李克钰

chlorophyll@sjtu.edu.cn | @chlorophtllwzh | 13370856505
weizhihao1.github.io | +86 13370856505 | 小红书 4183134930
Google Scholar | weizhihao1 | weizhihao1

个人优势

- 时刻追踪最前沿人工智能, 金融市场资讯, 保持对最新消息的敏感度
- 熟练使用 claude code, openai codex, cursor 等提高效率, 熟悉当前最先进 AI 模型的能力边界
- 熟练掌握 linux, pytorch, verl, llama-factory, slime 等框架, 对新事物学习能力强, 上手快
- 有三年以上二级市场经验，包括 A 股，美股，港股，加密货币等

教育经历

2024 – 2029 **博士研究生, 上海交通大学** 计算机科学与技术.
研究方向: *AI 智能体, 多智能体系统, 大语言模型后训练.*
导师：王德泉老师, 刘鹏飞老师.

2020 – 2024 **本科生, 上海交通大学** 数学与应用数学.
上海市优秀毕业生, 上海交大郭晨晨奖学金, 本科生奖学金, 优秀团员, 三好学生.
学积分 92.6/100, GPA4.03/4.3, 排名 4/45.
毕业论文: 贝叶斯流网络的理解和可视化研究.

2017 – 2020 **山东省青岛第二中学**
青岛二中优秀毕业生, 三好学生, 优秀团员.

实习经历

2025.5 – 现在 **LLM Agent 算法研究**
上海创智学院刘鹏飞老师，生成式人工智能实验室 (GAIR Lab)

2024.6 – 2024.9 **量化研究员**
东吴金融科技，负责模型开发与训练，因子挖掘等工作

研究成果

Articles

1. D. Fu, Y. Wu, X. Cai, L. Ye, S. Xia, Z. Huang, W. Si, T. Xu, J. Sun, **K. Li**, et al., "Argo: Asynchronous rollout with human guidance for research agent optimization,"
2. M. Jiang, D. Fu, J. Shi, J. Zeng, W. Si, **K. Li**, X. Li, Y. Xiao, W. Li, D. Wang, et al., "Davinci-agency: Unlocking long-horizon agency data-efficiently," *arXiv preprint arXiv:2602.02619*, 2026.

3. **K. Li**, J. Shi, Y. Xiao, M. Jiang, J. Sun, Y. Wu, S. Xia, X. Cai, T. Xu, W. Si, et al., "Agencybench: Benchmarking the frontiers of autonomous agents in 1m-token real-world contexts," *arXiv preprint arXiv:2601.11044*, 2026.

4. D. Fu, Y. Wu, X. Cai, L. Ye, S. Xia, Z. Huang, W. Si, T. Xu, J. Sun, **K. Li**, et al., "Interaction as intelligence part ii: Asynchronous human-agent rollout for long-horizon task training," *arXiv preprint arXiv:2510.27630*, 2025.

5. **K. Li**, M. Jiang, D. Fu, Y. Wu, X. Hu, D. Wang, and P. Liu, "Datasetresearch: Benchmarking agent systems for demand-driven dataset discovery," *arXiv preprint arXiv:2508.06960*, 2025.

6. Y. Wu, D. Fu, W. Si, Z. Huang, M. Jiang, **K. Li**, S. Xia, J. Sun, T. Xu, X. Hu, et al., "Innovatorbench: Evaluating agents' ability to conduct innovative llm research," *arXiv preprint arXiv:2510.27598*, 2025.

7. Y. Xiao, M. Jiang, J. Sun, **K. Li**, J. Lin, Y. Zhuang, J. Zeng, S. Xia, Q. Hua, X. Li, et al., "Limi: Less is more for agency," *arXiv preprint arXiv:2509.17567*, 2025.

Conference Proceedings

1. J. Gao, J. Zhao, **K. Li**, and D. Wang, "Kan-mixer: Kolmogorov-arnold networks for gene expression prediction in plant species," in *European Conference on Computer Vision*, Springer, 2024, pp. 135–150.

项目经历

- **[ICLR 2026] Aligned Agents, Biased Swarm: Measuring Bias Amplification in Multi-Agent Systems**
  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。

- **[Submitted to COLM 2026] The Cost of Knowing: Hallucination Quest Game in Resource-Constrained Multi-Agent Systems**
  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。

项目经历 (continued)

- **[Submitted to ECCV 2026] LU-500: A Logo Benchmark for Concept Unlearning**
  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**
  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**
  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。