

Junhao Liu

Address: Peking University (Xinyanyuan Campus), Beijing, China

Email: liujunhao@pku.edu.cn

Homepage: <https://outerform.site>

RESEARCH INTERESTS

My research area is Explainable Artificial Intelligence (XAI), dedicated to developing machine learning technologies with transparency and interpretability that are easier for humans to understand. Specifically, I focus on developing explainability technologies that help humans understand, trust, utilize, and control complex artificial intelligence systems.

EDUCATION

- **Peking University, School of Computer Science** Ph.D. Student – Computer Software and Theory
Advisor: Prof. Xin Zhang Sep. 2022 – Jun. 2027 (Expected)
- **Peking University, School of EECS** Bachelor – Computer Science and Technology
GPA ranked top 11%, Outstanding Graduate of Beijing Sep. 2018 – Jun. 2022
- **Chongqing No. 8 Secondary School** Secondary Education
Sep. 2012 – Jun. 2018

PUBLICATIONS

- **Liu, Junhao**, Haonan Yu, Zhenyu Yan, and Xin Zhang. Revitalizing Black-Box Interpretability: Actionable Interpretability for LLMs via Proxy Models. *ACL 2026, Main Conference (Oral, 4.12%)*.
Abstract: Model-agnostic explanations for large language models are often prohibitively expensive due to extensive perturbation costs. We propose a proxy-based explanation framework that transfers explanations from budget-friendly models to expensive LLMs with a screen-and-apply strategy. Our approach achieves over 90% fidelity at only 9.5% of the cost, and remains effective for downstream tasks such as prompt compression and poisoned example removal.
- **Liu, Junhao**, Haonan Yu, Zhenyu Yan, and Xin Zhang. Focus-LIME: Surgical Interpretation of Long-Context Large Language Models via Proxy-Based Neighborhood Selection. *IJCAI-ECAI 2026*.
Abstract: As LLMs scale to handle massive context windows, achieving surgical feature-level interpretation is essential for high-stakes tasks. We propose Focus-LIME, a coarse-to-fine framework that utilizes a proxy model to curate the perturbation neighborhood, allowing the target model to perform fine-grained attribution exclusively within the optimized context. Empirical evaluations demonstrate that our method makes surgical explanations practicable and faithful.
- **Liu, Junhao**, and Xin Zhang. ReX: A framework for incorporating temporal information in model-agnostic local explanation techniques. *AAAI 2025 (Oral, 4.68%)*.
Abstract: Existing model-agnostic local explanation techniques perform poorly on models with variable input lengths because they fail to consider temporal information. To address this, we propose a general framework REX that incorporates temporal information by optimizing the sampling process and surrogate features. We implement REX on Anchors, LIME, and Kernel SHAP, and validate its effectiveness across six models on three tasks. Results show that REX significantly improves explanation fidelity, surpassing state-of-the-art model-specific techniques and helping users better understand model behavior.
- Li, Tianchi, Zhenyu Yan^{*}, **Junhao Liu^{*}**, Peng Di, and Xin Zhang. Guiding LLM-based Loop Invariant Synthesis via Feedback on Local Reasoning Errors. *TOPLAS, 2026*.
Abstract: We propose a novel framework that provides constructive feedback to an LLM by formally verifying its own thinking process and detecting local reasoning errors. Applied to loop invariant synthesis, our tool LORIS achieved an overall success rate of 93.1% on a benchmark suite of 460 C programs and demonstrates robustness on challenging non-linear benchmarks.

- Yu, Haonan, **Junhao Liu**, and Xin Zhang. MAnchors: Memorization-Based Acceleration of Anchors via Rule Reuse and Transformation. ICML 2026.

Abstract: Anchors is a widely used model-agnostic explanation method but suffers from high computational cost. We propose a memorization-based acceleration framework that generates general rules as initialization and refines them through feature transformations. Experiments on tabular, text, and image data show significant speedups while preserving explanation fidelity and interpretability.

*Equal contribution

PREPRINTS

- **Liu, Junhao**, Haonan Yu, and Xin Zhang. Concept-Based Local Unified Explanations. arXiv preprint arXiv:2410.12439 (2024).

Abstract: Existing concept-based model-agnostic explanation methods are limited to attribution and cannot support richer explanation forms. We propose ConLUX, a general framework that extends local model-agnostic techniques to concept-based explanations via large pre-trained model perturbations. ConLUX supports attributions, sufficient conditions, and counterfactuals, and provides more faithful explanations for text, image, and multimodal models.

- Yu, Haonan, **Junhao Liu**, Zhenyu Yan, Haoran Lin, and Xin Zhang. WASD: Locating Critical Neurons as Sufficient Conditions for Explaining and Controlling LLM Behavior. arXiv preprint arXiv:2603.18474 (2026).

Abstract: We propose WASD, a novel framework that explains model behavior by identifying sufficient neural conditions for token generation. Our method represents candidate conditions as neuron-activation predicates and iteratively searches for a minimal set that guarantees the current output under input perturbations. Experiments demonstrate that our approach produces explanations that are more stable, accurate, and concise than conventional attribution graphs.

INTERNSHIP EXPERIENCE

- **Research Intern (ACE top-talent intern), Xiaohongshu** Jun. 2026 – Present
Dots Team Beijing, China
 - **Research Focus:** Participating in large language model pre-training.
- **Research Intern (Project Up), Tencent** Jul. 2025 – May 2026
Hunyuan Multimodal Model Team Beijing, China
 - **Research Focus:** Conducting research and development on HunyuanImage models, focusing on improving model interpretability and controllability, enabling more transparent understanding and precise manipulation of model behavior.

COMPETITION AWARDS AND SCHOOL HONORS

- **Competition Awards**
 - ICPC EC-Final Gold Medal, Asia Regional Gold Medals 2018 – 2021
 - CCPC Finals Gold Medal, Regional Gold Medals 2019 – 2021
 - NOI Gold Medal 2017
- **School Awards**
 - Outstanding Research Award 2023
 - Outstanding Graduate of Beijing 2022
 - National Scholarship, PKU First-Class Scholarship, PKU Merit Student Model 2019 – 2021

OTHER PROJECT EXPERIENCE

- **MTML: A Multi-threaded Language without Data Races and Deadlocks** Mar. 2023 – Jun. 2023
Designed a multi-threaded programming language based on OCaml, leveraging a type system to statically prevent data races and deadlocks. Open-sourced on GitHub: <https://github.com/outerform/DPPL-project>.
- **User-based Collaborative Filtering (Distributed)** May 2023 – Jun. 2023
Implemented user-based collaborative filtering using Spark and Hadoop, with a comparative study showing Spark's superior efficiency on large-scale workloads.

- **EasyFile: Automated Document Processing Tool** Sep. 2021 – Dec. 2021
 Developed an automated tool for Office and PDF processing, supporting format editing and information extraction.
 Open-sourced on GitHub: <https://github.com/Yibo-He/EasyFile>.
- **Heuristic EuSolver-based Program Synthesizer** Dec. 2021 – Jan. 2022
 Built a syntax-guided program synthesizer with heuristic rules for CLIA, improving efficiency over standard SMT-based approaches.
- **Java Pointer Analyzer** Sep. 2021 – Nov. 2021
 Implemented a Java pointer analysis tool supporting flow-, context-, and field-sensitive analysis for memory-related bug detection.

TEACHING EXPERIENCE (TEACHING ASSISTANT)

Introduction to Probabilistic Programming (Graduate Course)	Spring 2024
Introduction to Discrete Mathematics	Fall 2024
Programming Practice	Spring 2023
Introduction to Computation (B)	Fall 2022
Data Structures and Algorithms Practice	Fall 2020

PROFESSIONAL SKILLS AND HOBBIES

- **Programming Skills:** Proficient in C/C++, Python; Familiar with Linux, Git, Docker and other development tools
- **Language Skills:** CET-6: 628
- **Hobbies:** Swimming, Long-distance Running, Sim Racing