

Hongbin Zhong

📍 Atlanta, US ✉ hzhong81@gatech.edu ☎ 470-437-8071 🔗 <https://rjzhb.github.io/> in Hongbin Zhong
🌐 rjzhb

Research Interest


AI Agents & Cognitive Architectures for Planning, Reasoning, and Memory
High-Performance and Secure Data Retrieval & Storage Systems
End-to-End LLM Inference with Information Retrieval

Education

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

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

Preprints And Draft

- **Beyond Screenshots: A Dynamic State-Machine Memory and Global Programmatic Planner for Web Agents**
Hongbin Zhong* with Microsoft Research collaborators
draft
- **HoneyBee: Efficient Role-based Access Control for Vector Databases via Dynamic Partitioning**
Hongbin Zhong, Matthew Lentz, Nina Narodytska, Adriana Szekeres, Kexin Rong
Under Revision for **SIGMOD 2026**
[arXiv](#) 

Publications

- **Fast Hypothetical Updates Evaluation**
Haneen Mohammed*, Alexander Yao*, Charlie Summers*, Hongbin Zhong, Gromit Yeuk-Yin Chan, Subrata Mitra, Lampros Flokas, Eugene Wu
SIGMOD 2025 Demo
- **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
VLDB 2025
- **Accelerating Deletion Interventions on OLAP Workload**
Haneen Mohammed, Alexander Yao, Lampros Flokas, Hongbin Zhong, Charlie Summers, Eugene Wu
ICDE 2024
- **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

Experience

- AI-Focused Work • Systems-Focused Work
- **Microsoft Research**, Redmond, WA May 2025 – Aug 2025
Research Intern
Mentors: Adriana Szekeres, Suman Nath

- Architected the Beyond Screenshots WebAgent planner/memory stack so a single LLM pass emits a full sketch program for execution.
- Lifted WebArena task success to **~90%** (up from ~50%) by coupling state-machine memory with programmatic planning.

• **Georgia Institute of Technology**, Atlanta, GA

Aug 2024 – Present

Research Assistant

Advisor: Kexin Rong; Collaboration: VMware Systems Group

- Built a dynamic partitioning framework for RBAC-secure vector databases, reaching **13.5x** faster queries with **90%** less memory.
- Co-designed an RBAC-aware batching scheduler that aligns vector retrieval with LLM inference to boost RAG throughput.

• **InfiniFlow (Vector Database Startup)**

Mar 2024 – Apr 2024

Database Internals Engineer Intern

- Reworked timestamp persistence to avoid redundant ‘txn_map’ access, reducing critical-path latency.
- Streamlined bulk deletion cleanup, cutting file I/O and compaction cost for vector segments.

• **4Paradigm**

Feb 2024 – Apr 2024

Full Stack Software Engineer Intern (Part-Time)

- Tuned cache strategy for the AI assistant service, lowering response latency and server overhead.
- Delivered community feature workflows and async schedulers to automate data refresh.

• **Columbia University**, New York City, NY

Jul 2023 – Nov 2023

Research Assistant

Advisor: Eugene Wu

- Optimized sparse matrix evaluations for FADE, tightening end-to-end throughput for hypothetical queries.
- Deployed SIMD and multithreaded execution for provenance workloads, achieving near-linear 8x speedups.

• **Rutgers University**, New Brunswick, NJ

Jun 2023 – Sep 2023

Research Assistant

Advisor: Dong Deng

- Implemented baseline similarity search pipelines and orchestrated large experiment runs.
- Parallelized group-function analytics, improving CPU utilization across data partitions.

• **Nanyang Technological University / 4Paradigm**, Singapore

Jan 2023 – Jul 2023

Research Assistant

Advisors: Mian Lu, Shuhao Zhang

- Engineered a low-latency stream processing stack resilient to disorder and out-of-order arrivals.
- Applied Bayesian variational inference with transformer encoders to model complex event streams.

• **Meituan**, Beijing, China

Apr 2022 – Sep 2022

Backend Software Engineer Intern

- Shipped foundational backend for the Meituan short-video product, supporting new content formats.
- Built Kafka/Hive reporting pipelines to feed recommendation models and periodic refresh logic for low-bandwidth users.

Technologies

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

Mastery (First-Author + only student Author experience): AI agent cognitive architectures (memory, planners, LLM Reasoning), vector database, LLM Inference-RAG co-optimization, large-scale optimization modeling for data systems