包云泰

(+86) 156-0082-5199 · baoyuntai@outlook.com · GitHub colored-dye

教育背景

浙江大学 (软件学院), 人工智能, *在读博士研究生* 2023.09–2028.06

2024–2025 学年获院设优秀实践奖学金。

浙江大学 (计算机科学与技术学院), 信息安全, 工学学士 2019.09–2023.06

2020–2021 学年和 2021–2022 学年获优秀学生称号；2021-2022 学年获浙江大学二等奖学金；2019–2020、2020–2021 学年获浙江大学三等奖学金。

学术成果及经历

- **[ICLR 2026]** Faithful Bi-Directional Model Steering via Distribution Matching and Distributed Interchange Interventions
  Yuntai Bao, Xuhong Zhang, Jintao Chen*, Ge Su, Yuxiang Cai, Hao Peng, Bing Sun, Haiqin Weng, Liu Yan, Jianwei Yin
- **[IJCAI 2025]** Scalable Multi-Stage Influence Function for Large Language Models via Eigenvalue-Corrected Kronecker-Factored Parameterization
  Yuntai Bao, Xuhong Zhang*, Tianyu Du*, Xinkui Zhao, Jiang Zong, Hao Peng, Jianwei Yin
- **[Findings of ACL 2025]** Probing the Geometry of Truth: Consistency and Generalization of Truth Directions in LLMs Across Logical Transformations and Question Answering Tasks
  Yuntai Bao, Xuhong Zhang*, Tianyu Du, Xinkui Zhao, Zhengwen Feng, Hao Peng, Jianwei Yin
- **[ACM MM 2022]** SongDriver: Real-time Music Accompaniment Generation without Logical Latency nor Exposure Bias
  第十作者 (共 11 人)，负责评估指标的设计和实验数据预处理。
- **[学术服务]** 曾在 NeurIPS 2025 Mechanistic Interpretability Workshop 担任志愿审稿人。

实习经历

浙江大学嘉兴研究院 2022.11–2023.06

本科毕业设计 赵永望教授祥云高安全无人机团队成员

- **基于 seL4 微内核实现无人机加密通信**。本人负责机载任务计算机端及地面控制站端加密通信系统的实现，包括密钥交换算法、AES 加解密算法及驱动。

社会实践/其他

灵隐街道挂职锻炼 2021.07-2021.08

参与了计算机学院与灵隐街道联合举办的**大学生挂职锻炼社会实践活动**，担任黄龙社区书记助理，协助社区治理日常事务，包括慰问空巢老人、消防巡查等。

IT 技能

- **编程语言**: Python, C/C++, Go, Rust
- **机器学习库**: PyTorch, Transformers, sklearn

个人陈述

本人导师为张旭鸿老师；我的**主要研究领域**为 LLM 机制可解释性 (mechanistic interpretability)，尤其是借助可解释性的技术和见解实现有效 (effective)、高效 (efficient) 且鲁棒 (robust) 的模型控制，当下主要研究基于表征控制的模型引导 (representation steering)。作为以可解释性见长的研究者，我拥有一般 LLM 研究者不具备的**比较优势**：我对 LLM 的内部工作机制有较深入的技术积累和见解，这些经验可以进而帮助我提出比传统机器学习路线更有针对性 (targeted)、更精细 (granular)、数据效率 (data efficiency) 更高的解决方案。

References

Yuntai Bao, Xuhong Zhang, Tianyu Du, Xinkui Zhao, Zhengwen Feng, Hao Peng, and Jianwei Yin. Probing the geometry of truth: Consistency and generalization of truth directions in LLMs across logical transformations and question answering tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 682–700, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.38. URL https://aclanthology.org/2025.findings-acl.38/.

Yuntai Bao, Xuhong Zhang, Tianyu Du, Xinkui Zhao, Jiang Zong, Hao Peng, and Jianwei Yin. Scalable multi-stage influence function for large language models via eigenvalue-corrected kronecker-factored parameterization. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 8022–8030. International Joint Conferences on Artificial Intelligence Organization, 8 2025b. doi: 10.24963/ijcai.2025/892. URL https://doi.org/10.24963/ijcai.2025/892. Main Track.

Yuntai Bao, Xuhong Zhang, Jintao Chen, Ge Su, Yuxiang Cai, Hao Peng, Bing Sun, Haiqin Weng, Liu Yan, and Jianwei Yin. Faithful bi-directional model steering via distribution matching and distributed interchange interventions. *arXiv preprint arXiv:2602.05234*, 2026.

Zihao Wang, Kejun Zhang, Yuxing Wang, Chen Zhang, Qihao Liang, Pengfei Yu, Yongsheng Feng, Wenbo Liu, Yikai Wang, Yuntao Bao, and Yiheng Yang. Songdriver: Real-time music accompaniment generation without logical latency nor exposure bias. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pp. 1057–1067, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3548368. URL https://doi.org/10.1145/3503161.3548368.