Junyuan "Jason" Hong

University of Texas at Austin

+1 (517) 668-5790 jyhong@utexas.edu https://jyhong.gitlab.io

I am a joint postdoctoral fellow in the Institute for Foundations of Machine Learning (IFML) and Wireless Networking and Communications Group (WNCG), and also affiliated with the UT AI Health Lab as well as the Good System Challenge. I was recognized as one of the MLSys Rising Stars in 2024 and received a Best Paper Nomination at VLDB 2024. Part of my work is funded by OpenAI Research Access Program.

Education and Professional Experience

Postdoctoral Fellow, IFML, UT Austin 2023 - Present

- Supervisor: Dr. Zhangyang "Atlas" Wang
- Main Collaborators: Dr. Bo Li (University of Chicago), Dr. Ying Ding (UT Austin), and Dr. Hiroko H. Dodge (Havard Medical School)
- Lead multiple GenAI projects (funded by NSF, the state of Texas, and OpenAI), focusing on algorithmic and theoretical advancements of LLMs, as well as innovative GenAI applications in healthcare.

Ph.D. in Computer Science and Engineering 2018 - 2023
Michigan State University, East Lansing, MI, USA

- Supervisor: Dr. Jiayu Zhou (currently at University of Michigan)
- Committee: Dr. Anil K. Jain (NAE Member), Dr. Sijia Liu, Dr. Atlas Wang, Dr. Jiayu Zhou
- Thesis: Data-Centric Privacy-Preserving Machine Learning

Research Intern, Sony AI, NY, USA Feb - Aug 2022

- Mentors: Dr. Micheal Spranger, Dr. Lingjuan Lyu
- Work on privacy-preserving, low-computation training algorithms, and memory-efficient model adaptation algorithms for dynamical test-time environments, both optimized for edge devices.

M.S. in Computer Science 2015 - 2018

B.S. in Physics & Computer Science (double major) 2011 - 2015
University of Science and Technology of China, Hefei, China

Research Interests

AI Trust: Algorithms, Theories, and Benchmarks for assessing the trustworthiness of AI/ML systems, encompassing privacy, security, fairness, robustness, and efficiency, with a recent focus on **GenAI**.

Human-centric AI for Healthcare Applications: Trustworthy and affordable medical predictive modeling and intervention for cognition dementia diseases, leveraging recent GenAI advances.

Selected Honors & Awards

2024 **Best Paper Nomination, VLDB 2024**

ML and Systems Rising Stars, ML Commons 2024

2023 **3rd place Winner, US-UK Privacy-Enhancing Technologies (PETs) Challenge**

- Press Release by **The White House**: "At Summit for Democracy, the United States and the United Kingdom Announce Winners of Challenge to Drive Innovation in Privacy-enhancing Technologies That Reinforce Democratic Values"

- University Headline Coverage: "Privacy-enhancing Research Earns International Attention", by MSU Engineering News and MSU Office Of Research And Innovation

Dissertation Completion Fellowship, Michigan State University

Research Enhancement Award, Michigan State University

Top Reviewer, NeurIPS 2023

2021 Carl V. Page Memorial Graduate Research Fellowship, Michigan State University

Selected Publications

My research vision is to **harmonize, understand, and deploy Responsible AI**. As of Oct 2024, my work has amassed **1,100+** citations, with H-index = **12** [Google Scholar]. My selected publications are organized in three thrusts below († indicates the equal contribution, * indicates (co-)advised students.)

(1) Harmonizing Multifaceted AI Trusts: Privacy, Fairness, Robustness, and Efficiency

[NACCL'25] Runjin Chen, Gabriel Jacob Perin, Xuxi Chen, Xilun Chen, Yan Han, Nina S. T. Hirata, **Junyuan Hong**, Bhavya Kailkhura. Extracting and Understanding the Superficial Knowledge in Alignment. *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics.*

[ICLR'24] **Junyuan Hong**, Jiachen T. Wang, Chenhui Zhang, Zhangheng Li, Bo Li, Zhangyang Wang. DP-OPT: Make Large Language Model Your Privacy-Preserving Prompt Engineer. *Proceedings of International Conference on Learning Representations* (**Spotlight, top-5%**).

[ICLR'24] Shuyang Yu*, **Junyuan Hong**, Haobo Zhang, Haotao Wang, Zhangyang Wang, Jiayu Zhou. Safe and Robust Watermark Injection with a Single OoD Image. *Proceedings of the Twelfth International Conference on Learning Representations.*

[ICML'23] **Junyuan Hong**†, Yi Zeng†, Shuyang Yu$^{\dagger *}$, Lingjuan Lyu, Ruoxi Jia, Jiayu Zhou. Revisiting Data-Free Knowledge Distillation with Poisoned Teachers. *Proceedings of Fortieth International Conference on Machine Learning.*

[ICLR'23] Shuyang Yu*, **Junyuan Hong**, Haotao Wang, Zhangyang Wang and Jiayu Zhou. Turning the Curse of Heterogeneity in Federated Learning into a Blessing for Out-of-Distribution Detection. *Proceedings of the Eleventh International Conference on Learning Representations* (**Spotlight, top-5%**).

[ICLR'23] **Junyuan Hong**, Lingjuan Lyu, Jiayu Zhou, Michael Spranger. MECTA: Memory-Economic Continual Test-Time Model Adaptation. *Proceedings of the Eleventh International Conference on Learning Representations.*

[TMLR'23] Haotao Wang, **Junyuan Hong**, Jiayu Zhou, and Zhangyang Wang. How Robust is Your Fairness? Evaluating and Sustaining Fairness under Unseen Distribution Shifts. *Transactions on Machine Learning Research.*

[AAAI'23] **Junyuan Hong**, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Federated Robustness Propagation: Sharing Adversarial Robustness in Heterogeneous Federated Learning. *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence.*

[NeurIPS'22] **Junyuan Hong**, Lingjuan Lyu, and Jiayu Zhou, Micheal Spranger. Outsourcing Training without Uploading Data via Efficient Collaborative Open-Source Sampling. *Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems.*

[NeurIPS'22] Haotao Wang, **Junyuan Hong**, Aston Zhang, Jiayu Zhou and Zhangyang Wang. Trap and Replace: Defending Backdoor Attacks by Trapping Them into an Easy-to-Replace Subnetwork. *Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems.*

[ICLR'22] **Junyuan Hong**, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Efficient Split-Mix Federated Learning for On-Demand and In-Situ Customization. *Proceedings of the Tenth International Conference on Learning Representations*

[FAccT'22] **Junyuan Hong**, Zhangyang Wang, and Jiayu Zhou. Dynamic Privacy Budget Allocation Improves Data Efficiency of Differentially Private Gradient Descent. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.*

[AAAI'21] **Junyuan Hong**, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Learning model-based privacy protection under budget constraints. *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence.*

(2) Understanding Multi-faceted Emerging Risks in GenAI Trust

[ICML'24] **Junyuan Hong**[†], Jinhao Duan[†], Chenhui Zhang[†], Zhangheng Li[†], Chulin Xie, Kelsey Lieberman, James Diffenderfer, Brian Bartoldson, Ajay Jaiswal, Kaidi Xu, Bhavya Kailkhura, Dan Hendrycks, Dawn Song, Zhangyang Wang, Bo Li. Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression. *The Forty-first International Conference on Machine Learning.*

[VLDB'24] Qinbin Li[†], **Junyuan Hong**[†], Chulin Xie[†], Jeffrey Tan[*], Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, Bo Li, Bingsheng He, Dawn Song. LLM-PBE: Assessing Data Privacy in Large Language Models. *International Conference on Very Large Data Bases.* (**Best Paper Nomination**)

[SaTML'24] Zhangheng Li[*], **Junyuan Hong**, Bo Li, Zhangyang Wang. Shake to Leak: Amplifying the Generative Privacy Risk through Fine-Tuning. *The 2nd IEEE Conference on Secure and Trustworthy Machine Learning.*

[ICML'24] Yihua Zhang[†], Pingzhi Li[†], **Junyuan Hong**[†], Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D. Lee, Wotao Yin, Mingyi Hong, Zhangyang Wang, Sijia Liu, Tianlong Chen. Revisiting Zeroth-Order Optimization for Memory-Efficient LLM Fine-Tuning: A Benchmark. *The Forty-first International Conference on Machine Learning.*

[NeurIPS'23] Haobo Zhang[*†], **Junyuan Hong**[†], Yuyang Deng, Mehrdad Mahdavi, Jiayu Zhou. Understanding Deep Gradient Leakage via Inversion Influence Functions. *Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems.*

[arXiv'24] Zhen Xiang, Linzhi Zheng, Yanjie Li, **Junyuan Hong**, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Carl Yang, Dawn Song., Bo Li. GuardAgent: Safeguard LLM Agents by a Guard Agent via Knowledge-Enabled Reasoning. *arXiv Preprint.*

(3) Deploying AI Aligned with Human Norms in High-Stake Applications

[arXiv'25] Shrey Pandit, Jiawei Xu, **Junyuan Hong**, Zhangyang Wang, Tianlong Chen, Kaidi Xu, Ying Ding. MedHallu: A Comprehensive Benchmark for Detecting Medical Hallucinations in Large Language Models. *arXiv Preprint.*

[NACCL'25] Jinhao Duan, Xinyu Zhao, Zhuoxuan Zhang, Chenan Wang, Tianhao Li, Alexander Rasgon, **Junyuan Hong**, Qi Long, Ying Ding, Tianlong Chen, and Kaidi Xu. An Exploration of LLM-Guided Conversation in Reminiscence Therapy. *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*

[ICLRW'24] **Junyuan Hong**†, Wenqing Zheng†, Han Meng, Siqi Liang, Anqing Chen, Hiroko Dodge, Jiayu Zhou, Zhangyang Wang. A-CONECT: Designing AI-based Conversational Chatbot for Early Dementia Intervention. *ICLR 2024 Workshop on LLM Agents.*

[🏆KDDW'23] Haobo Zhang*, **Junyuan Hong**, Fan Dong, Steve Drew, Liangjie Xue, Jiayu Zhou. A Privacy-Preserving Hybrid Federated Learning Framework for Financial Crime Detection. *International Workshop on Federated Learning for Distributed Data Mining.* (**3rd Place Winner at PETs Prize Challenge**)

[KDD'21] **Junyuan Hong**, Zhuangdi Zhu, Shuyang Yu*, Zhangyang Wang, Hiroko Dodge, and Jiayu Zhou. Federated Adversarial Debiasing for Fair and Transferable Representations. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

[AD'20] **Junyuan Hong**, Jeffrey Kaye, Hiroko H Dodge, Jiayu Zhou. Detecting MCI Using Real-time, Ecologically Valid Data Capture Methodology: How to Improve Scientific Rigor in Digital Biomarker Analysis. *Alzheimer's & Dementia*

Teaching Experiences

Fall 2024	Project Mentor, "VRT-CHAT: Designing Reminiscence-Therapy Chatbots with Culturally-Sensitive Visual Stimulation for Mental Health", RAI4Ukraine Program
	Project Mentor, "Designing AI-based Conversational Chatbot for Early Dementia Intervention" (undergraduate level), Directed Reading Program (DiRP), UT Austin CS
Fall 2023	Project Mentor, "Trustworthy Large Language Models" (undergraduate level), Directed Reading Program (DiRP), UT Austin CS
Spring 2023	Invited Lectures on privacy and federated learning at "CSE847: Machine Learning" (graduate level), Michigan State University CSE
Spring 2021	Teaching Assistant and Invited Lecture, "CSE847: Machine Learning" (graduate level), Michigan State University CSE
Fall 2020	Teaching Assistant, "CSE404: Introduction to Machine Learning" (undergraduate level), Michigan State University CSE

Invited Talks & Guest Lectures

2024	(Invited Talk) Harmonizing, Understanding, and Deploying Responsible AI: A Tale About Privacy. CS@University of Maryland & CS@Rutgers University, October, 2024
	(Invited Talk) GenAI-Based Chatbot for Early Dementia Intervention. *Rising Star Symposium*, IEEE TCCN Special Interest Group for AI/ML in Security, September, 2024
	(Guest Lecture) Building Conversational AI for Affordable and Accessible Early Dementia Intervention. *AI Health* Course, iSchool@UT Austin, April, 2024
	(Invited Talk) Amplifying the Generative Privacy Risk through Fine-Tuning. *Good Systems Symposium: Shaping the Future of Ethical AI*, UT Austin, March, 2024
2023	(Guest Lecture) Foundation Models Meet Data Privacy: Risks and Countermeasures. *Trustworthy Machine Learning* Course, CS@Virginia Tech, November, 2023
	(Invited Talk) Backdoor Meets Data-Free Learning. *TMLR Group*, Hong Kong Baptist University, September, 2023

(Invited Talk) MECTA: Memory-Economic Continual Test-Time Model Adaptation. *Computer Vision Talks* (YouTube webinar), April, 2023

(Invited Talk) Split-Mix Federated Learning for Model Customization, *TrustML Young Scientist Seminars*, RIKEN & University of Tokyo, Japan, July, 2022

(Invited Talk) Efficient Split-Mix Federated Learning for On-demand and In-situ Model Customization, *Sony AI*, Virtual, February, 2022.

Professional Activities

- Lead Program Chair at the NeurIPS 2024 Workshop on GenAI for Health: Potential, Trust and Policy Compliance (genai4health.github.io/).
- Co-Organizer at NeurIPS 2024 LLM Privacy Challenge (llm-pc.github.io/)
- Co-Organizer at NeurIPS 2024 LLM & Agent Safety Competition (www.llmagentsafetycomp24.com)
- Lead Program Chair at ACM SIGKDD 2024 Workshop on Federated Learning for Data Mining and Graph Analytics (FedKDD) (fedkdd.github.io)
- Lead Program Chair at the ACM SIGKDD 2023 Workshop on Federated Learning for Distributed Data Mining (fl4data-mining.github.io)
- Technical Program Committee: NeurIPS (2022 - 2024), ICLR (2022 - 2024), ICML (2022 - 2024), KDD (2022- 2024), AISTATS (2022 - 2023), WSDM 2022, AAAI (2021 - 2023)
- Journal Reviewer: IEEE TKDE, JAIR, ACM TKDD, ACM Health, Neurocomputing

Grant Writing Experience

- *NAIRR Pilot: Developing Engaging AI Chatbots to Enhance Senior Well-being*, ($25k + 400 hrs of GH100, **Co-PI** with PI: Dr. Zhuangdi Zhu)
- *AI-based conversational engagement for dementia healthcare*, OpenAI Research Access ($8k, **sole PI**)
- *NSF III: Medium: A consolidated framework of computational privacy and machine learning* ($266k, **primary proposal writing assistant** to PI Dr. Atlas Wang)
- *AI-CONECT: Conversational AI for Early Dementia Prevention in Socially-Isolated Senior Adults*, A2 Pilot Award funded by the National Institute on Aging (NIA) (**lead PI**, under Round 2 review)
- *Remi: Reminiscence Therapy by the Retrieval-Augmented LLM Chatbot for the Memory Reactivation of Socially Isolated Senior Immigrants*, UT Austin IC2 Grant (**co-PI**, under Phase 2 review)

Student Advising

2023 - Now	**Zhangheng Li**, Ph.D. student, University of Texas at Austin SaTML 2024 (first author), ICML 2024 (co-first author), ICLR 2024
2023 - Now	**Runjin Chen**, Ph.D. student, University of Texas at Austin NACCL 2025 (first author)

2023 - Now	**Gabriel Jacob Perin**, Undergraduate student, University of São Paulo, Brazil NACCL 2025
2023 - 2024	**Jeffrey Tan**, Undergraduate student, University of California, Berkeley VLDB 2024 (Best Paper Nomination)
2020 - 2023	**Shuyang Yu**, Ph.D. student, Michigan State University ICLR 2024 (first author), ICLR 2023 (spotlight; first author), NeurIPSW 2023 (first author), ICML 2023, KDD 2021
2022 - 2023	**Haobo Zhang**, Ph.D. student, Michigan State University NeurIPS 2023 (first author), KDDW 2023 (first author) Team member, 3rd place winner at US-UK PETs (Privacy-enhancing technologies) Prize Challenge, 2023.
2022 - 2023	**Siqi Liang**, Ph.D. student, Michigan State University KDDW 2023 (first author)