Tianshu Wang

contact@wang.tianshu.me | github.com/tshu-w | Beijing, China

Summary

PhD candidate at ISCAS, graduating in Sep. 2025. Research interests include LLM, data integration, and RAG. Published **6 papers at CCF A/B conferences (3 oral presentations)** including ACL, IJCAI, and EDBT. Recipient of **EDBT 25 Best Paper Runner-Up Award** and **ACL 24 Area Chair's Award**. Won **5 ICPC/CCPC Regional Contest gold medals** and **1 Invitational Contest third-place**. Experienced in both academic research and engineering practice through internships at SenseTime Research and ByteDance AI Lab.

Education

Institute of Software, Chinese Academy of Sciences
Hangzhou Institute for Advanced Study, UCAS — Sep 2021 – Sep 2025
PhD of Computer Science and Technology, Supervisor: Prof. Le Sun — Beijing, China
- Research Interests: Data Integration, Large Language Models, Retrieval-Augmented Generation (RAG)
- GPA: 3.73/4, Honors: **CAS President Award**, Outstanding Student, First-Class Academic Scholarship

Institute of Software, Chinese Academy of Sciences — Sep 2019 – Sep 2021
Master of Computer Software and Theory, Supervisor: Prof. Xianpei Han — Beijing, China
- Research Interests: Entity Resolution, Structured Data Retrieval

Beijing Jiaotong University — Sep 2015 – Jun 2019
Bachelor of Computer Science and Technology — Beijing, China
- GPA: 91.2/100 (Top 10%), English Level: CET-6 513

Research Experience

DBCopilot: Natural Language Querying over Massive Databases via Schema Routing — EDBT 2025
(🏆 ***1st Runner-Up for Best Research Paper Award***) tshu-w/DBCopilot ☆ 100+ stars
- First work to apply generative retrieval to LLM graph RAG, won **Best Paper Runner-Up Award**
- Propose DBCopilot framework with small-large model collaboration and relation-aware joint retrieval, extending NL2SQL to massive databases
- Successfully integrated into Alibaba Cloud Bailian product suite
- **First author** published at **CCF-B** conference EDBT 2025, **Oral Presentation**

Spiral of Silence: How is Large Language Model Killing Information Retrieval?—A Case Study on Open Domain Question Answering (🏆 *Area Chair's Award*) — ACL 2024
- Build simulation framework to analyze catastrophic impacts of LLM-generated text on information retrieval and QA, and explore mitigation strategies
- **Second student author** published at **CCF-A** conference ACL 2024, won **Area Chair Award**

Bridging the Gap between Reality and Ideality of Entity Matching: A Revisiting and Benchmark Re-Construction — IJCAI 2022
- Analyze erroneous assumptions implied in entity matching benchmarks and construct first open entity, imbalanced label, and multimodal benchmarks to mitigate evaluation bias
- **First author** published at **CCF-A** conference IJCAI 2022, **Oral Presentation**

Match, Compare, or Select? An Investigation of Large Language Models for Entity Matching — COLING 2025
- First work to introduce L2R pointwise, pairwise, and listwise concepts into entity matching
- Investigate different LLM entity matching strategies and design more effective composite framework
- **First author** published at **CCF-B** conference COLING 2025, **Oral Presentation**

ARise: Towards Knowledge-Augmented Reasoning via Risk-Adaptive Search — ACL 2025
- Propose knowledge-augmented reasoning framework based on risk-adaptive Monte Carlo tree search, outperforming existing methods by 20+% on open-domain complex reasoning tasks
- **Co-first author** published at **CCF-A** conference ACL 2025

Work Experience

LLM Algorithm Intern | *SenseTime Research* Jun 2024 – Apr 2025

- Led research on LLM knowledge-augmented reasoning, outperforming on open-domain complex reasoning tasks
- Participated in "Data Pioneer" product algorithm research, implementation, and delivery, supporting WAIC showcase and Ningbo Bank Phase-1 project
- Received SenseTime 2024 Outstanding Team Award

Algorithm Intern | *ByteDance AI Lab* Feb 2019 – Aug 2019

Assisted in development and optimization of "Precision Answering" project for Toutiao Search

- Used Spark & Hive to analyze daily Top 10 precision QA coverage statistics, supporting forward mining
- Filtered 700M+ CQA data using MapReduce based on rules, answer types, question intent, and answer quality
- Optimized answer summary generation module efficiency by over 50% using dynamic programming

Honors & Awards

International Collegiate Programming Contest (ICPC) Asia Regional Contest 2016 – 2018

- **Gold Medal × 2**: 42nd Beijing Site, 43rd Jiaozuo Site
- **Silver Medal × 3**: 42nd Urumqi Site, 42nd, 43rd East Continent Final Contest
- **Bronze Medal × 1**: 41st East Continent Final Contest

China Collegiate Programming Contest (CCPC) 2017 – 2018

- **Third Place × 1**: 4th National Invitational Contest Xiangtan Site
- **Gold Medal × 3**: 3rd Qinhuangdao Site, 4th Guilin Site, Jilin Site

Skills

- *Technical Expertise*: Familiar with LLMs, Agent, IR, NLP, software development, data mining, and DevOps
- *Programming Languages*: Proficient in Python, C++, Bash; familiar with SQL, Lua, TypeScript, Java
- *System Tools*: Proficient in macOS/Linux, Git, PyTorch, Docker, Ray, Spark, Elasticsearch, Emacs, LaTeX, etc.
- *General Abilities*: Strong literature reading and rapid learning skills, excellent analysis and problem-solving abilities, good communication and collaboration skills

Additional Activities

- *Open Source*: GitHub 400+ Stars, contributed to PyTorch Lightning (12 commits) and 44 other repositories
- *Academic Service*: ACL ARR - Feb./May 2025, KDD 2024/2025 Research Track, NeurIPS 2024 TRL Reviewer
- *Technical Community*: Stack Overflow Reputation 758, Emacs Stack Exchange Reputation 1889
- *Teaching Assistant*: "Intelligent Information Retrieval" at ISCAS, "C Language Programming" at Beijing Jiaotong University, ACM-ICPC Team Summer Training