Keyu Li 李克钰

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

Personal Strengths

- Actively track cutting-edge AI and financial market advances, with strong sensitivity to the latest news.
- Skilled in leveraging Claude Code, OpenAI Codex, and Cursor to enhance productivity; well-versed in the capability boundaries of state-of-the-art AI models.
- Proficient in Linux, PyTorch, Verl, LLaMA-Factory, Slime; quick to learn and adapt to new tech.
- Passionate about market research, with more than three years of experience in secondary markets.

Education

2024 – 2029 **Ph.D. Candidate, Shanghai Jiao Tong University**
Computer Science and Technology.
Research Interests: *AI Agents, Multi-agent Systems, LLM Post-training.*
Advisors: Prof. Dequan Wang, Prof. Pengfei Liu.

2020 – 2024 **Bachelor's Degree, Shanghai Jiao Tong University**
Mathematics and Applied Mathematics.
Shanghai Outstanding Graduate, Guo Chenchen Scholarship of SJTU, Undergraduate Scholarship, Excellent League Member, Three-Good Student Award.
Academic Score: 92.6/100, GPA: 4.03/4.3, Rank: 4/45.
Bachelor Thesis: *Understanding and Visualization of Bayesian Flow Networks.*

2017 – 2020 **Qingdao No.2 High School, Shandong Province**
Outstanding Graduate of Qingdao No.2 High School, Three-Good Student Award, Excellent League Member.

Internship Experience

May 2025 – Present **LLM Agent Algorithm Research**
Prof. Pengfei Liu, Shanghai Innovation Institute, Generative AI Research

Jun 2024 – Sep 2024 **Quantitative Researcher**
Soochow Securities Fintech, responsible for model development and training, factor discovery, and related tasks

Research Achievements

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.

Project Experience

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

  We studied bias amplification in multi-agent systems (MAS) across gender, age, and race. We introduced the Discrim-Eval-Open benchmark and a new metric to quantify extremity in system outputs. Experiments show that, although individual models are minimally biased, MAS often amplify bias—favoring younger individuals, women, and Black populations. Additional objective inputs can worsen this effect, highlighting vulnerabilities. Our work underscores the need to understand collective bias dynamics to develop fairer MAS.

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

  We proposed MAS-HQ, a benchmark to evaluate hallucinations in multi-agent systems (MAS). Unlike static leaderboards, MAS-HQ places agents in a resource-constrained competitive environment, balancing accuracy and information output. Experiments show agents generate fewer hallucinations and use resources more efficiently, offering a robust framework for evaluating and mitigating hallucinations.

Project Experience (continued)

- **[Submitted to ECCV 2026] LU-500: A Logo Benchmark for Concept Unlearning**
  We introduced LU-500, a benchmark for company logo unlearning with tasks of varying difficulty. Using five metrics spanning local logos to global image attributes in pixel and latent space, we provide quantitative analysis in complex scenarios. Experiments show traditional unlearning methods (NP, SLD, SEGA) perform poorly, while prompt-based approaches with large language models achieve significant improvement.

- **Towards Visualizing and Understanding Bayesian Flow Networks**
  We studied Bayesian Flow Networks (BFNs) for long-horizon path planning tasks, such as Maze-2D. Visual analysis of trajectories during training and sampling provides insights into BFN mechanisms. Parameter tuning and comparisons with diffusion models highlight key components and demonstrate BFNs' advantages in path planning, offering guidance for their optimization and interpretability.

- **MuBiC: Multi-view Contrastive Fusion of Biological Large Models Across Multiple Domains for Enhanced Protein Fitness Prediction**
  Proposed the MuBiC framework, which integrates embeddings from DNA and protein models. Compared to single-domain models, MuBiC significantly improves prediction performance, validating the advantages of multi-domain fusion.