Hongyi Wang

Head of Infrastructure

GenBio.ai
✉ *hongyi.wang@genbio.ai*
hwang595.github.io

Positions

8/2024- **Head of Infrastructure** *GenBio.ai*.

10/2023-7/2024 **Senior Project Scientist** *Carnegie Mellon University*.
Hosted by Eric P. Xing

02/2023-09/2023 **Senior Researcher** *Carnegie Mellon University*.
Hosted by Eric P. Xing

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

Summer 2020 **Research Intern** *Microsoft, DeepSpeed Team*.
Hosted by Minjia Zhang and Yuxiong He

Summer 2019 **Research Intern** *IBM Research*.
Hosted by Mikhail Yurochkin and Yasaman Khazaeni

Research interests

I am interested in co-designing systems and algorithms for efficient and trustworthy large-scale machine learning. I am particularly interested in applying my research in the development and deployment of foundation models, such as GPT and LLaMA.

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] 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.

[2] 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.

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

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

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

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

[7] **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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

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

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

[16] 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*.

[17] **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.

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

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

[20] 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.

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

[22] 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.

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

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

[25] 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.

Grants

NSF CNS2414087 (Senior Personnel, PI: Eric P. Xing) "CSR: RI: Small: Sustainable Large Scale Machine Learning via Multi-Level Optimization on Algorithm, System, and Meta Learning", 10/01/2024-09/31/2027.

NSF IIS2311990 (Senior Personnel, PI: Eric P. Xing) "III: Small: Multiple Device Collaborative Learning in Real Heterogeneous and Dynamic Environments", 09/01/2023-08/31/2026.

Semiconductor Research Corp. Artificial Intelligence Hardware Program (Project Co-lead, PI: Eric P. Xing) "Co-designing Distributed ML Systems and Algorithms for Foundation Models for AI-for-Science", 01/01/2024-12/31/2026.

Honors & Awards

2024 **Best Demo Paper Runner Up** *NAACL 2024.*

2024 **The Rising Stars Award at the Conference on Parsimony and Learning (CPAL)**.

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

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

2020 **Top Reviewer Award** *ICML 2020.*

2019 **Top Reviewer Award** *NeurIPS 2019.*

2015 **National Scholarship of China (Top 2%)**.

Open-Source Projects

LLM360 *An initiative to fully open-source LLMs, which advocates for all training code and data, model checkpoints, and intermediate results to be made available. The goal of LLM360 is to support open and collaborative AI research by making the end-to-end LLM pre-training process transparent and reproducible by everyone.*

Mentoring

Zheyu Shen (Ph.D. student at University of Maryland ECE), co-advised with Prof. Ang Li.

Han Guo (Ph.D. student at CMU LTI).

Jinyu Hou (M.Sc. student at CMU MLD).

Junbo Li (Research intern at MBZUAI).

Dacheng Li (M.Sc. student at CMU), now a Ph.D. student at EECS at UC Berkeley.

Rulin Shao (M.Sc. student at CMU), now a Ph.D. student at UWashington CS.

Professional Service

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): *ICML 2019-23, NeurIPS 2019-24, ICLR 2021-24, CVPR 2021-24, ICCV 2021-23.*

Conference session chair: *Tutorial session ICML 2022, Federated learning session MLSys 2023.*

Workshop Organizer: *Federated Learning Systems (FLSys) Workshop @ MLSys 2023.*

Teaching Experience

Spring 2023 **Guest Lecturer** *MBZUAI ML710: Parallel and Distributed ML Systems.*

Fall 2022 **Guest Lecturer** *MBZUAI ML710: Parallel and Distributed ML Systems.*

Spring 2022 **Guest Lecturer** *UW-Madison ECE826: Theoretical Foundations of Large-scale ML.*