

Qiyang He

☎ (765)337-0786 | @ qiyanghe1998@outlook.com | 🔗 LinkedIn | 📄 Google Scholar | 🏠 Homepage

EDUCATION

Purdue University

West Lafayette, IN, USA

M.Sc. in Computer Science; GPA: 3.91/4.00

Jan. 2021 – Dec. 2023

Southern University of Science and Technology (SUSTech)

Shenzhen, Guangdong, China

B.Eng. in Computer Science and Engineering; GPA: 3.86/4.00; Rank: 3/146

Sep. 2016 – Jul. 2020

EXPERIENCE

Microsoft AI – Agentic Coding RL Team, Member of Technical Staff

Apr. 2026 – Present

- **Built large-scale coding reinforcement learning environment (RLE) generation pipelines**, converting real-world GitHub pull requests, issues, commits, repository states, dependencies, and tests into reproducible executable environments, scaling toward **1M+ coding environments**.
- **Led trajectory-based RL environment generation from VS Code and Copilot CLI agent trajectories**, reconstructing task context, repository state, execution history, and evaluation signals from real agent interactions to create reliable, replayable and verifiable coding environments.
- **Built production infrastructure and validation for coding RLEs**—including dependency resolution, reproducibility checks, execution/test validation, and quality filtering—and contributed this work to the development and release of MAI-Thinking-1.

Snowflake – Metadata Team, Software Engineer

Feb. 2024 – Apr. 2026

- **Led metadata file clustering**, enabling compact storage of metadata based on customer-defined clustering keys. This significantly improved pruning efficiency and reduced query compilation time by **30%** on targeted workloads; the feature was adopted by major customers including **OpenAI**.
- **Built adaptive metadata file clustering**, a background optimization that intelligently balances performance and resource usage, cutting resource consumption by **40%** while maintaining similar pruning benefits.
- **Developed a hidden-table testing framework** to safely validate metadata/data-change features against live workloads, enabling faster, lower-risk releases.
- **Shipped new metadata APIs** that support approximate metadata requests, allowing the compiler to bypass precision-heavy lookups when safe; improved planning/compile latency and throughput on large **Iceberg** workloads.

Cockroach Labs – SQL Query Team, Research Intern

May 2023 – Aug. 2023

- **Developed a workload-aware index recommender** that analyzes real query traces to suggest optimal indexes. Delivered **5× throughput improvement** on analytical workloads (TPC-H) and **15% latency reduction** on transactional workloads (TPC-C) over manually tuned baselines.
- **Designed a cost-based index merging algorithm**, consolidating overlapping indexes to reduce storage footprint and ongoing maintenance overhead.

Pinterest – Distributed Storage System Team, Research Intern

May 2022 – Aug. 2022

- **Built a space-efficient incremental backup** using Multi-Paxos that stored only delta updates instead of full files, achieving a **10× speedup** in backup and restore performance.

Purdue University – Programming language Group, Research Assistant

Jan. 2021 – May 2023

- **Developed a C code generation backend for SparkSQL** using (LMS)-based generative programming to support incremental view maintenance.
- Identified inefficiencies in DBToaster and **designed tree- and hash-based index structures**, improving correlated nested aggregate query performance by up to **1000×**

SUSTech – Database Group, Research Assistant

Dec. 2019 – Jul. 2020

- **Built a high-dimensional vector search engine** using product/residual vector quantization; implemented custom CUDA kernels for batched inner-product computation and on-GPU sorting to generate top-k results, delivering low-latency, high-throughput retrieval while reducing memory footprint vs. exact search.

PUBLICATIONS & PROFESSIONAL SERVICE

SIGMOD 2022

Efficient Incrementalization of Correlated Nested Aggregate Queries using Relative Partial Aggregate Indexes (RPAI)

OOPSLA 2021

Reachability types: tracking aliasing and separation in higher-order functional programs

ACM CSUR 2025

A Survey of Learned Indexes for the Multi-dimensional Space

Shadow Reviewer, Proceedings of the VLDB Endowment (PVLDB), 2026

Student Volunteer, PLDI 2021

TECHNICAL EXPERTISE

Research Areas: Agentic coding systems; reinforcement learning and post-training infrastructure; scalable AI/data infrastructure; database systems; distributed systems; metadata management; query optimization

Core Expertise: Coding RL environment generation; agent trajectory processing; large-scale AI training/post-training systems; distributed systems; indexing; performance optimization

Programming Languages: C++, Python, Go, Java, Scala, Rust, C, SQL, Bash

Systems & Frameworks: Linux; RocksDB; SparkSQL; Apache Iceberg; AWS S3

Awards: ACM-ICPC Regional (Gold/Silver); China National Olympiad in Informatics (Bronze)