Yuxuan Zhu

+1 5189619609 · zhuy27@rpi.edu · Google Scholar · linkedin · Homepage

Education

Rensselaer Polytechnic Institute (RPI) **Aug. 2023 – Now**
***Ph.D** in Computer Science* *Troy, New York, USA*

Leiden University **Feb. 2021 – Mar. 2023**
***Master of Science** in Computer Science* *Leiden, Netherlands*

Sichuan University **Sept. 2016 – June 2020**
***Bachelor of Engineering** in Electrical Engineering* *Chengdu, China*

Research Interest

Efficient LLM Inference, Flow Matching for Reasoning, Anomaly Detection, Graph Neural Networks

Publications

SentenceKV: Efficient LLM Inference via Sentence-Level Semantic KV Caching. [**Conference**]
Zhu, Y., Falahati, A., Yang, D. H., & Amiri, M. M.
Conference on Language Modeling (COLM), Montreal, Canada, Oct. 2025.
https://arxiv.org/abs/2504.00970

On the Robustness of Graph Reduction Against GNN Backdoor. [**Conference**]
Zhu, Y., Mandulak, M., Wu, K., Slota, G., Jeon, Y., Chow, K. H., & Yu, L.
AISec@CCS, Proceedings of the Workshop on Artificial Intelligence and Security, Salt Lake City, Nov. 2024.
https://dl.acm.org/doi/10.1145/3689932.3694762

A Survey on Explainable Anomaly Detection. [**Journal**]
Li, Z., **Zhu, Y.** & van Leeuwen, M.
ACM Transactions on Knowledge Discovery from Data (TKDD) - Top 5 downloaded papers.
https://doi.org/10.1145/3609333

Scalable, Explainable and Provably Robust Anomaly Detection with One-Step Flow Matching
Li, Z., Huang, Q., **Zhu, Y.**, Yang, L., Amiri, M. M., van Stein, N. & van Leeuwen, M.
[**Under Review**]. May 2025.

Research Experience

Research Assistant, Amiri Lab **Sept. 2024 – Now**
Advisors: Mohammad Mohammadi Amiri (Assistant Professor) *Troy, USA*
- Research on efficient LLM inference via **KV cache compression** to optimize memory usage and latency.
- Research on efficient LLM reasoning via **CoT compression and pruning**.

Research Assistant, Data Security and Privacy Lab **Aug. 2023 – Aug. 2024**
Advisors: Lei Yu (Assistant Professor) *Troy, USA*
- Research on **backdoor attacks**, faced by Graph Neural Networks when applied to large-scale graph data.
- Research the training data leakage problem (which is called **Membership Inference Attack**) on LLMs

Research Assistant, Explanatory Data Analysis group **Nov. 2021 – Sept. 2022**
Advisors: Matthijs van Leeuwen (Associate Professor), Zhong Li (PhD candidate) *Leiden, Netherlands*
- Research on **interpretable anomaly detection** methods

Work Experience

Student Researcher, Online Low-Rank KV-Cache Compression **May. 2025 – Aug. 2025**
Advisors: Mohammad Mohammadi Amiri, Pin-Yu Chen *IBM, NY, USA*
- Provide an online KV cache dimension compression method for efficient LLM inference.
- This work led to a conference paper submission.

Research Intern, Inkjet Failures Detection and Classification **Sept. 2022 – Mar. 2023**
Advisors: Fatima Abidine, Matthijs van Leeuwen *Canon, Netherlands*
- Identified anomalies using industrial time-series data; Labeled outlier clusters with expert knowledge.

Skills

Languages: Python, C++, Matlab
Frameworks: PyTorch, HuggingFace Transformers
Tools: Git, LaTeX, Linux