Zebin Yao

M.S. Student, Nankai-Baidu Joint Laboratory Group, Nankai University
zebin_yao@126.com — +86 18107096525 — https://github.com/YReddice

RESEARCH INTERESTS

Approximate Nearest Neighbor Search • Vector Database • Retrieval-Augmented Generation
My primary research interests lie in the efficiency and optimization of ANNS systems and their applications in large language models, such as retrieval-augmented generation and search agents.

EDUCATION

M.S. in Computer Science and Technology **Sep 2023 — Present**
Nankai University (NKU)
Supervisor: Prof. Gang Wang

B.S. in Computer Science and Technology **Sep 2019 — Jun 2023**
Nankai University (NKU)

PUBLICATIONS AND MANUSCRIPTS

ALGAS: A Low-latency GPU-Accelerated Approximate Nearest Neighbor Search System.
IEEE International Parallel & Distributed Processing Symposium (**IPDPS**), 2025. (Third Author)

Demystifying and Enhancing the Efficiency of Large Language Model Based Search Agents. arXiv
Annual Conference on Neural Information Processing Systems (**NeurIPS**), 2025. (Submitted, Co-First Author)

Dynamic Detect and Fix Hardness for Efficient Approximate Nearest Neighbor Search.
Proceedings of the ACM SIGMOD International Conference on Management of Data (**SIGMOD**), 2026.
(Under Review, Third Author)

Embedding, Retrieval, and Generation: A Comprehensive Survey of Efficient Retrieval-Augmented LLMs.
(Manuscript)

RESEARCH AND PROJECT EXPERIENCE

Alibaba Group **Nov 2024 — Jan 2025**
Research Intern Beijing, China

- Addressed the long-tail recall issue in multimodal image retrieval by optimizing the ANN graph index structure, significantly improving retrieval accuracy.
- Implemented an intra-query parallel algorithm leveraging multi-core CPU architecture to accelerate single-query performance under large-scale data conditions.

Infinity GitHub
Open Source Contributor

- Infinity is a database designed for LLM applications, offering hybrid search across dense vectors, sparse vectors, tensors, and full-text data.
- Designed and implemented a disk-based vector index architecture that eliminates split-table complexity and enables efficient end-to-end indexing and querying; successfully merged into the core repository.

TEACHER ASSISTANT

- **COSC0025** Parallel Programming Spring 2025, NKU
- **COSC0017** Compiler Principles Autumn 2022, NKU

SELECTED AWARDS

- Gongneng Scholarship, Nankai University Sep 2024
- Gongneng Scholarship, Nankai University Sep 2023
- Huawei Smart Pedestal Scholarship (sponsored by Huawei) Oct 2022
- Innovation and Academic Excellence Scholarship, Nankai University Oct 2021