HE, PENGFEI

Homepage: https://pengfeihepower.github.io/ Email: hepengf1@msu.edu

EDUCATION

- **Michigan State University**,
  - Ph.D. in Computer Science and Engineering. Advisor: Dr. Jiliang Tang Sep 2022 - Aug 2026 (expected)
  - Ph.D. in Statistics and Probability. Dr. Yuehua Cui and Dr. Haolei Weng Sep 2020 - Aug 2026 (expected)
  - Research Interest: Trustworthy LLMs and LLM-based agents. AI safety (data poisoning) and privacy. LLM and agents reasoning.
- **University of Wisconsin-Madison**, Sep 2019 - May 2020
  - M.Sc. in Statistics, Department of Statistics
- **Nankai University**, Sep 2015 - May 2019
  - B.Sc. in Statistics, School of Mathematics

RESEARCH EXPERIENCES

Sharpness-aware Data Poisoning Attack (ICLR 2024 Spotlight)

- Data poisoning attacks are perturbations injected into the dataset so that trained model will perform poorly. Existing attacks are developed for one or a few victim models, and have poor transferability across different victim models and training algorithms.
- We develop a new attacking method (SAPA), inspired by the loss landscape sharpness, to search for poisoning samples that can maximize the poisoning effect on the most robust victim model. Our method is shown as a general strategy that can be combined with existing attacks and improve their performance on various models and real datasets like Cifar10 and ImageNet.

Data Poisoning on In-context Learning (NAACL 2025)

- In-context learning (ICL) is a powerful ability that emerged from large language models (LLMs). However, ICL suffers from a high risk against potential data poisoning attacks.
- We develop the first practical data poisoning attack against ICL to evaluate the robustness of commonly used LLMs. We add perturbations to the demonstration examples to distort the hidden states extracted from LLMs to interrupt its prediction.

Spectral Analysis on Multi-Manifold Clustering (JMLR)

- Multi-Manifold Clustering (MMC) aims to retrieve the multi-manifold structure underlying a given Euclidean data set. The challenge is that manifolds are usually intersected and have different dimensions.
- We investigate the sufficient conditions on similarity of graphs to ensure their corresponding graph Laplacians capturing the correct geometric information when solving the MMC problem. We provide high probability error bounds for the spectral approximation of a tensorized Laplacian on manifolds with a suitable graph Laplacian built from the observations.

PUBLICATIONS

- **Pengfei He**, Han Xu, Jie Ren, Yingqian Cui, Charu C. Aggarwal, Jiliang Tang. *Sharpness-Aware Data Poisoning Attack*, International Conference on Learning Representations (**ICLR**), 2024, **Spotlight 5%**.
- **Pengfei He**, Nicolas Garcia Trillos, Chenghui Li. *Large Sample Spectral Analysis of Graph-based Multi-manifold Clustering*. Journal of Machine Learning Research (**JMLR**), 2023.
- Han Xu,**Pengfei He**, Jie Ren, Yuxuan Wan, Zitao Liu, Jiliang Tang. *Probabilistic Categorical Adversarial Attack & Adversarial Training* , In the Proceedings of 40th International Conference on Machine Learning (**ICML**), 2023.
- **Pengfei He**, Yue Xing, Han Xu, Jie Ren, Yingqian Cui, Shenglai Zeng, Jiliang Tang, Makoto Yamada, Mohammad Sabokrou.*Stealthy Backdoor Attack via Confidence-driven Sampling*. Transactions on Machine Learning Research(**TMLR**), 2024.
- **Pengfei He**, Haochen Liu, Xiangyu Zhao, Jiliang Tang. *PROPN: Personalized Probabilistic Strategic Parameter Optimization in Recommendations*. 31st ACM International Conference on Information & Knowledge Management (**CIKM**), 2022.
- **Pengfei He**, Yingqian Cui, Han Xu, Hui Liu, Makoto Yamada, Jiliang Tang, Yue Xing.*Towards the Effect of Examples on In-Context Learning: A Theoretical Case Study*. M3L and SFLLM Workshop **NeruIPS** 2024. Appear on journal **Stat**.
- **Pengfei He**, Han Xu, Yue Xing, Makoto Yamada, Jiliang Tang. *Data Poisoning for In-context Learning*. **NAACL** 2025.
- **Pengfei He***, Yuping Lin*, Han Xu, Yue Xing, Makoto Yamada, Hui Liu, Jiliang Tang. *Towards Understanding Jailbreak Attacks in LLMs: A Representation Space Analysis*. **EMNLP** 2024
- **Pengfei He**, Yupin Lin, Shen Dong, and Han Xu, Yue Xing, Hui Liu. *Red-Teaming LLM Multi-Agent Systems via Communication Attacks*. **ACL** 2025.
- Bo Wang, Weiyi He, Shenglai Zeng, Zhen Xiang, Yue Xing, Jiliang Tang, **Pengfei He**. *Unveiling Privacy Risks in LLM Agent Memory*. **ACL** 2025.
- **Pengfei He**, Zitao Li, Yue Xing, Yaling Li, Jiliang Tang, Bolin Ding. *Make LLMs better zero-shot reasoners: Structure-orientated autonomous reasoning*. ArXiv.
- **Pengfei He**, Yue Xing, Han Xu, Zhen Xiang, Jiliang Tang. *Position: Multi-Faceted Studies on Data Poisoning can Advance LLM Development*. Preprint 2025.
- **Pengfei He**, Yue Xing, Shen Dong, Juanhui Li, Zhenwei Dai, Xianfeng Tang, Hui Liu, Han Xu, Zhen Xiang, Charu C. Aggarwal, Hui Liu. *Comprehensive Vulnerability Analysis is Necessary for Trustworthy LLM-MAS*. Preprint 2025.
- **Pengfei He**, Zhenwei Dai, Xianfeng Tang, Yue Xing, Hui Liu, Jingying Zeng, Qiankun Peng, Shrivats Agrawal, Samarth Varshney, Suhang Wang, Jiliang Tang, Qi He. *Attention Knows Whom to Trust: Attention-based Trust Management for LLM Multi-Agent Systems*. Preprint 2025.

- Yingqian Cui, **Pengfei He**, Jingying Zeng, Hui Liu, Xianfeng Tang, Zhenwei Dai, Yan Han, Chen Luo, Jing Huang, Zhen Li, Suhang Wang, Yue Xing, Jiliang Tang, Qi He. *Stepwise Perplexity-Guided Refinement for Efficient Chain-of-Thought Reasoning in Large Language Models*. **ACL** 2025.
- Shenglai Zeng, **Pengfei He**, Kai Guo, and Tianqi Zheng, Hanqing Lu, Yue Xing, Hui Liu. *Towards Context-Robust LLMs: A Gated Representation Fine-tuning Approach*. **ACL** 2025.
- Yingqian Cui, **Pengfei He**, Xianfeng Tang, Qi He, Chen Luo, Jiliang Tang, Yue Xing. *A theoretical understanding of chain-of-thought: Coherent reasoning and error-aware demonstration*. **AISTATS** 2025.
- Shen Dong, Shaochen Xu, **Pengfei He**, Yige Li, Jiliang Tang, Tianming Liu, Hui Liu, Zhen Xiang. *A practical memory injection attack against llm agents*. Preprint 2025.
- Shenglai Zeng, Yaxin Li, Jie Ren, Yiding Liu, Han Xu, **Pengfei He**, Yue Xing, Shuaiqiang Wang, Jiliang Tang, Dawei Yin. *Exploring Memorization in Fine-tuned Language Models*. **ACL** 2024.
- Shenglai Zeng, Jiankun Zhang, **Pengfei He**, Yiding Liu, Yue Xing, Han Xu, Jie Ren, Yi Chang, Shuaiqiang Wang, Dawei Yin, Jiliang Tang. *The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG)*. **ACL** 2024.
- Yingqian Cui, Jie Ren, **Pengfei He**, Jiliang Tang, Yue Xing. *Superiority of Multi-Head Attention in In-Context Linear Regression*. **AISTATS** 2025.
- Han Xu, Jie Ren, **Pengfei He**, Shenglai Zeng, Yingqian Cui, Amy Liu, Hui Liu, Jiliang Tang. *On the Generalization of Training-based ChatGPT Detection Methods*. **EMNLP** 2024.
- Yingqian Cui, Jie Ren, Han Xu, **Pengfei He**, Hui Liu, Lichao Sun, Yue Xing, Jiliang Tang *Diffusionshield: A watermark for copyright protection against generative diffusion models*. **ACM SIGKDD Explorations Newsletter** 2025.
- Yingqian Cui, Jie Ren, Yuping Lin, Han Xu, **Pengfei He**, Yue Xing, Lingjuan Lyu, Wenqi Fan, Hui Liu, Jiliang Tang *Ft-shield: A watermark against unauthorized fine-tuning in text-to-image diffusion models*. **ACM SIGKDD Explorations Newsletter** 2025.

PROFESSIONAL EXPERIENCES

Applied Scientist Intern at Amazon.com Inc., **Trustworthy LLM Multi-agent system&LLM agent tool usage**, Mar 2025 - present
Research Intern at Alibaba Group (U.S.) Inc., **LLM reasoning & Multi-agent development**, June 2024 - Sept 2024
Visiting research scholar at Okinawa Institute of Science and Technology (OIST), May 2023 - Jul 2023, Dec 2023 - present
Research Scientist at James Madison College, Michigan State university, December 2021 - August 2022

AWARDS

- ICLR 2025 Notable reviewer
- SDM2025 Doctoral Forum Travel Award
- Awarded the Professor James Stapleton Prize in Statistics, Michigan State University, Fall 2021.
- Travel Grant. Michigan State University, Graduate School, 2023, 2024.
- KDD-2022 Student Registration Award.
- CIKM-2022 Student Registration Award.

OTHER EXPERIENCE

- Serve as PC Member: AAAI(2022-2025), KDD(2023-2025), SDM(2023-2025), ICDM-2023, PAKDD-2023, WWW-2024, ACL(2024,2025), EMNLP(2024.2025), ICML(2024,2025), NeruIPS(2024,2025), ICLR-2025, AISTAT-2025
- Serve as Journal Reviewer: Journal of the American Statistical Association (JASA), Transactions on Knowledge and Data Engineering (TKDE), Transactions on Knowledge Discovery from Data (TKDD), Transactions on Machine Learning Research (TMLR)
- Serve as SPC for TKDD
- Serve as conference volunteers: KDD-2022

PROFESSIONAL SKILLS

- **Programming languages**: Python, R, MATLAB.
- **Softwares and systems**: Pytorch, Numpy, Pandas, Sklearn, Git, Docker, Linux.
- **AI skills**: Large language models (Qwen, DeepSeek, Llama, Vicuna, Mixtral, Pythia, Gemma, GPT), computer vision (Vision Transformer, ResNet), data science, data visualization.