

Shuhao Zhang

Professor, School of Computer Science and Technology, HUST
shuhao_zhang@hust.edu.cn ORCID: 0000-0002-9927-6925 home.shuhao.sage.org.ai

Profile

Shuhao Zhang is a Professor in the School of Computer Science and Technology at Huazhong University of Science and Technology. He leads a task under a major national science and technology project of the Ministry of Science and Technology of China and is a recipient of the National Natural Science Foundation of China Excellent Young Scientists Fund (Overseas).

His research focuses on state-management-driven optimization for inference systems, spanning parallel and distributed systems, LLM inference infrastructure, inference serving systems, and memory-agent middleware. Current work emphasizes state-aware scheduling, hardware-aware memory management, KV and intermediate-state reuse, serving stability, and the impact of these mechanisms on throughput, P99 latency, and service reliability.

Appointments

Huazhong University of Science and Technology School of Computer Science and Technology.	Professor
Nanyang Technological University Before joining HUST.	Assistant Professor
Technische Universität Berlin Postdoctoral research on data systems and distributed stream processing.	Postdoctoral Researcher

Education

National University of Singapore Research area: database systems, stream processing, parallel and distributed systems.	Ph.D., Computer Science, 2019
Nanyang Technological University	B.Eng., Computer Engineering, 2014

Research Themes

- State-aware orchestration and resource governance for LLM serving systems.
- Hardware-aware memory management for KV cache, prefix state, vector memory, and multi-level memory systems.
- State-reuse inference for long-context workloads, MoE serving, and domestic heterogeneous accelerators.
- Benchmark-driven systems methodology, reproducible performance claims, and open engineering workflows.

Major Projects and Funding

- Task Lead, major national science and technology project of the Ministry of Science and Technology of China.
- Principal Investigator, National Natural Science Foundation of China Excellent Young Scientists Fund (Overseas).
- Principal Investigator and collaborator on previous research grants from Singapore MOE, Singapore NRF, Temasek Laboratories, and university-level research programs.

Representative Systems

- **vLLM-HUST**: HUST-maintained LLM inference engine stack and benchmark ecosystem for domestic heterogeneous serving practice.
- **SAGE**: A dataflow-native framework for modular, controllable, and transparent LLM-augmented reasoning.
- **Neuromem**: Middleware and benchmark line for external memory and memory-agent systems.
- **MorphStream / TStream / CStream**: Stream processing systems for multicore, stateful, and hardware-conscious execution.

Selected Recent Publications

- **BidKV: Utility-Guided Preemption Scheduling for KV-Pressure LLM Serving**. SC 2026. Corresponding Author.
- **SAGE: A Dataflow-Native Framework for Modular, Controllable, and Transparent LLM-Augmented Reasoning**. ICML 2026. Corresponding Author.
- **Neuromem: A Granular Decomposition of the Streaming Lifecycle in External Memory for LLMs**. ICML 2026. Corresponding Author.
- **CANDOR-Bench: Benchmarking In-Memory Continuous ANNS under Dynamic Open-World Streams**. SIGMOD 2026. Corresponding Author.
- **MorphStream: Adaptive Scheduling for Scalable Transactional Stream Processing on Multicores**. SIGMOD 2023 / TKDE 2025 line of work. Corresponding Author.
- **Towards Concurrent Stateful Stream Processing on Multicore Processors**. ICDE 2020. First Author.
- **BriskStream: Scaling Data Stream Processing on Shared-Memory Multicore Architectures**. SIGMOD 2019. First Author.

Honors and Service

- HUST Excellent Young Scholar; Huawei Donghu Young Scholar.
- VLDB Review Board Member; SC Program Committee Member; ICDE/TKDE Poster Track Chair.
- Program committee service for KDD, CIKM, DASFAA, SDM, and related venues.

Teaching and Open Materials

- Large Language Model Inference Systems and Practice, a practice-oriented graduate course connected to vLLM-HUST development.
- Graduate Paper Writing, public lecture materials for research writing and systems-paper methodology.