Yiwei Chen

chenyiw9@msu.edu | Homepage | LinkedIn | Google Scholar

Education

Michigan State University *East Lansing, MI, USA*

- **Ph.D. in Computer Science** Aug. 2024 -
  Advisor: Prof. Sijia Liu

Xi'an Jiaotong University *Xi'an, Shaanxi, China*

- **M.S. in Computer Science and Technology.** Ranking: 7/173 Sept. 2021 - Jun. 2024
- **B.Eng. in Automation.** Sept. 2017 - Jun. 2021
  Honor Class in Qian Xuesen Honors College. Ranking: 3/24
- **Special Class for Gifted Young.** Sept. 2015 - Jun. 2017
  A honor program for nationwide selected gifted young students.

Research Interests

- Reasoning, Post-training, Large Language Models, Multi-modal Language Models
- AI Safety, Alignment, Trustworthy Algorithms, Interpretability, Machine Unlearning

Publications & Manuscripts

(* means Equal Contribution)

[1] **Y. Chen***, S. Pal*, Y. Zhang, Q. Qu, S. Liu. "Unlearning Isn't Invisible: Detecting Unlearning Traces in LLMs from Model Outputs." *The International Conference on Learning Representations (ICLR)*, 2026. [Paper]

[2] **Y. Chen***, Y. Yao*, Y. Zhang, B. Shen, G. Liu, S. Liu. "Safety Mirage: How Spurious Correlations Undermine VLM Safety Fine-Tuning and Can Be Mitigated by Machine Unlearning." *The International Conference on Learning Representations (ICLR)*, 2026. [Paper]

[3] B. Shang*, **Y. Chen***, Y. Zhang, B. Shen, S. Liu. "Forgetting to Forget: Attention Sink as A Gateway for Backdooring LLM Unlearning." In Submission to ICML'26. [Paper]

[4] Z. Zhang*, **Y. Chen***, W. Zhang, C. Yan, Q. Zheng, Q. Wang, W. Chen. "Tile Classification Based Viewport Prediction with Multi-modal Fusion Transformer." *ACM Multimedia (ACM-MM)*, 2023. [Paper]

[5] 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. *Conference on Neural Information Processing Systems(NeurIPS)*, 2025, Spotlight (3.2% of 21575 submissions). [Paper]

Internships

Cisco, Research Intern San Francisco, US
Host: Lichi Li May 2025 - Oct. 2025
Research Project: Reasoning Large Language Models for CVE Exploit Generation in Cybersecurity

- *Exploit Generation*: Developed a benchmark with 6 input levels and 7 evaluation metrics, providing the first systematic framework for evaluating LLMs on exploit generation.
- *LLM Reasoning*: Collected a large-scale dataset of CVE–exploit pairs and fine-tuned Qwen3-8B using SFT + GRPO, yielding significant performance gains over SOTA LLMs (GPT, Claude, Qwen, etc).

Research Experience

OPTML, Michigan State University *East Lansing, US*
Research Assistant. Advisor: Prof. Sijia Liu Aug. 2024 -

- *MLLM Reasoning*: Existing improvements in mathematical reasoning for MLLMs are largely transferred from LLMs, with insufficient utilization of visual information during training. Propose visual-biased GRPO training that enhances visual grounding and improves reasoning performance.
- *LLM Unlearning* [1]: Investigated how machine unlearning leaves persistent, detectable "fingerprints" in both outputs and internal activations. Designed classifiers could identifying unlearned models, revealing new risks of reverse-engineering forgotten information.
- *MLLM Unlearning* [2]: Conventional safety fine-tuning of MLLMs suffers from a "safety mirage" caused by training bias, leading to spurious correlations and over-rejections under one-word attacks. Unlearning algorithms effectively remove harmful content and mitigate these issues.
- *Backdoor Unlearning* [3]: Demonstrated that LLM unlearning can be compromised through attention-sink-guided backdoor unlearning, where triggers placed at attention sinks enable models to recover forgotten knowledge while maintaining normal behavior without triggers.

Awards & Honors

SCHOLARSHIPS

- Michigan State University Travel Award Oct. 2025
- Outstanding Master Graduate, Xi'an Jiaotong University (Top 5% of all graduates) Jun. 2024
- Outstanding Bachelor Graduate, Xi'an Jiaotong University (Top 5% of all graduates) Jun. 2021
- Outstanding Freshman Scholarship, Xi'an Jiaotong University (Top 10%) Nov. 2021
- First Prize Scholarship, Xi'an Jiaotong University (Top 10%) Nov. 2016-2019, 2022-2023

COMPETITIONS

- Meritorious Winner in Mathematical Contest in Modelling (top 8% in 25370 teams) Mar. 2019
- The First Prize in Shaanxi of China Undergraduate Mathematical Contest in Modeling Oct. 2018

Academic Service

- **Conference Reviewer:** ICLR (25–26), ICASSP (25–26), ACM MM (23–24), IJCNN (25)
- **Journal Reviewer:** Journal of Machine Learning Research (JMLR)

Language & Skills

Language	Mandarin (Native)
	English (TOEFL iBT 103, Reading 28 Listening 28 Speaking 25 Writing 22)
Programming	Python, C/C++, Bash, LaTeX, MATLAB, HTML/CSS
Model Frameworks	LLMs (e.g., Llama, Qwen), MLLMs (e.g., LLaVA, Qwen-VL)
Libraries / Softwares	Pytorch, Tensorflow, Deepspeed, Huggingface, OpenCV
Developer Tools	Git, Docker, Vim, VSCode, Cursor, Claude Code, PyCharm