Yuntai Bao

✉ baoyuntai@outlook.com   📍 Zhejiang, China

Education

School of Software Technology, Zhejiang University, Artificial Intelligence — Zhejiang, China
2023 – 2028

College of Computer Science and Technology, Zhejiang University, Information Security — Zhejiang, China
2019 – 2023

- Top 25% of the class (9/38).

Publications

Faithful Bi-Directional Model Steering via Distribution Matching and Distributed Interchange Interventions
Yuntai Bao
openreview.net/forum?id=LoisXFZL3k

Scalable Multi-Stage Influence Function for Large Language Models via Eigenvalue-Corrected Kronecker-Factored Parameterization
Yuntai Bao
doi.org/10.24963/ijcai.2025/892

Probing the Geometry of Truth: Consistency and Generalization of Truth Directions in LLMs Across Logical Transformations and Question Answering Tasks
Yuntai Bao
aclanthology.org/2025.findings-acl.38

Skills

Programming languages

Languages

Chinese
Native speaker

English
Fluent

Interests

Mechanistic interpretability

Representation steering