

陈航

HANG CHEN · 南洋理工大学 博士后研究员

性别 男
出生日期 1998.01.23
电子邮箱 chen.hang@ntu.edu.sg
微信 Zodiark_C
个人主页 zodiarkhangchen.com
谷歌学术 scholar.google.com/citations?user=8-PmU7QAAAAJ



研究方向

机制可解释性 Mechanistic Interpretability 后训练 Post-training 大语言模型 LLMs 因果发现 Causal Discovery

目前致力于**机制可解释性 (Mechanistic Interpretability)**的实用性研究，包括如何将机制可解释性与 SFT、RL 等后训练技术相结合，以及如何利用机制可解释性分析和指导大语言模型的后训练。例如：多任务 SFT 时如何确定任务间冲突；如何在训练前预估模型 SFT 后的内部机制；如何利用可解释性信号监督微调过程等。此前也研究过语言模型的因果机制、涌现现象的数学建模，以及情感推理等下游任务。

教育经历

2021 – 2026 博士 · 计算机科学与技术, 西安交通大学
计算机科学与技术学院
2016 – 2020 学士 · 计算机科学与技术, 西安交通大学
计算机科学与技术学院

工作与实习经历

2026.07 – 至今 博士后研究员, 南洋理工大学 (新加坡)
计算与数据科学学院 (CCDS) GLINT Lab · 合作导师: Wenya Wang 教授
2020 – 2021 算法工程师 (实习) · 对话智能, 度小满金融有限公司
负责构建对话情感原因提取模型

学术论文

下划线标注为本人

机制可解释性 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*

大语言模型 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

情感推理 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 提名**

[2] Learning a Structural Causal Model for Intuition Reasoning in Conversation. [Hang Chen](#), Bingyu Liao, Jing Luo, Wenjing Zhu, Xinyu Yang. *IEEE TKDE 2024*

因果发现 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*

荣誉奖励

2025	国家奖学金
2017, 2018, 2019	西安交通大学校级奖学金

学术服务

会议审稿人	NeurIPS、ICML、ICLR、ACL Rolling Review (ARR)、AAAI 等
期刊审稿人	IEEE TKDE、IEEE TNNLS、IEEE TMM、IEEE Transactions on Affective Computing (TAC)、Knowledge-Based Systems (KBS) 等