

Jinghan Jia

+1-352-870-5374 | jiajinh@msu.edu | jinghan.com

 [your-profile](#) |  [jinghanjia](#) |  [jia_jinghan](#)

Okemos, Michigan - 48864, United States

RESEARCH FOCUS

- **Foundation Models (LLM/Diffusion Model):** Trustworthiness (Machine Unlearning, Alignment & RLHF, Privacy), Efficiency (Model Sparsification, Memory-Efficient Fine-Tuning, Parameter-Efficient Fine-Tuning).
- **Machine Learning:** Zeroth-order Optimization, Bi-level Optimization, Convex/Non-convex Optimization

INDUSTRIAL EXPERIENCE

• ByteDance Research

AI Research Intern, Supervisor: [Xiaojun Xu](#)

May 2024 - Current

San Jose, United States

- Developed an innovative text watermarking system utilizing LLMs for paraphrasing and RLHF.
- Achieved a detection accuracy of 0.9993 AUROC in watermark, significantly enhancing system reliability.
- Enhanced semantic preservation in watermarked texts to maintain content integrity and readability.

• Amazon

Applied Scientist Intern, Supervisor: [Aram Galstyan](#)

May 2023 - August 2023

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

Ph.D. Candidate in Computer Science

August 2021 - Current

East Lansing, United States

• University of Florida

M.S. in Electrical and Computer Engineering

August 2019 - July 2021

Gainesville, United States

• University of Science and Technology of China

B.Eng in Computer Science

August 2015 - July 2019

Hefei, China

PUBLICATIONS

C=CONFERENCE, J=JOURNAL, P=PATENT, S=IN SUBMISSION, T=THESIS

Jinghan Jia has co-authored 18 papers in top-tier machine learning, computer vision, NLP venues (NeurIPS, ICLR, CVPR, ECCV, EMNLP, etc.) and published 9 first-authored papers. Below are his publications: * indicates an equal contribution, and ‡ denotes the author is his mentee. Full list of publications at [Google Scholar](#) (Citation 301).

- [S.1] Chongyu Fan^{*,‡}, Jiancheng Liu^{*}, Licong Lin^{*}, Jinghan Jia, et al. **Simplicity Prevails: Rethinking Negative Preference Optimization for LLM Unlearning**. ICLR'2025 Submitted.
- [C.1] Jinghan Jia, et al. **WAGLE: Strategic Weight Attribution for Effective and Modular Unlearning in Large Language Models**. NeurIPS'24.
- [C.2] 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.
- [S.2] Sijia Liu, Yuanshun Yao^{*}, Jinghan Jia^{*}, et al. **Rethinking Machine Unlearning for Large Language Models**. Manuscript submitted for publication in Nature Machine Intelligence.
- [C.3] 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.4] Yimeng Zhang, Xin Chen, Jinghan Jia, et al. **Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models**. NeurIPS'24.
- [C.5] Jinghan Jia^{*}, Yimeng Zhang^{*}, et al. **"To Generate or Not? Safety-Driven Unlearned Diffusion Models Are Still Easy To Generate Unsafe Images... For Now"**. ECCV'24.
- [C.6] Jinghan Jia, et al. **Leveraging LLMs for Dialogue Quality Measurement**. NAACL'24.
- [C.7] Aochuan Chen^{*}, Yimeng Zhang^{*}, Jinghan Jia, et al. **DeepZero: Scaling up Zeroth-order Optimization for Deep Model Training**. ICLR'24

- [C.8] **Jinghan Jia***, Jiancheng Liu*, et al. **Model Sparsity can Simplify Machine Unlearning**. NeurIPS'23 - Spotlight.
- [C.9] Yihua Zhang*, Yimeng Zhang*, Aochuan Chen*, **Jinghan Jia**, et al. **Selectivity Drives Productivity: Efficient Dataset Pruning for Enhanced Transfer Learning**. NeurIPS'23.
- [C.10] **Jinghan Jia***, Shashank Srikant*, et al. **Having Both: Robust and Accurate Code Models**. IEEE SANER'23.
- [C.11] Bairu Hou, **Jinghan Jia**, et al. **TextGrad: Advancing Robustness Evaluation in NLP by Gradient-Driven Optimization**. ICLR'23.
- [C.12] Yimeng Zhang, Xin Chen, **Jinghan Jia**, et al. **Text-Visual Prompting for Efficient 2D Temporal Video Grounding**. CVPR'23.
- [C.13] Hui Li[‡], **Jinghan Jia**, et al, **SMUG: Towards robust MRI reconstruction by smoothed unrolling**. ICASSP'23.
- [C.14] **Jinghan Jia**, et al. **Robustness-preserving Lifelong Learning via Dataset Condensation**. ICASSP'23.
- [C.15] **Jinghan Jia**, et al. **On the Robustness of deep learning-based MRI Reconstruction to image transformations**. TSRML'22.
- [C.16] Yimeng Zhang, Yuguang Yao, **Jinghan Jia**, et al. **How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective**. ICLR'22 - Spotlight.
- [C.17] **Jinghan Jia***, Chi Zhang*, Burhaneddin Yaman*, et al. **On Instabilities of Conventional Multi-Coil MRI Reconstruction to Small Adversarial Perturbations**. ISMRM'21 - Oral.

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
<i>Conference on Neural Information Processing Systems</i> | 2023 |
| • Herbert Wertheim College of Engineering Achievement Award Scholarship
<i>University of Florida</i> | 2019&2020 |
| • USTC Outstanding Student Scholarship
<i>University of Science and Technology of China</i> | 2018 |
| • USTC Newly Enrolled Students Scholarship
<i>University of Science and Technology of China</i> | 2015 |

SKILLS

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

SERVICES

Conference Reviewer: ICLR'22/23/24, NeurIPS'22/23/24, ICML'22, AISTATS'23, ICASSP'23, ICASSP'24

Workshop Student Chair: Workshop Series: AdvML: New Frontiers in Adversarial Machine Learning [ICML'23].

MENTEES

- | | |
|---|------------------------------|
| • Chongyu Fan (Phd, MSU)
<i>ICLR'25 Submitted</i> | <i>May. 2024 - Oct. 2024</i> |
| • Hui Li (Undergraduate, HUST)
<i>ICASSP'23</i> | <i>May. 2022 - Oct. 2022</i> |