Chongyu Fan

fanchon2@msu.edu Website Github Google Scholar Citation 442

Education

Michigan State University, East Lansing, USA 2024.08 - Present
Doctor of Computer Science
Advisor: Prof. Sijia Liu @ OPTML Lab

Huazhong University of Science and Technology, Wuhan, China 2020.09 - 2024.06
Bachelor of Engineering
Outstanding Graduate
GPA: 3.96/4.0

Industrial Experience

ByteDance, Full-time Summer Internship, San Jose, USA 2025.05 – 2025.08
Research Scientist Intern, Mentor: Jian Du
Project: Multi-agent System and Model Context Protocol
- Improve multi-agent reasoning with reinforcement learning
- Designe counterfactual rewards to enhance collaboration in multi-agent systems
- Defined privacy issues in Model Context Protocol
- Proposed *Log-to-Leak*, an attack framework for agent systems
- Research outcome: Paper submitted to **ICLR'26**

Research Project Highlights

• Knowledge Editing, Machine Unlearning, and Alignment
My research focuses on developing methods to mitigate the influence of undesired knowledge in foundation models, particularly during the post-training stage of LLMs **[NeurIPS'25a]**, **[EMNLP'25]** and diffusion models **[ICLR'24]**, **[NeurIPS'24a]**, **[NeurIPS'24b]**. Through these efforts, I aim to enhance the trustworthiness, robustness, and safety of next-generation AI systems.

• Bi-level and Smooth Optimization for Trustworthy LLMs and Diffusion Models
My research leverages bi-level optimization to improve both the training and inference of LLMs and diffusion models, enhancing their robustness and interpretability **[ECCV'24]**, **[NeurIPS'25b]**, and explores smooth optimization to strengthen the safety and reliability of LLMs **[ICML'25]**.

• Resource-Aware Inference and Training for Scalable Reasoning Models
I investigate optimized test-time compute strategies to enhance reasoning accuracy while mitigating over- and under-thinking issues **[arXiv'25a]**, and design dataset condensation techniques to improve the efficiency and scalability of reasoning model training **[arXiv'25b]**.

Publications

I have published more than ten papers in top-tier machine learning and computer vision venues (*e.g.*, NeurIPS, ICML, ICLR, ECCV, EMNLP), with five of them as first author. As of September 25, 2025, my research has garnered 442 citations on Google Scholar.

First-Authored Publications (* indicates equal contribution)

[ICLR'24] **C. Fan**, J. Liu, Y. Zhang, D. Wei, E. Wong, S. Liu, *"SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation."* **(Spotlight, acceptance rate 5%; 200+ citations)**

[ECCV'24] **C. Fan**, J. Liu, A. Hero, S. Liu, *"Challenging forgets: Unveiling the worst-case forget sets in machine unlearning."*

[ICML'25] **C. Fan**, J. Jia, Y. Zhang, A. Ramakrishna, M. Hong, S. Liu, *"Towards LLM Unlearning Resilient to Relearning Attacks: A Sharpness-Aware Minimization Perspective and Beyond."*

[EMNLP'25] **C. Fan***, C. Wang*, Y. Zhang, J. Jia, D. Wei, P. Ram, N. Baracaldo, S. Liu, *"Reasoning Model Unlearning: Forgetting Traces, Not Just Answers, While Preserving Reasoning Skills."*

[NeurIPS'25a] **C. Fan**, J. Liu, L. Lin, J. Jia, R. Zhang, S. Mei, S. Liu, *"Simplicity Prevails: Rethinking Negative Preference Optimization for LLM Unlearning."*

Co-Authored Publications

[NeurIPS'24a] Y. Zhang, **C. Fan**, Y. Zhang, Y. Yao, J. Jia, J. Liu, G. Zhang, G. Liu, R. Kompelia, X. Liu, S. Liu, *"UnlearnCanvas: Stylized Image Dataset for Enhanced Machine Unlearning Evaluation in Diffusion Models."*

[NeurIPS'24b] Y. Zhang, X. Chen, J. Jia, Y. Zhang, **C. Fan**, J. Liu, M. Hong, K. Ding, S. Liu, *"Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models."*

[ICML'25W] J. Lee, Z. Mai, **C. Fan**, W.L. Chao, *"An Empirical Exploration of Continual Unlearning for Image Generation."*

[NeurIPS'25b] Y. Zhang, C. Wang, Y. Chen, **C. Fan**, J. Jia, S. Liu, *"The Fragile Truth of Saliency: Improving LLM Input Attribution via Attention Bias Optimization."* **(Spotlight, acceptance rate 3%)**

Preprint Papers

[arXiv'25a] **C. Fan**, Y. Zhang, J. Jia, A. Hero, S. Liu, *"CyclicReflex: Improving Large Reasoning Models via Cyclical Reflection Token Scheduling."*

[arXiv'25b] J. Jia, H. Reisizadeh, **C. Fan**, N. Baracaldo, M. Hong, S. Liu, *"EPiC: Towards Lossless Speedup for Reasoning Training through Edge-Preserving CoT Condensation."*

Community Services

- **Workshop Student Co-Organizer:** New Frontiers in Adversarial ML [NeurIPS'24]
- **Reviewer:** NeurIPS, ICLR, ICML, AISTATS

Honors

Graduate Travel Fellowship, Michigan State University 2024