

刘俊豪

地址: 北京大学新燕园校区
主页: <https://outerform.site>

电子邮件: liujunhao@pku.edu.cn

研究兴趣

我的研究方向为可解释人工智能 (XAI), 致力于开发透明、可理解的机器学习技术。具体而言, 我关注研发能帮助人们理解、信任、有效利用和控制复杂 AI 系统的可解释性方法。

教育背景

- **北京大学 · 计算机学院 · 程序设计语言研究室** 博士研究生 | 计算机软件与理论
导师: 张昕 2022.9 – 2027.6 (预计)
- **北京大学 · 信息科学技术学院** 学士 | 计算机科学与技术
GPA 排名前 11%, 北京市优秀毕业生 2018.9 – 2022.6
- **重庆市第八中学校** 初中及高中
2012.9 – 2018.6

论文发表

- **Liu, Junhao**, Haonan Yu, Zhenyu Yan, and Xin Zhang. Revitalizing Black-Box Interpretability: Actionable Interpretability for LLMs via Proxy Models. ACL 2026 主会 (Oral, 4.12%).
摘要: 大语言模型的模型无关解释因高昂的扰动成本而难以实现。我们提出基于代理模型的解释框架, 通过筛选-应用机制将低成本模型的解释迁移至目标 LLM, 以仅 9.5% 的成本实现 90% 以上的解释保真度, 并在提示压缩和有毒样本移除等下游任务中验证了其有效性。
- **Liu, Junhao**, Haonan Yu, Zhenyu Yan, and Xin Zhang. Focus-LIME: Surgical Interpretation of Long-Context Large Language Models via Proxy-Based Neighborhood Selection. IJCAI-ECAI 2026.
摘要: 随着 LLM 上下文窗口不断扩大, 在高风险任务中实现精细的特征级解释愈发重要。我们提出 Focus-LIME, 通过代理模型筛选扰动邻域, 使目标模型仅在优化后的上下文内进行细粒度归因, 实验验证了该方法的可行性与忠实性。
- **Liu, Junhao**, and Xin Zhang. ReX: A framework for incorporating temporal information in model-agnostic local explanation techniques. AAI 2025 (Oral, 4.68%).
摘要: 现有模型无关局部解释技术未能考虑时序信息, 在处理变长输入模型时效果不佳。我们提出通用框架 ReX, 通过改进采样过程与代理特征统一引入时序信息, 并在 Anchors、LIME 和 Kernel SHAP 上进行了实例化。实验结果表明, ReX 显著提升了解释保真度, 超越了现有模型专用解释技术。
- Li, Tianchi, Zhenyu Yan*, **Junhao Liu***, Peng Di, and Xin Zhang. Guiding LLM-based Loop Invariant Synthesis via Feedback on Local Reasoning Errors. TOPLAS, 2026.
摘要: 我们提出一种新框架, 通过形式化验证 LLM 的推理过程来检测局部逻辑错误, 并据此提供针对性反馈。将该框架应用于循环不变式合成问题, 所实现的工具 LORIS 在 460 个 C 程序上取得了 93.1% 的成功率, 并在非线性属性基准上表现出良好的鲁棒性。

- Yu, Haonan, **Junhao Liu**, and Xin Zhang. MAnchors: Memorization-Based Acceleration of Anchors via Rule Reuse and Transformation. ICML 2026.

摘要: Anchors 因计算开销大而难以实际部署。我们提出基于记忆的加速框架，通过缓存与复用历史解释中的中间结果，经横向和纵向转换适配新输入。在表格、文本与图像数据上的实验表明，该方法大幅缩短生成时间，同时保持解释质量。

* 共同贡献

预印本

- **Liu, Junhao**, Haonan Yu, and Xin Zhang. ConLUX: Concept-Based Local Unified Explanations. arXiv:2410.12439 (2024).

摘要: 现有基于概念的模型无关解释方法局限于归因形式。我们提出通用框架 ConLUX，利用大模型扰动将现有局部解释技术统一扩展为基于概念的解释，支持归因、充分条件和反事实三种形式。在文本、图像与多模态模型上的评估表明，ConLUX 显著提升了解释的保真度与可理解性。

- Yu, Haonan, **Junhao Liu**, Zhenyu Yan, Haoran Lin, and Xin Zhang. WASD: Locating Critical Neurons as Sufficient Conditions for Explaining and Controlling LLM Behavior. arXiv:2603.18474 (2026).

摘要: 我们提出 WASD 框架，通过定位 token 生成的充分神经元条件来解释和控制模型行为。该方法将候选条件建模为神经元激活谓词，并迭代搜索保证当前输出的最小集合。实验表明，所得解释比传统归因图更稳定、准确且简洁。

实习经历

- **小红书 · 研究实习生 (Ace 「顶尖实习生」计划)** 2026.6 – 至今
Dots 团队 北京
 - **研究内容:** 参与大语言模型预训练工作。
- **腾讯 · 研究实习生 (青云计划)** 2025.7 – 2026.5
混元多模态模型团队 北京
 - **研究内容:** 参与混元图像模型 (HunyuanImage) 的研发，专注于模型可解释性与可控性研究，探索对模型行为的透理解与精准调控方法。

竞赛获奖与在校荣誉

- **竞赛获奖**
 - ICPC EC-Final 金牌, 亚洲区域赛金牌 2018 – 2021
 - CCPC 总决赛金牌, 区域赛金牌 2019 – 2021
 - NOI 金牌 2017
- **学校奖励**
 - 优秀科研奖 2023
 - 北京市优秀毕业生 2022
 - 国家奖学金、北京大学一等奖学金、北京大学三好学生标兵 2019 – 2021

其他项目经历

- **MTML: 无数据竞争与死锁的多线程语言** 2023.3 – 2023.6

基于 OCaml 开发多线程编程语言 MTML，通过类型系统从根源上避免死锁与数据争用。项目开源于 Github:
<https://github.com/outerform/DPPL-project>.

- **基于用户的协同过滤算法的分布式实现与效率对比** 2023.5 – 2023.6
分别使用 Spark 和 Hadoop 实现基于用户的协同过滤算法，并对两者的运行效率进行比较。结果表明 Spark 在整体上表现出更高的执行效率。
- **EasyFile: 自动化电子文档处理工具** 2021.9 – 2021.12
针对 Office 文档、PDF 等处理中的重复性工作，开发自动化工具 EasyFile。项目开源于 GitHub:
<https://github.com/Yibo-He/EasyFile>。
- **基于启发式 EuSolver 的语法制导程序合成器** 2021.12 – 2022.1
实现结合 SMT 求解器的语法制导程序合成方法，针对 CLIA 与程序结构设计启发式规则，提升合成效率与效果。
- **Java 指针分析器** 2021.9 – 2021.11
开发支持流敏感、上下文敏感与域敏感分析的 Java 指针分析工具，用于辅助检测内存相关缺陷。

教学经历 (助教)

概率编程导论 (研究生课)	2024 春
离散数学导论	2024 秋
程序设计实习	2023 春
计算概论 (B)	2022 秋
数据结构与算法实习	2020 秋

专业技能与兴趣爱好

- **开发能力:** 常用 C/C++、Python; 熟悉 Linux、Git、Docker 等工具的使用
- **语言能力:** 英语六级 628
- **兴趣爱好:** 游泳, 长跑, 模拟赛车