Zixuan (Eve) Yi

zixy@upenn.edu zixy17.github.io

Education

University of Pennsylvania, PA, United States — Aug 2023 - Present
Ph.D. in Computer and Information Science, GPA: 4.0/4.0
Advisors: Ryan Marcus and Zack Ives
M.S. in Computer and Information Science — Aug 2023 - Dec 2025

Tsinghua University, Beijing, China — Aug 2019 - Jun 2023
B.Eng. in Computer Science and Technology

University of California, Berkeley, CA, United States — Jan 2022 - May 2022
Visiting Undergraduate Student

Research Interests

Efficient and Reliable AI Systems; Agentic Systems; Machine Learning Systems.
I study how learning and systems techniques can make intelligent systems more efficient, adaptive, and reliable under real-world constraints.

Publications

Zixuan Yi, Yuanming Shao, Shaun Wallace, Zachary G. Ives, and Ryan Marcus. "Data Canvas: A Provenance-Guided Harness for Agentic Data Engineering.", **COLM'26**

Marcus J. Min, Mike He, Zhaoyu Li, **Zixuan Yi**, Sharad Malik, Aarti Gupta, Xujie Si, and Osbert Bastani. "Theory-Level Autoformalization: From Isolated Statements to Unified Formal Knowledge Bases." **ICML'26 Position Track (Spotlight)**.

Zixuan Yi, Yao Tian, Zachary G. Ives, and Ryan Marcus. "LimeQO: Low-Rank Learning for Offline Query Optimization." **SIGMOD'25**.

Peter Akioyamen, **Zixuan Yi**, and Ryan Marcus. "The Unreasonable Effectiveness of LLMs for Query Optimization." **MLSystems@NeurIPS'24 (Spotlight)**.

Zixuan Yi, Sami Abu-El-Haija, Teja Vemparala, Yawen Wang, Yannis Chronis, Yu Gan, Michael Burrows, Carsten Binnig, Bryan Perozzi, Ryan Marcus, and Fatma Ozcan. "Is It Bigger than a Breadbox: Efficient Cardinality Estimation for Real-World Workloads." **In Submission**.

Research Experience

University of Pennsylvania with Prof. Ryan Marcus and Zack Ives

Data Canvas – Reliable Agentic Systems — Feb 2025 - April 2026

- Developed Data Canvas, a provenance-guided harness that turns open-ended information-seeking tasks into structured semantic-operator workflows with explicit intermediate state.
- Designed fine-grained provenance linking sources, agent actions, intermediate results, rows, and cells, making failures attributable and allowing sparse feedback to propagate across related outputs.
- Built targeted repair that replays only affected computation; on WideSearch, improved row/item F1 from 21.8/43.4 to 28.9/55.9, while repair was 35× cheaper and 9.7× faster than full re-execution.

LimeQO – Efficient Learning under Expensive System Feedback Sep 2023 - Nov 2024

- Framed offline workload optimization as learning under expensive execution feedback and a limited exploration budget.
- Modeled query–configuration runtimes as a partially observed low-rank matrix, using shared workload structure to infer unobserved outcomes and prioritize informative executions.
- Achieved up to 2× workload speedup with up to 100× lower learning overhead than deep-learning alternatives while enforcing a no-regressions guarantee.

UC Berkeley RISELab with Prof. Aditya Parameswaran Feb 2022 - Dec 2022
FormS: A Python Library for Scalable Spreadsheet Formula [project]

- Designed a scalable spreadsheet formula execution engine on distributed frameworks such as Ray and Dask, translating formulas into execution plans and rewrite rules for interactive execution on larger datasets.
- Improved end-to-end execution through load balancing, physical partitioning, and hierarchical clustering to reduce computation and communication costs, and evaluated performance across execution strategies.

Industry Experience

Microsoft Corporation May 2026 - Aug 2026
Research Intern at Microsoft Research Redmond, WA

- Developed cost-aware methods for selecting execution configurations when feedback is expensive or censored, using adaptive timeouts to allocate validation budget toward informative candidates.
- Evaluated configuration-selection strategies for groups of parameterized SQL queries on SQL Server, studying the trade-off between decision quality, validation cost, and end-to-end workload runtime.

Google May 2024 - Aug 2024
Student Researcher at SystemsResearch@Google Sunnyvale, CA

- Developed lightweight online learning methods that adapt from historical execution behavior on recurring workloads.
- Applied hierarchical models to history-based cardinality estimation, enabling fast updates and low-overhead inference and achieving 2×–10× better performance on data-warehouse-style workloads.

Google Jul 2021 - Sep 2021
Engineering Practicum Intern

- Built a human-in-the-loop analytics pipeline combining Search Language Specialist feedback with NLP models for evaluating Google Search and Assistant queries.
- Developed production tooling with a Go frontend, C++ backend, and Spanner storage; user feedback reported more than 50% reduction in workflow time.

Awards

AWS ASSET Ph.D. Fellowship 2026