Jinghan Jia

+1-352-870-5374 | jiajingh@msu.edu | Google Scholar Citation 1003

Jinghan Jia | jinghanjia | jia_jinghan | jinghan.com

Lansing, Michigan - 48910, United States

Research Focus

- **Foundation Models (LLMs / Diffusion Models)**: Trustworthy AI (Unlearning, Alignment, Privacy), Efficient Training (Sparsification, Memory-/Parameter-Efficient Fine-Tuning, MoE), LLM Reasoning (Test-Time Computing, Reasoning-Enhanced Training)
- **Machine Learning**: Zeroth-order Optimization, Bi-level Optimization, Convex/Non-convex Optimization

Industrial Experience

- **IBM Research** *May 2025 - current*
  Research Intern, Supervisor: Nathalie Baracaldo San Jose, United States
  - Assessing the safety and reliability of agent models that leverage reinforcement learning to employ tools, like search engines as demonstrated by DeepResearch.
- **ByteDance Seed** *May 2024 - Nov. 2024*
  AI Research Intern, Supervisor: Xiaojun Xu San Jose, United States
  - Developed a robust multi-bit text watermarking system using LLM-based paraphrasers, fine-tuned with reinforcement learning from human feedback (RLHF).
  - Achieved 0.9999 AUC in watermark bit detection with a lightweight sentence-level classifier, ensuring high reliability and stealthiness.
  - This work resulted in a patent and a paper, with code released to support reproducibility.
- **Amazon** *May 2023 - August 2023*
  Applied Scientist Intern, Supervisor: Aram Galstyan Los Angeles, United States
  - Evaluated task-oriented conversational AI using LLMs with zero-shot and few-shot capabilities, focusing on automated dialogue quality assessments.
  - Conducted experiments on public and proprietary datasets, optimizing model configurations and implementing 'chain-of-thought' reasoning for improved accuracy and performance.
  - Presented findings in a paper published at the NAACL conference, demonstrating that fine-tuned LLMs significantly enhance automated dialogue evaluation.

Education

- **Michigan State University** *August 2021 - Current*
  Ph.D. Candidate in Computer Science East Lansing, United States
- **University of Florida** *August 2019 - July 2021*
  M.S. in Electrical and Computer Engineering Gainesville, United States
- **University of Science and Technology of China** *August 2015 - July 2019*
  B.Eng in Computer Science Hefei, China

Selected Publications

C=Conference, J=Journal, P=Patent, S=In Submission, T=Thesis

Jinghan Jia has co-authored 21 papers in top-tier machine learning, computer vision, NLP venues (NeurIPS, ICLR, CVPR, ECCV, EMNLP, etc.) and published 11 first-authored papers. Below are part of his publications: * indicates an **equal contribution**, and ‡ denotes the author is his mentee. Full list of publications at **Google Scholar** (Citation 1003).

[S.1] Chongyu Fan‡, Yihua Zhang, **Jinghan Jia**, et al. **CyclicReflex: Improving Large Reasoning Models via Cyclical Reflection Token Scheduling**. NeurIPS'2025 Submitted.

[S.2] **Jinghan Jia**, et al. **EPiC: Towards Lossless Speedup for Reasoning Training through Edge-Preserving CoT Condensation**. NeurIPS'2025 Submitted.

[C.1] Haomin Zhuang, Yihua Zhang, Kehan Guo, **Jinghan Jia**, Gaowen Liu, Sijia Liu, Xiangliang Zhang. **UOE: Unlearning One Expert Is Enough For Mixture-of-Experts LLMs**. ACL'25.

[C.2] Chongyu Fan*,‡, **Jinghan Jia***, et al. **Towards LLM Unlearning Resilient to Relearning Attacks: A Sharpness-Aware Minimization Perspective and Beyond**. ICML'2025.

[P.1] Xiaojun Xu, **Jinghan Jia**, Hang Li, Yuanshun Yao. **Watermark processing**.

[C.3] Changsheng Wang‡, Yihua Zhang, **Jinghan Jia**, et al. **Invariance Makes LLM Unlearning Resilient Even to Unanticipated Downstream Fine-Tuning**. ICML'2025.

[C.4] Xiaojun Xu, **Jinghan Jia**, Yuanshun Yao, Yang Liu, Hang Li. **Robust Multi-bit Text Watermark with LLM-based Paraphrasers**. ICML'2025.

[S.3] Chongyu Fan*,‡, Jiancheng Liu*, Licong Lin*, Jinghan Jia, et al. **Simplicity Prevails: Rethinking Negative Preference Optimization for LLM Unlearning**. NeurIPS'2025 Submitted.

[C.5] **Jinghan Jia**, et al. **WAGLE: Strategic Weight Attribution for Effective and Modular Unlearning in Large Language Models**. NeurIPS'24.

[C.6] **Jinghan Jia**, Y. Zhang, Y. Zhang, J. Liu, B. Runwal, J, Diffenderfer, Bhavya Kailkhura, S. Liu. **SOUL: Unlocking the Power of Second-Order Optimization for LLM Unlearning**. EMNLP'24 Main Track.

[J.1] Sijia Liu, Yuanshun Yao*, **Jinghan Jia***, et al. **Rethinking Machine Unlearning for Large Language Models.** Nature Machine Intelligence.

[C.7] Yihua Zhang, Yimeng Zhang, Yuguang Yao, **Jinghan Jia**, Jiancheng Liu, Xiaoming Liu, Sijia Liu. **UnlearnCanvas: A Stylized Image Dataset to Benchmark Machine Unlearning for Diffusion Models.** NeurIPS'24 Dataset and Benchmark Track.

[C.8] Yimeng Zhang, Xin Chen, **Jinghan Jia**, et al. **Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models.** NeurIPS'24.

[C.9] Yimeng Zhang*, **Jinghan Jia***, et al. **"To Generate or Not? Safety-Driven Unlearned Diffusion Models Are Still Easy To Generate Unsafe Images... For Now".** ECCV'24.

[C.10] **Jinghan Jia**, et al. **Leveraging LLMs for Dialogue Quality Measurement.** NAACL'24.

[C.11] Aochuan Chen*, Yimeng Zhang*, **Jinghan Jia**, et al. **DeepZero: Scaling up Zeroth-order Optimization for Deep Model Training.** ICLR'24

[C.12] **Jinghan Jia***, Jiancheng Liu*, et al. **Model Sparsity can Simplify Machine Unlearning.** NeurIPS'23 - Spotlight.

[C.13] Yihua Zhang*, Yimeng Zhang*, Aochuan Chen*, **Jinghan Jia**, et al. **Selectivity Drives Productivity: Efficient Dataset Pruning for Enhanced Transfer Learning.** NeurIPS'23.

[C.14] **Jinghan Jia***, Shashank Srikant*, et al. **Having Both: Robust and Accurate Code Models.** IEEE SANER'23.

[C.15] Bairu Hou, **Jinghan Jia**, et al. **TextGrad: Advancing Robustness Evaluation in NLP by Gradient-Driven Optimization.** ICLR'23.

[C.16] Yimeng Zhang, Xin Chen, **Jinghan Jia**, et al. **Text-Visual Prompting for Efficient 2D Temporal Video Grounding.** CVPR'23.

[C.17] Hui Li‡, **Jinghan Jia**, et al, **SMUG: Towards robust MRI reconstruction by smoothed unrolling.** ICASSP'23.

[C.18] **Jinghan Jia**, et al. **Robustness-preserving Lifelong Learning via Dataset Condensation.** ICASSP'23.

[C.19] **Jinghan Jia**, et al. **On the Robustness of deep learning-based MRI Reconstruction to image transformations.** TSRML'22.

[C.20] Yimeng Zhang, Yuguang Yao, **Jinghan Jia**, et al. **How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective.** ICLR'22 - Spotlight.

[C.21] **Jinghan Jia***, Chi Zhang*, Burhaneddin Yaman*, et al. **On Instabilities of Conventional Multi-Coil MRI Reconstruction to Small Adversarial Perturbations.** Asilomar Conference 2021.

TUTORIAL AND INVITED TALKS

- **Tutorial at CVPR 2024:** Machine Unlearning in Computer Vision: Foundations and Applications.
- **Invited Talk at University of Minnesota (UMN):** Recent Progress and Advancements in Large Language Models Unlearning.
- **Tutorial at NeurIPS 2022:** Foundational Robustness of Foundation Models.

HONORS AND AWARDS

- **NeurIPS Scholar Award** *2023*
  Conference on Neural Information Processing Systems
- **Herbert Wertheim College of Engineering Achievement Award Scholarship** *2019&2020*
  University of Florida
- **USTC Outstanding Student Scholarship** *2018*
  University of Science and Technology of China
- **USTC Newly Enrolled Students Scholarship** *2015*
  University of Science and Technology of China

SKILLS

- **Programming Languages:** Python, Matlab, C, C++
- **Deep Learning Libraries:** Pytorch, Deepspeed, Huggingface, Verl

SERVICES

Conference Reviewer: ICLR, NeurIPS, ICASSP, AAAI, CVPR, etc
Workshop Student Chair: Workshop Series: AdvML: New Frontiers in Adversarial Machine Learning [ICML'23].

MENTEES

- **Chongyu Fan (Phd, MSU)** *May. 2024 - Jan. 2025*
  ICML'25
- **Changsheng Wang (Phd, MSU)** *May. 2024 - Oct. 2024*
  ICML'25
- **Hui Li (Undergraduate, HUST)** *May. 2022 - Oct. 2022*
  ICASSP'23