Hongbin Zhong

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

Education

Georgia Institute of Technology *Aug 2024–2028 (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

Work Experience

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

- Integrated a **Graph Memory system** into WebAgent for structured, low-latency state management and durable memory persistence within the agent's data layer.
- Re-architected the **planner/memory stack** as a queryable data store, enabling a single LLM pass to emit executable sketch programs over the stored state.
- Improved **WebArena** task success from ˜50% to ˜90% by coupling database-style state tracking with programmatic planning.

4Paradigm(Top AI company in China, Series-D Unicorn) (Remote) | Feb 2024 – Apr 2024
Full Stack AI 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.

4Paradigm (Top AI company in China, Series-D Unicorn) Beijing | Jan 2023 – Jul 2023
AI Database Research Intern

- Developed **PECJ**, a proactive error-compensation join algorithm in a **streaming database system**, leveraging **variational inference** for robust out-of-order stream processing.
- Achieved up to **70% lower query error** at equal latency and integrated the algorithm into the **OpenMLDB** project (`github.com/4paradigm/OpenMLDB`, ⭐ **1.7k stars**), a multi-threaded stream join engine within the database system.

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.

Open Source/Research Projects

InfiniFlow (Open-source Vector DB) ⭐ **4.2k stars** Remote | Mar 2024 – Apr 2024
Contributor, Storage & Indexing Optimization

- Reworked timestamp persistence layer to eliminate redundant `txn_map` access, reducing critical-path latency.
- Streamlined bulk-deletion and segment-compaction logic, cutting file I/O cost.

VMware Research - Gatech Atlanta, GA · **SIGMOD 2026** | Aug 2024 – Present
HoneyBee – Dynamic Partitioning for Role-secure Vector Databases

- Built a **dynamic partitioning framework** for RBAC-secure **vector databases**, achieving **13.5x faster queries** with **90%** less memory.
- Co-designed an **RBAC-aware batching and LLM serving** layer with cache-first scheduling and lightweight Privilege Graph Approximation.

Columbia University New York, NY · **VLDB 2025** | Jul 2023 – Nov 2023
FaDE – Provenance-driven What-if Engine (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 multi-core speedup, and contributed to an open-source **DuckDB** extension (see github.com/haneensa/fade).

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