Hongbin Zhong

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

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 2024–2027 (expected)*
Ph.D. in Computer Science
Advisor: Kexin Rong

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

Preprints And Publications

1. **Beyond Screenshots: A Dynamic State-Machine Memory and Global Programmatic Planner for Web Agents**
   Hongbin Zhong* with Microsoft Research collaborators
   Achieved about 90% accuracy on the WebArena benchmark, surpassing all existing SOTA methods; results to be released soon.

2. **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

3. **Fast Hypothetical Updates Evaluation**
   Haneen Mohammed*, Alexander Yao*, Charlie Summers*, Hongbin Zhong, Gromit Yeuk-Yin Chan, Subrata Mitra, Lampros Flokas, Eugene Wu
   ProvWeek at SIGMOD 2025

4. **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

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

6. **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

Microsoft Research Redmond, WA | May 2025 – Aug 2025
Research Intern
Mentors: Adriana Szekeres, Suman Nath

- Architected the **WebAgent planner/memory stack** so a single LLM pass emits a full sketch program for execution. Adopted a graph storage in WebAgent for efficient and accurate memory storage.

- 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: Microsoft Research Group; 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** and **LLM serving** layer that unifies vector retrieval with zero-risk docset grouping, cache-first scheduling, and proactive preheat via a lightweight Privilege Graph Approximation (PGA), improving TTFT and GPU utilization.

InfiniFlow (Vector Database Startup) Remote | 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(Top AI company in China, Series-D Unicorn) (Remote) | Feb 2024 – Apr 2024
Full Stack Software Engineer(Part-Time)

- Optimized backend and inference serving pipeline for the AI assistant application, tuning caching and scheduling to reduce end-to-end latency and improve GPU utilization.

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

- Accelerated **FaDE**, a provenance-driven what-if engine, via sparse-matrix and **SIMD-parallel** evaluation.
- Achieved **1M+ interventions/sec** throughput with 8× near-linear speedups on multi-core workloads.

4Paradigm(Top AI company in China, Series-D Unicorn) Beijing | Jan 2023 – Jul 2023
Research Intern
Mentors: Mian Lu, Shuhao Zhang

- Developed **PECJ**, a proactive error-compensation join for out-of-order streams using **variational inference**.
- Achieved up to **70% lower error** under equal latency and integrated it into a **multi-threaded SWJ engine**.

Meituan, Top 5 Chinese Internet Company Serving 600M+ Users Beijing, China | Apr 2022 – Sep 2022
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

Proficiency: Agent Reinforcement Learning (RL) Techniques, CUDA/GPU Programming, PyTorch/TensorFlow, Triton Inference Server, Distributed Training - NCCL/Ray

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