

Yuntai Bao

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

🎓 Education

School of Software Technology, Zhejiang University, Artificial Intelligence, Ph.D. student Sept. 2023–June 2028

Awarded the college's Outstanding Practice Scholarship for the 2024–2025 academic year.

College of Computer Science and Technology, Zhejiang University, Information Security, B.Eng. Sept. 2019–June 2023

Awarded Zhejiang University Third-Class Scholarship in the 2019–2020 and 2020–2021 academic years; Awarded Zhejiang University Second-Class Scholarship in 2021–2022; awarded the title of Outstanding Student in 2020–2021 and 2021–2022.

✍ Publications and Research Experience

- [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
Tenth author (11 authors total); responsible for designing evaluation metrics and preprocessing experimental data.
- [Service] Served as a volunteer reviewer for the NeurIPS 2025 Mechanistic Interpretability Workshop.

🔗 Internship Experiences

Jiaxing Research Institute, Zhejiang University

Nov. 2022–June 2023

Undergraduate thesis project Member of Prof. Yongwang Zhao's Xiangyun High-Security UAV team

- Encrypted communication for UAVs based on the seL4 microkernel. Responsible for implementing the encrypted communication system on the onboard mission computer and the ground control station, including key exchange protocols, AES encryption/decryption, and device drivers.

▣ Community Service

- July–August 2021, participated in a university-community internship program organized by the College of Computer Science and Lingyin Subdistrict, serving as assistant to the secretary of Huanglong Community; assisted with community governance tasks including visiting elderly living alone and conducting fire patrols.

⚙️ Technical Skills

- Programming languages: Python, C/C++, Go, Rust
- ML libraries: PyTorch, Transformers, scikit-learn

👤 Personal Statement

My **primary research area** is mechanistic interpretability for large language models (LLMs), particularly leveraging interpretability techniques and insights to achieve *effective, efficient, and robust* model control. My current work focuses on representation steering via representation-level interventions. As a researcher specializing in interpretability, I bring a **comparative advantage**: a deep technical understanding of LLM internal mechanisms that enables me to propose more *targeted, granular, and data-efficient* solutions than traditional machine-learning approaches.

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>.