Jinghan Jia

+1-352-870-5374 | jiajingh@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** *May 2024 - Nov. 2024*
  AI Research Intern, Supervisor: Xiaojun Xu 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** *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

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 343).

[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** *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

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)** *May. 2024 - Oct. 2024*
  ICLR'25 Submitted
- **Hui Li (Undergraduate, HUST)** *May. 2022 - Oct. 2022*
  ICASSP'23