Junhao Liu

Address: Peking University (Xinyannan Campus) **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, and utilize 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

Publications and Preprints

- **Liu, Junhao**, and Xin Zhang. ReX: A framework for incorporating temporal information in model-agnostic local explanation techniques. The 39th AAAI Conference on Artificial Intelligence (AAAI-25, 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.

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

  Abstract: Model-agnostic explainability techniques are gaining demand due to their applicability across different architectures, but their reliance on low-level features limits fidelity and understandability. To address this, we propose the general framework ConLUX, which leverages large language models to uniformly extend existing techniques with concept-based local explanations. We implement ConLUX on four methods: LIME, Kernel SHAP, Anchor, and LORE, applying them to text and image models. Evaluation results demonstrate that ConLUX significantly enhances both explanation fidelity and understandability, outperforming state-of-the-art concept explainability techniques designed specifically for text and image models.

- Yu, Haonan, **Junhao Liu**, and Xin Zhang. Accelerating Anchors via Specialization and Feature Transformation. arXiv preprint arXiv:2502.11068 (2025).
  Abstract: Anchors is a popular local model-agnostic explainability technique but suffers from low computational efficiency. To address this, we propose a pre-training-based acceleration method that improves efficiency without compromising explanation quality. This method leverages the iterative optimization characteristics of Anchors by pre-training to generate general explanations as initial rules. It includes two transformation steps: horizontal transformation replaces features to adapt to current inputs, and vertical transformation further optimizes explanations. Experiments cover tabular, text, and image data, showing that this method significantly reduces generation time while maintaining high fidelity and interpretability, enabling Anchors applications in time-sensitive scenarios.

- **Liu, Junhao**, Haonan Yu, and Xin Zhang. Towards Budget-Friendly Model-Agnostic Explanation Generation for Large Language Models. arXiv preprint arXiv:2505.12509 (2025).
  Abstract: Due to architectural differences among different LLMs and the closed-source nature of some models, model-agnostic explanation techniques appear particularly promising as they do not require access to model internal parameters. However, existing model-agnostic techniques typically require multiple LLM calls to obtain sufficient samples for generating reliable explanations, leading to high economic costs. This paper demonstrates through empirical studies

that high-fidelity explanations can still be generated for large LLMs by sampling from cost-effective models. Additionally, we find that these surrogate explanations perform well in downstream tasks. Our analysis provides a new paradigm for model-agnostic explanation methods for LLMs, utilizing information from budget-friendly models for explanation generation.

Internship Experience

- **Research Intern (Project Up), Tencent** Jul. 2025 – Present

  Description: Member of the **Hunyuan Multimodal Model Team**, conducting research and development on **HunyuanImage** models. My work focuses 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. [GitHub]
- **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. [GitHub]
- **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