shivaram Venkataraman, and Dimitris Papailiopoulos. On the utility of gradient compression in distributed training systems. *MLSys*, 2022.

[24] **Hongyi Wang**, Saurabh Agarwal, and Dimitris Papailiopoulos. Pufferfish: Communication-efficient models at no extra cost. *MLSys*, 2021.

[25] Saurabh Agarwal, **Hongyi Wang**, Kangwook Lee, Shivaram Venkataraman, and Dimitris Papailiopoulos. Accordion: Adaptive gradient communication via critical learning regime identification. *MLSys*, 2021.

[26] Chaoyang He, Songze Li, Jinhyun So, Mi Zhang, **Hongyi Wang**, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, Li Shen, Peilin Zhao, Yan Kang, Yang Liu, Ramesh Raskar, Qiang Yang, Murali Annavaram, and Salman Avestimehr. *NeurIPS SpicyFL workshop*.

[27] **Hongyi Wang**, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. *NeurIPS*, 2020.

[28] **Hongyi Wang**, Mikhail Yurochkin, YueKai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *ICLR*, 2020.

[29] Shashank Rajput*, **Hongyi Wang***, Zachary Charles, and Dimitris Papailiopoulos. Detox: A redundancy-based framework for faster and more robust gradient aggregation. *NeurIPS*, 2019.

[30] Lingjiao Chen, **Hongyi Wang**, Leshang Chen, Paraschos Koutris, and Arun Kumar. Demonstration of nimbus: Model-based pricing for machine learning in a data marketplace. In *SIGMOD 2019*, pages 1885–1888. ACM, 2019.

[31] **Hongyi Wang***, Scott Sievert*, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. In *NeurIPS*, 2018.

[32] Lingjiao Chen, **Hongyi Wang**, Jinman Zhao, Dimitris Papailiopoulos, and Paraschos Koutris. The effect of network width on the performance of large-batch training. In *NeurIPS*, 2018.

[33] Lingjiao Chen, **Hongyi Wang**, Zachary Charles, and Dimitris Papailiopoulos. Draco: Byzantine-resilient distributed training via redundant gradients. In *ICML*, 2018.

[34] Lingjiao Chen, **Hongyi Wang**, and Dimitris Papailiopoulos. Draco: Robust distributed training against adversaries. In *SysML*, 2018.

[35] Guru Subramani, Daniel Rakita, **Hongyi Wang**, Jordan Black, Michael Zinn, and Michael Gleicher. Recognizing actions during tactile manipulations through force sensing. In *IROS*, pages 4386–4393. IEEE, 2017.

Honors & Awards

AMD University Program AI & HPC Cluster Allocation Award *2026.*

Best Demo Paper Runner Up *NAACL 2024.*

The Rising Stars Award of Conference on Parsimony and Learning (CPAL) 2024.

Student Travel Award *ICML 2018, NeurIPS 2018, 2019, MLSys 2022.*

The Baidu Best Paper Award *SpicyFL workshop at NeurIPS 2020.*

Top Reviewer Awards *NeurIPS 2019, ICML 2020.*

National Scholarship of China (Top 2%).

Students Advising

Huanwei Di (Ph.D. student at Rutgers CS).

Daize Dong (Ph.D. student at Rutgers CS).

Haolong Jia (Ph.D. student at Rutgers CS).

Jiawei Wu (Ph.D. student at Rutgers CS).

Junlin Chen (Undergrad. student at Rutgers).

Ryan Cheng (Undergrad. student at Rutgers).

Professional Service

Area Chair: *NeurIPS 2026, MLSys 2025, CPAL 2025.*

Program committee: *DAC 2024, EuroSys 2024, SOSP 2023 (light PC), MLSys 2023-25, MLSys 2022 (Artifact Evaluation Committee), SIGKDD 2022-23, AAAI 2021-22.*

Reviewer (journal): *JMLR, TMLR, IEEE TNNLS, IEEE IoT-J, IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Reviewer (conference): *SC 2026, ICML 2019-25, NeurIPS 2019-26, ICLR 2021-26, CVPR 2021-24, ICCV 2021-23.*

Conference session chair: *Tutorial session ICML 2022, Federated learning session MLSys 2023, LLM and Diffusion Model Serving session MLSys 2025.*

Workshop Organizer: *Scalable Learning and Optimization for Efficient Multimodal AI Agents (SCALE) Workshop @ ICML 2026, Federated Learning Systems (FLSys) Workshop @ MLSys 2023.*

Teaching Experience

Spring 2026 **Rutgers** *439 Introduction to Data Science.*

Fall 2025 **Rutgers** *671 Recent Advances in Large Language Model.*