李克钰

133-7085-6505 · chlorophyll@sjtu.edu.cn · **上海交通大学** · **计算机科学与技术**

教育背景

上海交通大学, 计算机科学与技术, 博士 2024.9 -

研究方向 AI Agent, Large Language Models

上海交通大学, 数学与应用数学, 本科 2020.9 - 2024.6

本科学积分 92.6，GPA4.03, 排名 5/44

上海市优秀毕业生，上海交通大学郭晨晨奖学金，本科生奖学金，上海交通大学优秀团员、三好学生

个人优势

- 数理基础扎实，计算机编程能力强
- 学习能力强，能够短时间快速上手新项目和探索新领域
- 善于通过各种工具（如 LLM 等）优化自己的工作流程和提高工作效率
- 熟练掌握 linux 系统, python 编程, pytorch 框架等
- 追踪前沿 agent 产品，如 Deep Research, Cursor, Dify 等

项目经历

Measuring Bias Amplification in Multi-Agent Systems with Large Language Models ,

- 第一作者 [ICLR2026] 在投
- 该项目研究了多智能体系统（MAS）中偏见放大的现象，特别是在性别、年龄和种族等敏感属性上。我们提出了 Discrim-Eval-Open 基准，用于测量系统级偏见，并引入新指标量化系统输出的极端性。实验表明，尽管单个模型偏见较小，但在 MAS 中，偏见放大现象普遍存在，并且系统倾向于偏向年轻人、女性和黑人群体。此外，我们还发现，额外的客观输入可能加剧这种偏见放大，揭示了系统脆弱性。这项研究强调了理解 LLM 系统集体偏见动态的重要性，并为开发更公平的多智能体系统提供了重要见解。

The Cost of Knowing: Hallucination Quest Game in Resource-Constrained Multi-Agent Systems ,

- 第一作者 [NeurIPS2025] 在投
- 该项目提出了 MAS-HQ 基准，用于评估多智能体系统（MAS）中的幻觉（即看似可信却错误的陈述)。与传统的静态排行榜不同，MAS-HQ 通过在资源有限的竞赛环境中，迫使智能体在准确性和信息量之间做出平衡，从而引导更有效的策略。实验结果表明，在这种竞争性设置下，智能体能够减少幻觉生成并更高效地利用资源，为评估和减轻幻觉提供了更强健的框架。

LU-500: A Logo Benchmark for Concept Unlearning ,

- 第一作者 [AAAI2026] 在投
- 该项目旨在解决公司 Logo 去学习的问题，提出了 LU-500 基准，其中包含两种不同难度的任务，用于评估 Logo 去学习的效果。通过五种创新的评价指标，从局部 Logo 区域到全局图像属性，涵盖像素和潜在空间，为复杂视觉场景提供量化分析。实验结果表明，现有的去学习方法，如 NP、SLD、SEGA 等在 Logo 去学习上表现不佳，而基于大语言模型的提示生成方法显示出显著改进。

Towards Visualizing and Understanding Bayesian Flow Networks ,

- 第一作者 [ICLR2026] 在投
- 该项目旨在深入理解贝叶斯流网络（BFNs）的内部机制，并将其应用于长期决策场景中的路径规划任务（如 Maze-2D)。通过对 BFNs 的可视化分析，我们能够有效展示训练和采样过程中轨迹的变化，从而更清晰地理解其工作原理。通过参数调优和对比实验，我们揭示了 BFNs 结构中的关键组件，并与扩散模型（DMs）进行了比较，展示了 BFNs 在路径规划任务中的优势。该研究为 BFNs 的进一步优化和解释提供了宝贵的视角。

MuBiC: Multi-view Contrastive Fusion of Biological Large Models Across Multiple Domains for Enhanced Protein Fitness Prediction ,

- 第一作者 [ICLR2026] 在投
- 该项目提出了 MuBiC 框架，结合 DNA 和蛋白质模型的嵌入相较于单一领域模型，显著提高了预测效果，验证了多领域融合的优势。