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

LLM efficiency (Inference, Reasoning, Long-context, Memory); Multi-agent system efficiency.

Publications

[C1] **SentenceKV: Efficient LLM Inference via Sentence-Level Semantic KV Caching.**
Y Zhu, A Falahati, DH Yang, MM Amiri
Conference on Language Modeling (COLM), Oct. 2025. [paper] [code]

[C2] **On the Robustness of Graph Reduction Against GNN Backdoor.**
Y Zhu, M Mandulak, K Wu, G Slota, Y Jeon, KH Chow, L Yu
Proceedings of the Workshop on Artificial Intelligence and Security (AISec@CCS), Nov. 2024. [paper]

[C3] **Scalable, Explainable and Provably Robust Anomaly Detection with One-Step Flow Matching**
Z Li, Q Huang, **Y Zhu**, L Yang, MM Amiri, N van Stein, M van Leeuwen
Conference on Neural Information Processing Systems (NeurIPS), Dec. 2025. [paper]

[J1] **A Survey on Explainable Anomaly Detection.**
Z Li, **Y Zhu**, M Van Leeuwen
ACM Transactions on Knowledge Discovery from Data (TKDD) - Top 5 downloaded papers. [paper]

[P1] **OjaKV: Context-Aware Online Low-Rank KV Cache Compression with Oja's Rule.**
Y Zhu, DH Yang, MM Amiri, K Murugesan, T Pedapati, PY Chen
Under Review of ICLR 2026 [paper]

[P2] **ZoomKV: Memory Efficient Reasoning through Multi-Granularity Key Value Retrieval**
DH Yang, **Y Zhu**, MM Amiri, K Murugesan, T Pedapati, S Chaudhury, PY Chen
Under Review of EACL 2026

Research Experience

Research Assistant, Amiri Lab — **Sept. 2024 – Present**
Advisor: Mohammad Mohammadi Amiri (Assistant Professor) — *Troy, USA*

- Engineered and evaluated novel methods for LLM efficiency, focusing on KV cache compression and long-context inference.
- Led the design of SentenceKV, a semantic caching method that improved long-context latency by **up to 4x** while compressing the KV cache by **over 80%**. Prepared the open-sourced codes and presented at COLM 2025.
- Built and automated end-to-end evaluation pipelines for **4 diverse benchmarks** (LongBench, RULER, NIAH, lm-eval-harness), reducing evaluation time for new KV cache compression methods.

Research Collaborator, IBM Research **May. 2025 – Present**

Mentors: Pin-Yu Chen, Karthikeyan Murugesan, Mohammad Mohammadi Amiri *Yorktown Heights, USA*

- Led OjaKV, an online low-rank KV cache update rule that reduced memory usage by **40%** with less than **5%** impact on model accuracy. Prepared submission package and open-sourced codes.
- Coordinated multi-institutional research efforts between teams, ensuring synchronized experiments for **1 top-tier conference submissions**.
- I am currently exploring memory-sharing mechanisms in multi-agent systems to improve communication efficiency; also working on enhancing LLM's reasoning efficiency.

Research Assistant, Data Security and Privacy Lab **Aug. 2023 – Aug. 2024**

Advisor: Lei Yu (Assistant Professor) *Troy, USA*

- Investigated the interaction between graph reduction techniques and backdoor attacks, improving defense detection rates by **15%** on sparse graphs. Published findings at **AISec@CCS 2024**.
- Analyzed privacy risks in LLMs, demonstrating a novel membership inference attack that successfully identified training data presence.

Research Assistant, Explanatory Data Analysis Group **Nov. 2021 – Sept. 2022**

Advisors: Matthijs van Leeuwen, Zhong Li *Leiden, Netherlands*

- Contributed key analyses and writing to a comprehensive survey on explainable anomaly detection, published in **ACM TKDD** and cited **150+ times**.

Research Collaborator, Explanatory Data Analysis Group **Jan. 2025 – Present**

Collaborators: Zhong Li, Matthijs van Leeuwen *Leiden, Netherlands / Remote*

- Co-developed a novel one-step flow matching method for anomaly detection that achieved state-of-the-art results on over **40 benchmark datasets and baseline methods**, leading to a publication at **NeurIPS 2025**.

Work Experience

Research Intern, Inkjet Failures Detection and Classification **Sept. 2022 – Mar. 2023**

Advisors: Fatima Abidine, Matthijs van Leeuwen *Canon, Netherlands*

- Developed a data-driven K-Nearest Neighbors (KNN) based algorithm to detect anomalies in industrial piezoelectric self-sensing time-series data.
- Applied unsupervised and semi-supervised learning techniques to cluster detected anomalies, which helped identify potential limitations of the company's existing health monitoring system.
- Analyzed clustering results with domain experts and proposed actionable recommendations to improve the accuracy and robustness of the existing nozzle failure detection methods.

Academic Services

Reviewer, Conference on Neural Information Processing Systems (NeurIPS), 2024,2025
Reviewer, International Conference on Distributed Computing Systems (ICDCS), 2024
Reviewer, Transactions on Information Forensics & Security
Reviewer, The International Conference on Learning Representations (ICLR), 2025
Reviewer, Empirical Methods in Natural Language Processing (EMNLP), 2025

Skills

Programming & Frameworks: Python, PyTorch, Hugging Face Transformers.
LLMs & VLMs: LLaMA 3.1, Qwen 2.5, Mistral, LongChat, DeepSeek-R1 (inference).
Fine-tuning & Adaptation: LoRA, Prefix/Prompt Tuning.
Inference & Systems: FlashAttention, CUDA, quantization (INT8/4-bit).
KV Cache & Efficiency: KV-cache compression and management, long-context optimization.
Tooling & Ops: Git, GitHub, Weights & Biases, Linux, Slurm.
Evaluation & Benchmarks: lm-eval-harness, LongBench, NIAH, RULER, GSM8K.