

YUXUAN ZHU

☎ +1 5189619609 · ✉ zhuy27@rpi.edu · 📄 [Google Scholar](#) · 🔗 [linkedin](#) · 🌐 [Homepage](#)

Education

Rensselaer Polytechnic Institute (RPI)

Ph.D in Computer Science

Aug. 2023 – Now

Troy, New York, USA

Leiden University

Master of Science in Computer Science

Feb. 2021 – Mar. 2023

Leiden, Netherlands

Sichuan University

Bachelor of Engineering in Electrical Engineering

Sept. 2016 – June 2020

Chengdu, China

Research Interest

LLM efficiency (Inference, Reasoning, Memory, KV Cache); 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] **OjaKV: Context-Aware Online Low-Rank KV Cache Compression.**

Y Zhu, DH Yang, MM Amiri, K Murugesan, T Pedapati, PY Chen

Annual Meeting of the Association for Computational Linguistics (**ACL 2026**) [\[paper\]](#)

[C4] **ZoomR: Memory Efficient Reasoning through Multi-Granularity Key Value Retrieval.**

DH Yang, **Y Zhu**, MM Amiri, K Murugesan, T Pedapati, S Chaudhury, PY Chen

Annual Meeting of the Association for Computational Linguistics (**ACL 2026 Oral**) [\[paper\]](#)

[C5] **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. [\[paper\]](#)

[C6] **MM-OptBench: A Solver-Grounded Benchmark for Multimodal Optimization Modeling.**

Z Li, Q Huang, **Y Zhu**, M Amiri, N van Stein, T Back, M Van Leeuwen, Z Wen, L Yang

Conference on Neural Information Processing Systems (**NeurIPS**), Dec. 2026. [\[paper\]](#)

[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. [\[paper\]](#)

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%**.
- 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, Keerthiram 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.

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**.

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**.

Work Experience

Research Intern, IBM Research

May 2026 – Aug. 2026

Advisor: Keerthiram Murugesan

Yorktown Heights, NY, USA

- Conduct research on efficient inference and memory management for agentic LLM systems.
- Explore reusable internal representations to reduce repeated computations.
- Build and evaluate prototype systems for studying the accuracy, latency, and memory trade-offs of long-context LLM workloads.

Research Intern, Canon

Sept. 2022 – Mar. 2023

Advisors: Fatima Abidine, Matthijs van Leeuwen

Venlo, 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), 2024,2025,2026

Reviewer, Conference on Language Modeling (COLM), 2026

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.