

Xinnan Dai

East Lansing, Michigan State, US

Mobile: +1 (517)455-5079

Email: daipigeion1@gmail.com

Homepage: <https://ddigimon.github.io/>

Scholar: <https://scholar.google.com.hk/citations?user=LGKDd2AAAAAJ>

Research Keywords: Graphs, LLMs, Reasoning, Explanation

Education

2023 – 2028: PH.D in Computer Science

Michigan State University, US

Supervised by Jiliang Tang.

Thesis: Graph for LLMs. (In Progress)

2019 – 2023: MSc in Computer Science

ShanghaiTech University, China

Thesis: Graph for single-cell analysis

2019 – 2023: B.Eng in Computer Science

Southwest University of Science and Technology, China

Industry Experience

May 2026 – Dec 2026: Research Scientist Intern

ByteDance

- Data Selection for Continue Pretraining (27B model with selected tiktok data)

Sept 2022 – May 2023: Research Assistant

Microsoft Research Asia, China

- Graph Neural Network based Single-cell analysis for drug discovery (Nature Communication, 2024.).
- Stars of Tomorrow Internship Program (Awards)

Research Highlights

Graphs for LLMs understanding

- GRAFT(internship project in ByteDance): a data selection framework guided by latent knowledge graph topology for continual pretraining, validated across three models and ten benchmarks, improving Qwen2.5-3B's average accuracy by 15.64 percentage points and achieving approximately 2× the long-document scoring throughput of baseline model. (first author, under review)
- GraphGhost: Exploring implicit graph structures inside LLMs during step-by-step reasoning across multiple datasets. We show that LLM reasoning behavior can be controlled by manipulating graph properties (first author, under review, the extension for RAG has been accepted by KDD 2026).
- Graph-based modeling of dataset distributions and hallucinations. We use a unified graph framework to model dataset distributions and characterize LLM hallucinations as path reuse and path compression, observing these phenomena during transformer training. (first author, ICML 2026)

LLMs for Graphs Understanding

- Evaluating LLM capabilities on graph-structured data. Beyond path-searching tasks, we show that LLMs can recognize and reason about substructures within the given context (first author, ICLR 2025).
- Mechanistic analysis of substructure understanding in LLMs. We identify why LLMs can capture graph substructures and model this phenomenon via a substructure filter mechanism (first author, NeurIPS 2025).
- Circuit-level explanation of graph structure understanding. Using circuit tracing, we provide a comprehensive account of graph reasoning in LLMs, identifying a token-merging mechanism underlying this behavior (first author, NeurIPS Efficient Reasoning 2025, Spotlight).

Academic Publications

Selected Published

- “Why Retrieval-Augmented Generation Fails: A Graph Perspective”
Kai Guo, **Xinnan Dai**, Zhibo Zhang, Nuohan Lin, Shenglai Zeng, Jie Ren, Haoyu Han, Jiliang Tang
Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (**KDD**(*oral)), 2026

- “When Do Hallucinations Arise? A Graph Perspective on the Evolution of Path Reuse and Path Compression”
Xinnan Dai, Kai Yang, Cheng Luo, Shenglai Zeng, Kai Guo, Jiliang Tang
Forty-Third International Conference on Machine Learning(**ICML**), 2026
- “From Sequence to Structure: Uncovering Substructure Reasoning in Transformers”
Xinnan Dai, Kai Yang, Jay Revolinsky, Kai Guo, Aoran Wang, Bohang Zhang, Jiliang Tang
Neural Information Processing Systems (**Neurips**), 2025
- “Uncovering Graph Reasoning in Decoder-only Transformers with Circuit Tracing”
Xinnan Dai, Chung-Hsiang Lo, Kai Guo, Shenglai Zeng, Dongsheng Luo, Jiliang Tang
Neural Information Processing Systems (**Neurips**) Workshop on Efficient Reasoning (Spotlight), 2025
- “How Do Large Language Models Understand Graph Patterns? A Benchmark for Graph Pattern Comprehension”
Xinnan Dai, Haohao Qu, Yifen Shen, Bohang Zhang, Qihao Wen, Wenqi Fan, Dongsheng Li, Jiliang Tang, Caihua Shan
International Conference on Learning Representations (**ICLR**), 2025
- “AI-powered omics-based drug pair discovery for pyroptosis therapy targeting triple-negative breast cancer”
Boshu Ouyang*, Caihua Shan*, Shun Shen*, **Xinnan Dai**, Qingwang Chen, Xiaomin Su, Yongbin Cao, Xifeng Qin, Ying He, Siyu Wang, Ruizhe Xu, Ruining Hu, Leming Shi, Tun Lu, Wuli Yang, Shaojun Peng, Jun Zhang, Jianxin Wang, Dongsheng Li, Zhiqing Pang. Nature Communications, 2024, 15(1): 7560.

Selected Under review

- “Words to World: Benchmarking LLMs in Interactive Physical Environments”
Xinnan Dai, Kaiqi Yang, Yaxin Li, Cheng Luo, Wanli Li, Mingqing Xiao, Xinyang Jiang, Daoan Zhang, Kaitao Song, Fengze Liu, Zefan Cai, Caihua Shan arXiv:2610.15228, 2026.
- “GraphGhost: Tracing Structures Behind Large Language Models”
Xinnan Dai, Kai Guo, Chung-Hsiang Lo, Shenglai Zeng, Jiayuan Ding, Dongsheng Luo, Subhabrata Mukherjee, Jiliang Tang arxiv/2510.08613 (2025)
- “Revisiting the Graph Reasoning Ability of Large Language Models: Case Studies in Translation, Connectivity and Shortest Path.”
Xinnan Dai, Qihao Wen, Yifei Shen, Hongzhi Wen, Dongsheng Li, Jiliang Tang, Caihua Shan.
arxiv/2408.09529 (2024)

Skills

Python, PyTorch

Languages: English (Fluent), Chinese (Native)

Services

Reviewer for: CIKM 2024; KDD 2024-2025; Neurips 2024-; ICLR 2025-; ICML 2025-