Haoyu Li

Email : lhy@utexas.edu

About Me

I am a Ph.D. student at UT Austin, advised by Prof. Aditya Akella and Prof. Venkat Arun. My research applies AI techniques to improve performance (e.g., throughput, latency) and usability of modern computer systems, such as data analytics, ML training/serving frameworks, OS, and networks.

Education

- **The University of Texas at Austin** — Austin, TX
  Ph.D. in Computer Science. — *2023 – 2028*
- **Peking University** — Beijing, China
  B.S. in Computer Science. — *2019 – 2023*

Experience

- **Microsoft Research** — Redmond, WA
  Research Intern — *May 2024 - Aug 2024*
  - **LLMs for Data Warehousing**: Integrated LLMs into database query processing pipelines to enable copilot-style, type-as-you-go data analysis.
    - First proposed LLM-guided speculative execution for ad-hoc queries, enabling real-time feedback and result previews as users compose their queries.
    - Implemented the system in Amazon Redshift, integrating a closed-loop framework featuring LLM-based debugging and prediction, rule-based SQL rewriting, speculative query execution, multi-layer result caching, and an interactive user interface.
    - Conducted a user study with 24 participants (A/B test) for data analysis tasks, showing up to a 289× reduction in user-perceived latency and statistically significant improvements in task completion time ($p < 0.05$).
    Preprint: Speculative Ad-hoc Querying.
- **UT Austin** — Austin, TX
  Graduate Research Assistant — *Sep 2023 – Present*
  - **Machine Learning Systems**: Proposed a lossless gradient compression technique to accelerate distributed deep learning training. Implemented the system using CUDA and low-level networking components to accelerate the gradient aggregation phase. The algorithm shows a 6.33× improvement in throughput and a 3.74× increase in per-iteration training speed on DNN workloads.
    Preprint: Accelerating Distributed Deep Learning using Lossless Homomorphic Compression.
- **Peking University** — Beijing, China
  Research Assistant — *2022 – 2023*
  - **Information Theory**: Proposed a space decomposition theorem for general membership problems, which yields a clean and optimal mathematical characterization of this class. In particular, we first establish a space lower bound for binary classification with one-sided error, without making any distributional assumptions.
    Publication: ChainedFilter: Combining Membership Filters by Chain Rule.
  - **Stream Processing**: Proposed a compact counter encoding method that enables fast and accurate per-flow frequency measurement, improving accuracy by 50% and speed by 33% over state-of-the-art methods.
    Publication: Stingy Sketch: A Sketch Framework for Accurate and Fast Frequency Estimation.

Honors and Awards

- **SenseTime Scholarship, SenseTime**: 2022
- **National Scholarship, China**: 2020, 2021

Preprints

1. **Haoyu Li**, Srikanth Kandula, Maria Angels de Luis Balaguer, Aditya Akella, Venkat Arun. *Speculative ad-hoc querying.*, preprint arXiv:2503.00714, 2025.

2. **Haoyu Li**, Yuchen Xu, Jiayi Chen, Rohit Dwivedula, Wenfei Wu, Keqiang He, Aditya Akella, Daehyeok Kim. *Accelerating distributed deep learning using lossless homomorphic compression*, preprint arXiv:2402.07529, 2024.

Peer-reviewed Conference Publications

1. **Haoyu Li**, Liuhui Wang, Qizhi Chen, Jianan Ji, Yuhan Wu, Yikai Zhao, Tong Yang, Aditya Akella. *ChainedFilter: Combining membership filters by chain rule.* In Proceedings of the ACM SIGMOD/PODS International Conference on Management of Data (SIGMOD), 2024.

2. **Haoyu Li**, Qizhi Chen, Yixin Zhang, Tong Yang, Bin Cui. *StingySketch: A sketch framework for accurate and fast frequency estimation.* In Proceedings of the VLDB Endowment (PVLDB), 2022.

3. Zhuochen Fan, Bowen Ye, Ziwei Wang, Zheng Zhong, Jiarui Guo, Yuhan Wu, **Haoyu Li**, Tong Yang, Yaofeng Tu, Zirui Liu, Bin Cui. *Enabling space-time efficient range queries with REncoder.* The VLDB Journal, Volume 33, Issue 6, pp. 1837–1859, November 2024. Springer Berlin Heidelberg.

4. Ziyang Wang, Zihan Zhong, Jiayi Guo, Yuhan Wu, **Haoyu Li**, Tong Yang, Yuxuan Tu, Hongzhi Zhang, Bin Cui. *REncoder: A space-time efficient range filter with local encoder.* In Proceedings of the 2023 IEEE 39th International Conference on Data Engineering (ICDE), pp. 2036–2049.

5. Xinyi Li, Zhenyu Fan, **Haoyu Li**, Zihan Zhong, Jiayi Guo, Shouke Long, Tong Yang, Bin Cui. *SteadySketch: Finding steady flows in data streams.* In Proceedings of the 2023 IEEE/ACM 31st International Symposium on Quality of Service (IWQoS), pp. 01–09.

6. Ruoyu Qiu, Yuchen Ming, Yifan Hong, **Haoyu Li**, Tong Yang. *Wind-bell Index: Towards ultra-fast edge query for graph databases.* In Proceedings of the 2023 IEEE 39th International Conference on Data Engineering (ICDE), pp. 2090–2098.