Last updated in September 2024

Hongbin Zhong

Atlanta, US hzhong81@gatech.edu 470-437-8071 https://rjzhb.github.io/ Hongbin Zhong
rjzhb

Research Interest

Retrieval-Augmented Generation (RAG) Systems
Data-Centric AI
Data Systems for Machine Learning
Data Analytics Systems

Education

Georgia Institute of Technology *2024–2029 (expected)*
Ph.D. in Computer Science
Advisor: Kexin Rong

Northeastern University *2020–2024*
B.S. in Computer Science

Publications

1. **FaDE: More Than a Million What-ifs Per Second**
   Haneen Mohammed*, Alexander Yao*, Charlie Summers*, **Hongbin Zhong**, Gromit Yeuk-Yin Chan, Subrata Mitra, Lampros Flokas, Eugene Wu
   *Submitted to VLDB 2025(**Accept with Shepherding**)*

2. **PECJ: Stream Window Join on Disorder Data Streams with Proactive Error Compensation**
   Xianzhi Zeng[*], Shuhao Zhang, **Hongbin Zhong**, Hao Zhang, Mian Lu, Zhao Zheng, Yuqiang Chen
   SIGMOD 2024

3. **Accelerating Deletion Interventions on OLAP Workload**
   Haneen Mohammed, Alexander Yao, Lampros Flokas, **Hongbin Zhong**, Charlie Summers, Eugene Wu
   ICDE 2024

Research Experience

Research Assistant, Georgia Institute of Technology, Atlanta, GA *Aug 2024 – Present*
Advisor: Kexin Rong ; Collaboration: VMware System Group

- Led research on fine-grained access control in **vector databases** for **RAG**, enhancing data confidentiality in enterprise applications.
- Developed PostgreSQL and pgvector solutions using row-level security, pre-filtering, and post-filtering, optimizing storage cost and retrieval performance.
- Designed and solved **optimization models** to reduce storage redundancy and improve query speed through efficient document partitioning and indexing.

Research Assistant, Columbia University, New York City, NY *Jul 2023 – Nov 2023*
Advisor: Eugene Wu

- FADE Project - Developed optimization techniques for sparse matrix evaluations, improving performance.
- Applied SIMD and multithreading for sparse data evaluations, reducing disk I/O significantly.

Research Assistant, Rutgers University, New Jersey *Jul 2023 – Sep 2023*
Advisor: Dong Deng

- Implemented baseline methods for data similarity tasks and assisted with running experiments.

- Optimized parallelization for group function tasks in data processing.

Research Assistant, Nanyang Technological University / 4Paradigm, Singapore *Jan 2023 – Jul 2023*
Advisors: Mian Lu, Shuhao Zhang

- Developed high-accuracy, low-latency stream processing system for out-of-order data.
- Implemented Bayesian variational inference with transformers for complex data streams.

Industry Experience

Database Internals Engineer, InfiniFlow**(vector database startup)** *Mar 2024 – Apr 2024*

- Improved the mechanism for recording the oldest visible timestamp to avoid unnecessary access to ‘txn_map‘.
- Optimized the cleanup process for bulk deletion of files and records, significantly reducing file I/O operations.

Full Stack Software Engineer(part-time), 4Paradigm *Feb 2024 – Apr 2024*

- Enhanced AI assistant server performance by refining cache systems, reducing system overhead, and improving user access speed.
- Developed backend logic for community features, and implemented timed tasks for data updates using asynchronous programming.

Backend Software Engineer, Meituan, Beijing *Apr 2022 – Sep 2022*

- Contributed to the Meituan App’s short video project by building foundational features.
- Developed a data reporting pipeline using Kafka and Hive to support recommendation algorithms.
- Improved user experience under poor network conditions by implementing periodic data refreshes through scheduled tasks.

Technologies

Languages: C++, C, Java, Python, C#, SQL

Technologies: CUDA, Compiler, Database, Deep Learning System, .NET, OS