

Hang Chen

Postdoctoral Research Fellow · Nanyang Technological University

Email chen.hang@ntu.edu.sg
WeChat [Zodiark_C](#)
Homepage zodiarkhangchen.com
Google Scholar scholar.google.com/citations?user=8-PmU7QAAAAJ



RESEARCH INTERESTS

[Mechanistic Interpretability](#) [Post-training](#) [Large Language Models](#) [Causal Discovery](#)

My research focuses on making **mechanistic interpretability practically useful** for LLM post-training: integrating interpretability with SFT and RL, and using mechanistic analysis to diagnose and guide post-training. Representative questions include identifying conflicts when fine-tuning on multiple tasks, anticipating a model's post-SFT mechanisms before training, and using interpretability signals to supervise fine-tuning. Previously, I worked on causal mechanisms of language models, mathematical models of emergence, and downstream tasks such as affective reasoning.

EDUCATION

2021 – 2026 **Ph.D. in Computer Science and Technology, Xi'an Jiaotong University**
School of Computer Science and Technology

2016 – 2020 **B.Eng. in Computer Science and Technology, Xi'an Jiaotong University**
School of Computer Science and Technology

EXPERIENCE

2026.07 – Present **Postdoctoral Research Fellow, Nanyang Technological University, Singapore**
GLINT Lab, College of Computing and Data Science (CCDS) · Supervisor: Prof. Wenya Wang

2020 – 2021 **Algorithm Engineer (Intern), Dialogue Intelligence, Du Xiaoman Financial**
Built models for emotion-cause extraction in conversations

PUBLICATIONS

Underlined: the author of this CV.

Mechanistic Interpretability

- [1] Unifying and Beyond the Theory of Pervasive Capability Degradation in LLM Post-training SFT from Feature Interference. [Hang Chen](#), Jiaying Zhu, Wenya Wang. *arXiv 2026*
- [2] Rethinking Circuit Completeness in Language Models: AND, OR, and ADDER Gates. [Hang Chen](#), Jiaying Zhu, Xinyu Yang, Wenya Wang. *NeurIPS 2025*
- [3] CLUE: Conflict-guided Localization for LLM Unlearning Framework. [Hang Chen](#), Jiaying Zhu, Xinyu Yang, Wenya Wang. *ICLR 2026*
- [4] Navigating by Old Maps: The Pitfalls of Static Mechanistic Localization in LLM Post-Training. [Hang Chen](#), Jiaying Zhu, Hongyang Chen, Hongxu Liu, Xinyu Yang, Wenya Wang. *NeurIPS 2026*
- [5] Beyond Static Interpretability: Anticipating Post-SFT Mechanisms from Pre-SFT Parameters for Better Tuning. [Hang Chen](#), Jiaying Zhu, Wenya Wang. *arXiv 2026*

Large Language Models

- [1] Quantifying Semantic Emergence in Language Models. [Hang Chen](#), Xinyu Yang, Jiaying Zhu, Wenya Wang. *ACL 2025* Oral
- [2] Skill Path: Unveiling Language Skills from Circuit Graphs. [Hang Chen](#), Xinyu Yang, Jiaying Zhu, Wenya Wang. *AAAI 2026* Oral

Affective Reasoning

- [1] How to Enhance Causal Discrimination of Utterances: A Case on Affective Reasoning. [Hang Chen](#), Xinyu Yang, Jing Luo, Wenjing Zhu. *EMNLP 2023* Outstanding Paper Nomination
- [2] Learning a Structural Causal Model for Intuition Reasoning in Conversation. [Hang Chen](#), Bingyu Liao, Jing Luo, Wenjing Zhu, Xinyu Yang. *IEEE TKDE 2024*

Causal Discovery

- [1] Towards Causal Relationship in Indefinite Data: Baseline Model and New Datasets. [Hang Chen](#), Xinyu Yang, Keqing Du. *DMLR 2025*
- [2] SSL Framework for Causal Inconsistency between Structures and Representations. [Hang Chen](#), Xinyu Yang, Keqing Du, Wenya Wang. *arXiv 2025*
- [3] A Review and Roadmap of Deep Causal Model from Different Causal Structures and Representations. [Hang Chen](#), Xinyu Yang, Chenguang Li, Keqing Du. *arXiv 2023*

HONORS & AWARDS

2025	National Scholarship, Ministry of Education of China
2017, 2018, 2019	University Scholarship, Xi'an Jiaotong University

ACADEMIC SERVICE

Conference Reviewer	NeurIPS, ICML, ICLR, ACL Rolling Review (ARR), AACL, etc.
Journal Reviewer	IEEE TKDE, IEEE TNNLS, IEEE TMM, IEEE Transactions on Affective Computing (TAC), Knowledge-Based Systems (KBS), etc.