Yuzhang Luo

+86 137-6126-0452 | 2300010808@stu.pku.edu.cn | yuzhang527.github.io

Education

B.S. in Statistics Sep. 2023 – Jun. 2027 (Expected)
Peking University, School of Mathematical Sciences Beijing, China

- GPA: 3.67/4.00. Selected coursework: Mathematical Statistics, Applied Stochastic Processes, Applied Stochastic Calculus, Statistical Learning, Modern Statistical Modelling, Data Structures & Algorithms, Machine Learning.

Research Interests

Large language models, mechanistic interpretability, training data attribution, and alignment.

Publications and Preprints

Mechanistic Data Attribution: Tracing the Training Origins of Interpretable LLM Units
Jianhui Chen*, **Yuzhang Luo***, Liangming Pan
*International Conference on Machine Learning (ICML), 2026 — **Oral Presentation (Top 0.7%).***

From Reweighting to Rewriting: Unlocking the Intervention Effects of Influential Samples in Training Data Attribution
Yuzhang Luo*, Chenpeng Wang*, Jianhui Chen*, Liangming Pan
arXiv:2609.02771, 2026.
*Equal contribution.

Research Experience

Research Intern Jul. 2026 – Present
George Mason University — Supervisor: Prof. Ziyu Yao Remote

Mech Interp for Planning-Related World Modelling

- Designed controlled GridWorld tasks for Qwen-family agents and Q-value-derived planning signals to probe whether internal activations encode planning-relevant world-model information.
- Developed mechanistic interpretability analyses linking internal representations to planning variables, intermediate decision states, and action selection.
- Extending the framework from synthetic GridWorld environments to real-world web agents to study whether analogous planning representations emerge in long-horizon interaction.

Research Intern Sep. 2025 – Aug. 2026
PKU PILLAR Group — Supervisor: Prof. Liangming Pan Peking University

Mechanistic Data Attribution (MDA)

- Co-developed MDA, an influence-function-based framework that attributes interpretable LLM units to influential training samples, linking data provenance to the formation of internal mechanisms.
- Built controlled Pythia pre-training and counterfactual retraining experiments with data deletion/augmentation to track the emergence of induction and previous-token heads.
- Created an automated data-synthesis pipeline that generates targeted training examples to strengthen in-context learning and test how synthetic data reshapes interpretable mechanisms.

Influence-Guided Response Rewriting

- Characterized when influence estimates fail to predict the realized effect of reweighting training examples, exposing a gap between attribution scores and actual model interventions.
- Proposed influence-guided response rewriting, converting influential SFT examples into targeted content-level interventions rather than only changing sample weights.
- Implemented Tulu SFT/retraining experiments across open-weight LLMs and evaluated intervention effects on epistemic abstention and safety refusal.

Skills

Programming/ML: Python, PyTorch, TransformerLens, Hugging Face Transformers, Git; MATLAB, C.
Languages: Chinese (native), English (TOEFL 115).