Pengfei He

Contact Information

428 S Shaw Ln Rm 3308
East Lansing, MI 48824
E-mail: hepengf1@msu.edu

Homepage: https://pengfeihepower.github.io/
GitHub: https://github.com/PengfeiHePower
Phone: (+1) 6086223144

Research Interests

- Trustworthy LLMs and agents (inter-agent communication security, memory manipulation, etc)
- AI security (ML attack and defense, privacy)
- LLM reasoning, agentic tool usage; AI applications and interpretability

Education

Michigan State University, USA 09/2022 – 06/2026 (Expected)

- Ph.D., Computer Science and Engineering
- Advisor: Dr. Jiliang Tang

Michigan State University, USA 09/2020 – 06/2026 (Expected)

- Ph.D., Statistics
- Advisor: Dr. Yuehua Cui

University of Wisconsin-Madison, USA 09/2018 – 06/2020

- Master, Statistics

Nankai University, China 09/2015 – 06/2019

- Bachelor of Science, Statistics

Professional Experience

Data Science and Engineering Lab, Michigan State University 09/2022 – Present
Research Assistant, Department of Computer Science and Engineering

- Topic: Trustworthy LLM and agents, AI security
- Mentor: Dr. Jiliang Tang

CloudAI Research, Google 10/2025 – present
Student Researcher

- Topic: LLM Multi-agent for Cybersecurity, and auto-red-team for agentic prompt injection
- Mentor: Dr. Long T. Le

Ads Team, Amazon 03/2025 – 9/2025
Applied scientist intern

- Topic: Trust management for LLM multi-agent; benchmarking LLM agentic toolusage
- Mentor: Dr. Zhenwei Dai

Alibaba Cloud, Alibaba Inc Bellevue 06/2024 – 08/2024
Research Scientist Intern

- Topic: LLM reasoning
- Mentors: Dr. Zitao Li and Dr. Boling Ding

Machine learning and Data science Unit, OIST 05/2023 – 08/2023
Visiting scholar

- Topic: Stealthy backdoor attacks
- Mentor: Dr. Mohammad Sabokrou and Dr. Makoto Yamada

Honors and Awards

- Outstanding Graduate Student Award, CSE MSU (1/160) 2026
- SDM Doctoral Forum Travel Award 2025
- ICLR Notable reviewer 2025
- Graduate School Travel Fellowship, MSU 2025
- Graduate School Travel Fellowship, MSU 2024
- Graduate School Travel Fellowship, MSU 2023
- KDD Student Registration Award 2022
- CIKM Student Registration Award 2022
- Professor James Stapleton Prize in Statistics, MSU 2021
- Ranked Top 5% in the school of mathematics, Nankai University, 2019
- Second Prize, Nankai University 2016

Open-Source Projects

- **TRAJECT-Bench** [Link]. 09/2025 – Present
  The first trajectory-aware benchmark for LLM agentic tool usage, ICLR 2026
- **Multi-Agent Vulnerable (MAV) Framework** [Link] 10/2025 – Present
  The first benchmark for testing multi-agent systems against various security vulnerabilities and attack vectors, EACL 2026

Publication

Google Scholar: `https://scholar.google.com/citations?user=nsSrd6kAAAAJ&hl=en`

- **Impact: 1000+** citations, h-index: **17**

Conference and Journal Papers
* indicates equal contribution; name indicates mentored students

- "TRAJECT-Bench:A Trajectory-Aware Benchmark for Evaluating Agentic Tool Use"
  Pengfei He, Zhenwei Dai, Bing He, Hui Liu, Xianfeng Tang, Hanqing Lu, Juanhui Li, Jiayuan Ding, Subhabrata Mukherjee, Suhang Wang, Yue Xing, Jiliang Tang, Benoit Dumoulin
  International Conference on Learning Representations (**ICLR**, 2026)
- "Red-Teaming LLM Multi-Agent Systems via Communication Attacks"
  Pengfei He, Yupin Lin, Shen Dong, and Han Xu, Yue Xing, Hui Liu.
  The 63rd Annual Meeting of the Association for Computational Linguistics (**ACL**, 2025)
- "Unveiling Privacy Risks in LLM Agent Memory"
  Bo Wang, Weiyi He, Shenglai Zeng, Zhen Xiang, Yue Xing, Jiliang Tang, **Pengfei He**
  The 63rd Annual Meeting of the Association for Computational Linguistics (**ACL**, 2025)
- "Sharpness-Aware Data Poisoning Attack"
  Pengfei He, Han Xu, Jie Ren, Yingqian Cui, Charu C. Aggarwal, Jiliang Tang
  International Conference on Learning Representations (**ICLR**, 2024, **Spotlight <5%**)
- "Data Poisoning for In-context Learning"
  Pengfei He, Han Xu, Yue Xing, Makoto Yamada, Jiliang Tang
  Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (**NAACL**, 2025)

– "Memory Injection Attacks on LLM Agents via Query-Only Interaction"
Pengfei He*, Shen Dong*, Shaochen Xu*, Yige Li, Jiliang Tang, Tianming Liu, Hui Liu, Zhen Xiang
The Thirty-Ninth Annual Conference on Neural Information Processing Systems (**NeurIPS**, 2025)

– "Stealthy Backdoor Attack via Confidence-driven Sampling"
Pengfei He, Yue Xing, Han Xu, Jie Ren, Yingqian Cui, Shenglai Zeng, Jiliang Tang, Makoto Yamada, Mohammad Sabokrou.
Transactions on Machine Learning Research(**TMLR**, 2024)

– "Towards Understanding Jailbreak Attacks in LLMs: A Representation Space Analysis"
Pengfei He*, Yuping Lin*, Han Xu, Yue Xing, Makoto Yamada, Hui Liu, Jiliang Tang
Empirical Methods in Natural Language Processing (**EMNLP**, 2024)

– "Large Sample Spectral Analysis of Graph-based Multi-manifold Clustering"
Pengfei He*, Nicolas Garcia Trillos*, Chenghui Li*
Journal of Machine Learning Research (**JMLR**, 2023)

– "Probabilistic Categorical Adversarial Attack & Adversarial Training"
Han Xu,**Pengfei He**, Jie Ren, Yuxuan Wan, Zitao Liu, Jiliang Tang
In the Proceedings of 40th International Conference on Machine Learning (**ICML**, 2023)

– "Towards the Effect of Examples on In-Context Learning: A Theoretical Case Study"
Pengfei He, Yingqian Cui, Han Xu, Hui Liu, Makoto Yamada, Jiliang Tang, Yue Xing
M3L and SFLLM Workshop **NeruIPS** 2024. Appear on journal **Stat**. **Invited talk** on **JSM** 2025.

– "Make LLMs better zero-shot reasoners: Structure-orientated autonomous reasoning"
Pengfei He, Zitao Li, Yue Xing, Yaling Li, Jiliang Tang, Bolin Ding
Empirical Methods in Natural Language Processing(**EMNLP**, 2025)

– "PROPN: Personalized Probabilistic Strategic Parameter Optimization in Recommendations"
Pengfei He, Haochen Liu, Xiangyu Zhao, Jiliang Tang
31st ACM International Conference on Information & Knowledge Management (**CIKM**, 2022, **Oral**)

– "PEAR: Planner-Executor Agent Robustness Benchmark"
Pengfei He*, Shen Dong*, Mingxuan Zhang*, Li Ma, Bhavani Thuraisingham, Hui Liu, Yue Xing (**EACL**, 2026)

– "Stepwise Perplexity-Guided Refinement for Efficient Chain-of-Thought Reasoning in Large Language Models"
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
Annual Meeting of the Association for Computational Linguistics (**ACL**, 2025)

– "Towards Context-Robust LLMs: A Gated Representation Fine-tuning Approach"
Shenglai Zeng, **Pengfei He**, Kai Guo, and Tianqi Zheng, Hanqing Lu, Yue Xing, Hui Liu
Annual Meeting of the Association for Computational Linguistics (**ACL**, 2025)

– "A theoretical understanding of chain-of-thought: Coherent reasoning and error-aware demonstration"
Yingqian Cui, **Pengfei He**, Xianfeng Tang, Qi He, Chen Luo, Jiliang Tang, Yue Xing
Annual Conference on Artificial Intelligence and Statistics(**AISTATS**, 2025)

– "Can Multiple Responses from an LLM Reveal the Sources of Its Uncertainty?"
Yang Nan, **Pengfei He**, Ravi Tandon, Han Xu
Empirical Methods in Natural Language Processing(**EMNLP**, 2025)

– "Mitigating the Privacy Issues in Retrieval-Augmented Generation (RAG) via Pure Synthetic Data"
Shenglai Zeng*, Jiankun Zhang*, **Pengfei He**, Jie Ren, Tianqi Zheng, Hanqing Lu, Han Xu, Hui Liu, Yue Xing, Jiliang Tang
Empirical Methods in Natural Language Processing (**EMNLP**, 2025)

– "Exploring Memorization in Fine-tuned Language Models"
Shenglai Zeng, Yaxin Li, Jie Ren, Yiding Liu, Han Xu, **Pengfei He**, Yue Xing, Shuaiqiang Wang, Jiliang Tang, Dawei Yin
Annual Meeting of the Association for Computational Linguistics (**ACL**, 2024)

– "The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG)"
Shenglai Zeng, Jiankun Zhang, **Pengfei He**, Yiding Liu, Yue Xing, Han Xu, Jie Ren, Yi Chang, Shuaiqiang Wang, Dawei Yin, Jiliang Tang
Annual Meeting of the Association for Computational Linguistics (**ACL**, 2024)

– "Superiority of Multi-Head Attention in In-Context Linear Regression"
Yingqian Cui, Jie Ren, **Pengfei He**, Jiliang Tang, Yue Xing
Annual Conference on Artificial Intelligence and Statistics(**AISTATS**, 2025)

– "On the Generalization of Training-based ChatGPT Detection Methods"
Han Xu, Jie Ren, **Pengfei He**, Shenglai Zeng, Yingqian Cui, Amy Liu, Hui Liu, Jiliang Tang
Empirical Methods in Natural Language Processing (**EMNLP**, 2024)

– "Diffusionshield: A watermark for copyright protection against generative diffusion models"
Yingqian Cui, Jie Ren, Han Xu, **Pengfei He**, Hui Liu, Lichao Sun, Yue Xing, Jiliang Tang
ACM SIGKDD Explorations Newsletter, 2025

– "Ft-shield: A watermark against unauthorized fine-tuning in text-to-image diffusion models"
Yingqian Cui, Jie Ren, Yuping Lin, Han Xu, **Pengfei He**, Yue Xing, Lingjuan Lyu, Wenqi Fan, Hui Liu, Jiliang Tang
ACM SIGKDD Explorations Newsletter, 2025

– "Human-AI Collaboration for Knowledge-in-Use Assessment Design: Leveraging LLMs with RAG"
Juanhui Li, Tingting Li, Hang Li, Haoyu Han, **Pengfei He**, Peng He, Hui Liu
IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE), 2025

Preprints and Submissions

– "Co-RedTeam: Orchestrated Security Discovery and Exploitation with LLM Agents"
Pengfei He, Ash Fox, Lesly Miculicich, Stefan Friedli, Daniel Fabian, Burak Gokturk, Jiliang Tang, Chen-Yu Lee, Tomas Pfister, Long T Le
(Work at **Google**, Submitted to **ICML**, 2026)

– "Position: Multi-Faceted Studies on Data Poisoning can Advance LLM Development"
Pengfei He, Yue Xing, Han Xu, Zhen Xiang, Jiliang Tang

– "Comprehensive Vulnerability Analysis is Necessary for Trustworthy LLM-MAS"
Pengfei He*, Yue Xing*, Shen Dong, Juanhui Li, Zhenwei Dai, Xianfeng Tang, Hui Liu, Han Xu, Zhen Xiang, Charu C. Aggarwal, Hui Liu

– "Attention Knows Whom to Trust: Attention-based Trust Management for LLM Multi-Agent Systems"
Pengfei He, Zhenwei Dai, Xianfeng Tang, Yue Xing, Hui Liu, Jingying Zeng, Qiankun Peng, Shrivats Agrawal, Samarth Varshney, Suhang Wang, Jiliang Tang, Qi He
Submitted to Annual Meeting of the Association for Computational Linguistics (**ACL**, 2026)

– "Adaptive Test-Time Reasoning via Reward-Guided Dual-Phase Search" Yingqian Cui, Zhenwei Dai, **Pengfei He**, Bing He, Hui Liu, Xianfeng Tang, Jingying Zeng, Suhang Wang, Yue Xing, Jiliang Tang, Benoit Dumoulin
Submitted to (**ACL**, 2026)

– "Interpretable Probability Estimation with LLMs via Shapley Reconstruction"
Yang Nan, Qihao Wen, Jiahao Wang, **Pengfei He**, Ravi Tandon, Yong Ge, Han Xu
Submitted to (**ACL**, 2026)

– "How Memory Management Impacts LLM Agents: An Empirical Study of Experience-Following Behavior"
Zidi Xiong, Yuping Lin, Wenya Xie, **Pengfei He,** Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, Zhen Xiang
Submitted to Annual Meeting of the Association for Computational Linguistics (**ACL**, 2026)

Proposal Writing

– AI long form grading that reflects faculty style
PI: Dr. Jiliang Tang
Role: Designed and drafted the objective of utilizing in-context learning and active learning in grading
Result: Funded by the Gates Foundation in 2025

– Interactive, Individualized Professional Learning for Elementary School Teachers: Enhancing Content and Pedagogical Content Knowledge as a Basis for Improving Practice
PI: Dr. Jiliang Tang
Role: Wrote the part of building agents with access to the knowledge database in assisting teachers.
Result: Funded by the National Science Foundation (NSF) in 2024

– Intelligent, Adaptive Program with Just-in-time Feedback for Preservice Teachers
PI: Dr. Jiliang Tang
Role: Wrote the part on leveraging LLMs with students' feedback for better assistance.
Result: Funded by the National Science Foundation (NSF) in 2023

– Comprehensive Evaluation on the Vulnerability of LLM-based Multi-agent Systems
PI: Dr. Jiliang Tang, Dr. Hui Liu
Role: Designed and drafted the objective of formalizing attack objectives and building a multi-agent vulnerability benchmark.
Result: Submitted to the Google Faculty Award in 2025

– LLM-based Multi-agent System trust management
PI: Dr. Jiliang Tang, Dr. Yue Xing
Role: Designed and drafted the part of proposing a trust evaluation framework grounded in six trustworthiness dimensions
Result: Submitted to Open Philanthropy Technical AI Safety in 2025

– Model Context Protocol (MCP) vulnerabilities and defenses
PI: Dr. Jiliang Tang
Role: Designed and wrote the red teaming part to reveal MCP vulnerability
Result: Submitted to Open Philanthropy Technical AI Safety in 2025

– Backdoor reasoning models (Encoded reasoning in CoT and inter-model communication)
PI: Dr. Jiliang Tang
Role: Designed and wrote the threat models and detailed designs of backdoor attacks against LLM reasoning.
Result: Submitted to Open Philanthropy Technical AI Safety in 2025

Presentations

Invited Talks

– "TRAJECT-Bench: A Trajectory-Aware Benchmark for Evaluating Agentic Tool Use" 10/2025
Invited talk at Hippocratic AI

– "The evolution of AI security: from models to agents" 10/2025
Invited talk at University of Georgia

– "Understanding In-context-learning from the lens of Bayesian Statistics" 08/2025
Invited talk on JSM 2025

– "Probabilistic Categorical Adversarial Attack and Jailbreak" 09/2023
AI-Time Invited talk
– "Probabilistic Categorical Adversarial Attack and Jailbreak" 07/2023
Presentation at Okinawa Institute of Science and Technology OIST

Conference Oral Presentations

– "PROPN: Personalized Probabilistic Strategic Parameter Optimization in Recommendations" 10/2022
Presentation for CIKM 2022

Poster Presentations

– "PROPN: Personalized Probabilistic Strategic Parameter Optimization in Recommendations" 10/2022
Presentation for CIKM 2022

Teaching Experience

– Conference Tutors for KDD'22
  * Topics: Towards Adversarial Learning: from Evasion Attacks to Poisoning Attacks
– Teaching Assistant and Instructor for STT 381 Fundamentals of Data Science Methods 2021
  * Duties included discussion sessions each week, office hours and grading.
– Teaching Assistant for STT 422 Probability and Statistics II: Statistics & STT 465 Bayesian Statistical Methods 2021
  * Duties included office hours and grading.
– Teaching Assistant for STT 200 Statistical Methods & STT 231 Statistics for Scientists 2020-2021
  * Duties included online discussions, office hours and grading.
– Teaching Assistant for Introductory Applied Statistics for Engineers 2019
  * Duties included office hours and grading.

Mentoring

– Bo Wang (Female), visiting PhD student from Jilin University 08/2024 – 07/2025
Co-authored paper: ACL 2025
– Shen Dong (Male), Michigan State University, PhD student 08/2024 – present
Co-authored paper: NeurIPS 2025
Paper in submission to ACL 2025: Multi-agent vulnerability benchmark.
– Yuping Lin (Male), Michigan State University, PhD student 09/2023 – Present
Co-authored paper: EMNLP 2024
Ongoing Project: LLM retrieval head interpretation
– Jiankun Zhang (Male), visiting PhD student from Jilin University 08/2023 – present
Co-authored paper: ACL 2024, EMNLP 2025
– Yingqian Cui (Female), Michigan State University, PhD student 09/2022 – present
Co-authored paper: AISTATS 2025, ACL 2025

Services

- Serve as conference volunteers: KDD-2022

Senior Program Committee Member

– ACM Transactions on Knowledge Discovery from Data (TKDD)	2024

Program Committee Member & Conference Reviewer

– International Conference on Machine Learning (ICML)	2024-2025
– Annual Conference on Neural Information Processing Systems (NeurIPS)	2024-2025
– International Conference on Learning Representations (ICLR)	2024-2026
– Annual Meeting of the Association for Computational Linguistics (ACL)	2024-2025
– Empirical Methods in Natural Language Processing (EMNLP)	2024-2025
– AAAI Conference on Artificial Intelligence (AAAI)	2022-2026
– International Conference on Artificial Intelligence and Statistics (AISTATS)	2025-2026
– AAAI Conference on Artificial Intelligence (KDD)	2023-2025
– IEEE International Conference on Data Mining (ICDM)	2023
– SIAM International Conference on Data Mining (SDM)	2023-2025
– International World Wide Web Conference (WWW)	2024
– Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)	2023

Journal Reviewer

– Transactions on Knowledge and Data Engineering (TKDE)	2024-2025
– Transactions on Knowledge Discovery from Data (TKDD)	2024-2025
– Journal of the American Statistical Association (JASA)	2022
– Transactions on Machine Learning Research (TMLR)	2024

Volunteering

– SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)	2022
– International Conference on Machine Learning (ICML)	2023
– Conference on Neural Information Processing Systems (NeurIPS)	2023
– International Conference on Learning Representations (ICLR)	2024