

Lab members and achievements



Sihong Xie
Associate Professor
Thruster of AI
HKUST(GZ)
sihongxie@hkust-gz.edu.cn

Biography: Dr. Xie held the position of Assistant Professor from 2016 to 2023 and tenured Associate Professor starting from 2023 at the Computer Science and Engineering Department of Lehigh University in USA. Dr. Xie received his Ph.D. in 2016 from the Department of Computer Science at the University of Illinois at Chicago (advised by Philip Yu, ACM/IEEE Fellow). From 2019 to 2022, he received over 1 million USD in grants from National Science Foundation (NSF) in the US. He is exploring the foundations, techniques, and applications of reliable AI, with from the perspectives of explainability, uncertainty, and robustness. Dr. Xie has published over 70 papers in major AI conferences, such as NeurIPS, ICLR, KDD, UAI, ICDM, WWW, AAAI, IJCAI, and TKDE, with over 2,700 citations and an h-index of 20. He has served as an NSF panelist, senior PC or area chair for AAAI and WWW, and PC member for conferences such as NeurIPS, ICML, ICLR, AISTAT, KDD, WWW, AAAI, ICDM, and SIGIR, a reviewer for TPAMI and TKDE, an organizer for the KDD TrueFact/MIS2-TrueFact Workshops. Dr. Xie's research is recognized via the NSF CAREER Award (2022) and the Tencent Rhino-Bird in 2024. In 2024, he is recognized as excellent youth scholar on the national, Guangdong provincial, and Guangzhou municipal levels. He is serving as an executive committee member of CCF-AI, CCF-BigData, ACM SIGSPATIAL China, and CIPS.



Rui Xu
Bachelor from University of Pittsburgh and Sichuan University. Master from UC Berkeley.
Research Interest: Uncertainty quantification.



Yazheng Liu
Bachelor & Master from Beijing University of Posts and Telecommunications. Published on ICLR.
Research Interest: Explaining graph neural networks on dynamic graphs.



Xiaqiang Tang
Master from Tongji University, interned at Tencent AI Lab.
Research Interest: Trustworthy LLMs using data-centric and model alignment approaches.



Jiuju Chen
Master from UESTC. Published on NeurIPS. Interned at Tencent and HUAWEI.
Research Interest: graph foundation model in finance.



Rufeng Chen
Master from Hangzhou Dianzi University, published on AAAI, NIPS.
Research Interest: Safe Reinforcement Learning and generative models.



Zhaofan Zhang
Master from the University of Macau, published on AAAI.
Research Interest: safe decision-making under uncertainty for robotics.

Reliable Foundation Model

Reliable LLM: data-centric method

- Background & Challenge**
Unlike verifiable factual statements, no cognitive standard and corresponding fine-grained dataset to support reliable LLM research.
- Solution**

User: How about healthy and safe eating for someone with age requirements for setting alcohol, especially regarding assistance to their spouse?

Output: Your model should provide an answer to the question by focusing on the user's intent and not being misled by irrelevant details. The old saying of 'eat what you like and drink what you want' is not applicable here. The Alcohol consumption should be limited to the age of 21 or with an adult over the age of 21 in order to enter the store. The Alcohol stores should understand identification from all states, whether on, on-board and when you are on-board.

Abstract: **Reliable LLMs require data-centric methods to filter out unreliable information and enhance the model's ability to handle complex, multi-step tasks. This paper introduces a data-centric method for LLMs, which involves fine-tuning the model on a dataset of high-quality, verifiable factual statements. The method aims to improve the model's ability to generate accurate and reliable responses, particularly in domains where safety and accuracy are critical. The proposed method is evaluated on various benchmarks, demonstrating its effectiveness in improving the model's performance on tasks requiring reliable information. The method is implemented using a combination of data filtering and fine-tuning techniques, resulting in a more robust and reliable LLM. The method is applicable to a wide range of LLMs and can be adapted to different domains and tasks. The method is a significant step towards building more reliable and trustworthy LLMs.**

Factual statement Cognitive statement

Reliable AI foundation

Uncertainty Quantification

- Background & Challenge**
Conformal prediction uses calibration data to construct $1 - \alpha$ confidence set of the prediction of a test input x . When calibration and test samples are drawn from different distributions (non-exchangeable), the prediction interval can not hold $1 - \alpha$ confidence.

Solution: **Reduced W-distance** (Wasserstein distance) between calibration and test distributions.

Call. Res. Aligned Test Res. Test Res.

Reliable Agent & Robotics

Decision Making Under Uncertainty

- Background & Challenge**
Uncertainty in perception

- Solutions**
Distributionally robust Optimization
Distributionally RL

Reliable GFM: weakly supervised method

- Background & Challenge**
GFM are hard to train due to the lack of ground-truth labels for contrastive learning

Pretraining Datasets Adaptation Data

Protocol Task: Contrastive Learning

Weak supervision for positive and negative sample construction

Downstream Tasks
Node-level Edge-level Graph-level

Explanations of GNN on evolving graphs

- Background & Challenge**

Unbalanced node degree distribution Complex message aggregation process High-dimensional node features

Solution: **Important paths** extraction.

Robotics with Generative Model

- Background & Challenges**
Planning optimization replaces the traditional model-based approach and integrates learning dynamics and policy.
Diffusion model is time-consuming and difficult to deploy in real world.

- Solution**
Hierarchical & cost-sensitive planning

Abstractive Graph
high level low level

Applications

AI characters and education

Creativity: **AI Characters:** What They Can Do and What They Can't Do. **Reliability:** Focus on being understandable rather than accurate.

Reliability: **Daily AI agent:** What They Can Do and What They Can't Do. **Reliability:** Focus on being understandable rather than accurate. **AI Judge - Trader:** What They Can Do and What They Can't Do. **Reliability:** Focus on being understandable rather than accurate.

Graph anomaly detection on large-scale social-financial network.

Smart Healthcare

- Medical Image Segmentation**
Variations in scanning equipments, patient differences, the severity of diseases, all can lead to distribution shift in medical images.

Robust conformal prediction leads to better clinician decision-making.

- Explainable Multi-modal models**

Robotics for Social Good

- Drones**
River Patrol Inspection Urban and Rural Delivery
- Robotic Arm**
Vehicle Welding 3D Printing
- AMR**
Catering Services Cargo Transportation

Collaborators



Welcome to join ExRAIL

