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
This paper introduces Concept Distributed Alignment Search (CDAS), a steering method that employs a distribution matching objective and distributed interchange interventions to faithfully manipulate internal concept features without overfitting to external preferences. CDAS achieves stable bi-directional control—effectively overriding safety refusals and neutralizing backdoors—while preserving general model utility.
Yuntai Bao
openreview.net/forum?id=LoisXFZL3k

Scalable Multi-Stage Influence Function for Large Language Models via Eigenvalue-Corrected Kronecker-Factored Parameterization
This paper introduces a scalable multi-stage influence function that attributes the predictions of fine-tuned LLMs back to their pretraining data, and this approach efficiently scales to billion-parameter models.
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
This paper investigates the internal representation of truth in LLMs, revealing that consistent "truth directions" emerge primarily in capable models and generalize effectively across logical transformations and diverse question-answering tasks. The truthfulness probes can be practically applied to selective question answering, improving task accuracy by filtering out incorrect model outputs.
Yuntai Bao
aclanthology.org/2025.findings-acl.38

Skills

Programming languages

Languages

Chinese
Native speaker

English
Fluent

Interests

Mechanistic interpretability

Representation steering