Jianing Zhu

University of Texas at Austin, Austin, Texas

Homepage | Google Scholar | jianing.zhu@austin.utexas.edu | +1 412-475-5044

I am a Postdoctoral Fellow at the VITA Group in the Chandra Family Department of Electrical and Computer Engineering, UT Austin, working with Prof. Atlas Wang on the reliability of evolving AI systems and trustworthy ML, and am also affiliated with the Good Systems Challenge for Ethical AI at UT Austin. I received my PhD from the TMLR Group in the Department of Computer Science, HKBU, advised by Prof. Bo Han.

RESEARCH INTERESTS

My research aims to make AI systems reliable as they undergo degradation over their operational lifetime. As AI shifts from static models to continually evolving systems, reliability is no longer a property to be verified once, but one that erodes as failures accumulate and compound. I study when and why this degradation becomes self-reinforcing and irreversible, and how systems can be designed to resist it, diagnose it, and govern what they retain and expose. My goal is to make resilience a first-class property of evolving AI systems, so that they can be responsibly and safely deployed for societal benefit in high-stakes, real-world applications.

PROFESSIONAL EXPERIENCE

Postdoctoral Fellow, The University of Texas at Austin — **Sept. 2025 – Present**
VITA Group, Advisor: Prof. Zhangyang "Atlas" Wang — Austin

Visiting PhD Student, Carnegie Mellon University — **Jan. 2025 – Jun. 2025**
Neuro-Symbolic AI Group, Advisor: Prof. Pradeep Ravikumar — Pittsburgh

Research Intern, RIKEN Center for Advanced Intelligence Project — **Dec. 2023 – May 2024**
Imperfect Information Learning Team, Advisor: Prof. Masashi Sugiyama — Tokyo

EDUCATION

Hong Kong Baptist University (HKBU) — **Sept. 2021 – Jun. 2025**
Ph.D. in TMLR Group, Department of Computer Science — Advisor: Prof. Bo Han

Sichuan University (SCU) — **Sept. 2017 – Jun. 2021**
B.Eng. in CS, College of Computer Science — National Top-Notch Undergraduate Program

HONORS AND AWARDS

- TrustAI Rising Star Award — 2025
- NeurIPS Top Reviewer Award — 2023, 2025
- ICML Best Reviewer Award — 2021, 2024
- ICLR Notable Reviewer Award — 2025
- Yakun Scholarship Scheme for Mainland Postgraduate Students — 2024
- Computer Science Department Research Excellence Award at HKBU — 2023
- Computer Science Department RPg Performance Award at HKBU — 2022-2024
- Excellent Teaching Assistant Performance Award at HKBU — 2022-2023
- Nominated for the Hong Kong PhD Fellowship Scheme — 2021
- National Scholarship from the Ministry of Education — 2018

SELECTED PUBLICATIONS

The full list can be found on [Google Scholar]. Works grouped by theme (* indicates equal contribution):

Building Robustness Against Degradation

1. **ICML 2026** [link]**:** Jingwei Sun, **Jianing Zhu**, Yuanyi Li, Tongliang Liu, Xia Hu, Bo Han, "`AgentHijack: Benchmarking Computer Use Agent Robustness to Common Environment Corruptions`".

2. **ICLR 2026** [link]**:** Jingwei Sun*, **Jianing Zhu***, Jiangchao Yao, Gang Niu, Masashi Sugiyama, Bo Han, "`Bilateral Information-aware Test-time Adaptation for Vision-Language Models`".

3. **ICLR 2026** [link]: Zizhuo Zhang*, **Jianing Zhu***, Xinmu Ge*, Zihua Zhao*, Zhanke Zhou, Xuan Li, Xiao Feng, Jiangchao Yao, Bo Han, "`Co-rewarding: Stable Self-supervised RL for Eliciting Reasoning in Large Language Models`".

4. **TPAMI 2025** [link]: **Jianing Zhu**, Bo Han, Jiangchao Yao, Tongliang Liu, Quanming Yao, Jianliang Xu, "`Slack Federated Adversarial Training`".

5. **ICLR 2023** [link]: **Jianing Zhu**, Jiangchao Yao, Tongliang Liu, Quanming Yao, Jianliang Xu, Bo Han, "`Combating Exacerbated Heterogeneity for Robust Models in Federated Learning`".

6. **NeurIPS 2022 (Spotlight)** [link]: Jianan Zhou*, **Jianing Zhu***, Jingfeng Zhang, Tongliang Liu, Gang Niu, Bo Han, Masashi Sugiyama, "`Adversarial Training with Complementary Labels: On the Benefit of Gradually Informative Attacks`".

7. **ICLR 2022** [link]: **Jianing Zhu**, Jiangchao Yao, Bo Han, Jingfeng Zhang, Tongliang Liu, Gang Niu, Jingren Zhou, Jianliang Xu, Hongxia Yang, "`Reliable Adversarial Distillation with Unreliable Teachers`".

8. **ICLR 2021 (Oral)** [link]: Jingfeng Zhang, **Jianing Zhu**, Gang Niu, Bo Han, Masashi Sugiyama, Mohan Kankanhalli, "`Geometry-aware Instance-reweighted Adversarial Training`".

Diagnosing Failure and Degradation

1. **arXiv 2026** [link]: **Jianing Zhu***, Yeonju Ro*, John Robertson, Kevin Wang, Junbo Li, Haris Vikalo, Aditya Akella, Zhangyang Wang, "`Your Agents Are Aging Too: Agent Lifespan Engineering for Deployed Systems`".

2. **arXiv 2026** [link]: Boxuan Zhang*, **Jianing Zhu***, Zeru Shi, Dongfang Liu, Ruixiang Tang, "`AgentForesight: Online Auditing for Early Failure Prediction in Multi-Agent Systems`".

3. **NeurIPS 2024** [link]: Boxuan Zhang*, **Jianing Zhu***, Tongliang Liu, Masashi Sugiyama, "`What If the Input is Expanded in OOD Detection?`".

4. **NeurIPS 2024** [link]: Geng Yu, **Jianing Zhu**, Jiangchao Yao, Bo Han, "`Self-Calibrated Tuning of Vision-Language Models for Out-of-Distribution Detection`".

5. **NeurIPS 2023** [link]: **Jianing Zhu**, Geng Yu, Jiangchao Yao, Tongliang Liu, Gang Niu, Masashi Sugiyama, Bo Han, "`Diversified Outlier Exposure for Out-of-Distribution Detection via Informative Extrapolation`".

6. **ICML 2023** [link]: **Jianing Zhu**, Hengzhuang Li, Jiangchao Yao, Tongliang Liu, Jianliang Xu, Bo Han, "`Unleashing Mask: Explore the Intrinsic Out-of-Distribution Detection Capability`".

7. **ICML 2023** [link]: **Jianing Zhu**, Xiawei Guo, Jiangchao Yao, Chao Du, Li He, Shuo Yuan, Tongliang Liu, Liang Wang, Bo Han, "`Exploring Model Dynamics for Accumulative Poisoning Discovery`".

8. **NeurIPS 2023 Workshop** [link]: Xuan Li*, Zhanke Zhou*, **Jianing Zhu***, Jiangchao Yao, Tongliang Liu, Bo Han, "`DeepInception: Hypnotize Large Language Model to Be Jailbreaker`" (400+ citations).

Governing What Models Retain

1. **ICLR 2026** [link]: **Jianing Zhu**, Bo Han, Jiangchao Yao, Jianliang Xu, Gang Niu, Masashi Sugiyama, "`Decoupling the Class Label and the Target Concept in Machine Unlearning`".

2. **TMLR 2026** [link]: Zhangheng Li, **Jianing Zhu**, Junyuan Hong, Sungmin Eum, Shuowen Hu, Suya You, Zhangyang Wang, "`POPS: Recovering Unlearned Multi-Modality Knowledge in MLLMs with Fine-tuning and Prompt-based Attacks`".

3. **arXiv 2026** [link]: Jingwei Sun*, **Jianing Zhu***, Jiangchao Yao, Tongliang Liu, Bo Han, "`Rethinking How to Remember: Beyond Atomic Facts in Lifelong LLM Agent Memory`".

4. **arXiv 2026** [link]: Junfeng Liao, Qizhou Wang, **Jianing Zhu**, Bo Du, Rui Yan, Xiuying Chen, "`Belief Memory: Agent Memory Under Partial Observability`".

5. **ICLR 2025** [link]: Qizhou Wang, Bo Han, Puning Yang, **Jianing Zhu**, Tongliang Liu, Masashi Sugiyama, "Unlearning with Control: Assessing Real-world Utility for Large Language Model Unlearning".

6. **ICML 2025 Workshop** [link]: **Jianing Zhu**, Zongze Li, Chandler Squires, Qizhou Wang, Bo Han, Pradeep Ravikumar, "On the Fragility of Latent Knowledge: Layer-wise Influence under Unlearning in Large Language Model".

Additional Publications

1. **ICLR 2026** [link]: Zizhuo Zhang, Qizhou Wang, Shanshan Ye, **Jianing Zhu**, Jiangchao Yao, Bo Han, Masashi Sugiyama, "Towards Understanding Valuable Preference Data for Large Language Model Alignment".

2. **arXiv 2026** [link]: John T Robertson, **Jianing Zhu**, Haris Vikalo, Zhangyang Wang, "When Is Rank-1 Steering Cheap? Geometry, Granularity, and Budgeted Search".

3. **ACL 2026 Demo** [link]: Guangwei Zhang, **Jianing Zhu**, Cheng Qian, Neil Gong, Rada Mihalcea, Zhaozhuo Xu, Jingrui He, Jiaqi Ma, Yun Huang, Chaowei Xiao, Bo Li, Ahmed Abbasi, Dongwon Lee, Heng Ji, Denghui Zhang, "Copyright Detective: A Forensic System to Evidence LLMs Flickering Copyright Leakage Risks".

4. **NeurIPS 2024** [link]: Zhanke Zhou, Rong Tao, **Jianing Zhu**, Yiwen Luo, Zengmao Wang, Bo Han, "Can Large Language Models Reason Robustly with Noisy Rationales?".

5. **arXiv 2024** [link]: Zhanke Zhou*, **Jianing Zhu***, Fengfei Yu*, Xuan Li, Xiong Peng, Tongliang Liu, Bo Han, "Model Inversion Attacks: A Survey of Approaches and Countermeasures".

INVITED TALKS

- Surgical Machine Unlearning @ UIUC MLS — Mar. 2026
- Unlearning Sensitive Information @ Texas Symposium — Mar. 2026
- Surgical Machine Unlearning @ NTU CCDS — Jan. 2026
- (Tutorial) "Handling Out-of-Distribution Data in the Open World" @ AAAI 2026 — Jan. 2026
- (Lecture) "Trustworthy Out-of-Distribution Detection" @ SJTU Graduate Course — Nov. 2025
- How AI Leaks Information — and How It Can Forget @ UT Digital Trust Symposium — Nov. 2025
- Towards Trustworthy Machine Learning for Out-of-Distribution Data @ SCU CS — Dec. 2023
- Diversified Outlier Exposure for Out-of-Distribution Detection @ HKBU Graduate Seminar — Dec. 2023
- (Lecture) Computer Science Research Methods @ HKBU Graduate Course — Oct. 2023
- Slack Federated Adversarial Training @ XMU ASC — Aug. 2023
- Youth PhD Talk for Conference Work Sharing @ AI Time — Jun./Nov. 2023

SERVICES

Organization Committee:

Leading Organizer, NeurIPS'26 Workshop on AI-Native Academia — 2026
Leading Organizer, AAAI 2026 Tutorial on Handling OOD Data — 2026
Leading Organizer, TMLR Young Scientist Seminars — 2023-2025
Workshop Chair Assistant, NeurIPS 2024 — 2024
Workshop Proposal Reviewer, NeurIPS — 2025
Workshop Assistant, HKBU-COMP & RIKEN-AIP Joint Workshop — 2024
Founding Member of Executive Group, RIKEN TrustML Young Scientist Seminars — 2022-2023

Program Committee:

Area Chair

The Annual Conference on Neural Information Processing Systems (NeurIPS) — 2026
International Conference on Machine Learning (ICML) — 2026
International Conference on Learning Representations (ICLR) — 2026

Conference Reviewer

International Conference on Machine Learning (ICML) [Best Reviewer] — 2021-2025

Neural Information Processing Systems (NeurIPS) [Top Reviewer] 2021-2025
International Conference on Learning Representations (ICLR) [Notable Reviewer] 2022-2026
International Conference on Artificial Intelligence and Statistics (AISTATS) 2023-2025
AAAI Conference on Artificial Intelligence (AAAI) 2023-2025
International Joint Conference on Artificial Intelligence (IJCAI) 2022-2025
Asian Conference on Machine Learning (ACML) 2021-2025

Journal Editor

Editorial Board Member, Machine Learning Journal (MLJ) 2026-Present

Journal Reviewer

IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI) 2023-Present
IEEE Transactions on Neural Networks and Learning Systems (IEEE TNNLS) 2021-Present
IEEE Transactions on Information Forensics & Security (IEEE TIFS) 2024-Present
ACM Computing Surveys (ACM CSUR) 2023-Present
Machine Learning Journal (MLJ) 2021-2025
Journal of Artificial Intelligence Research (JAIR) 2022-Present
Neural Networks (NN) 2021-Present

TEACHING

Teaching assistant for the following courses @ HKBU:

COMP7240(PG): Recommender Systems, Autumn 2022-2023
COMP7160(PG): Research Methods in Computer Science, Autumn 2022-2023
COMP7250(PG): Machine Learning, Spring 2022
COMP4135(UG): Recommender Systems and Applications, Autumn 2022-2023
COMP3057(UG): Intro to AI and Machine Learning, Autumn 2022