

# Tianshu Wang

✉ [contact@wang.tianshu.me](mailto:contact@wang.tianshu.me) | [github.com/tshu-w](https://github.com/tshu-w) | 📍 Beijing, China

[Online Version](#)

## Education

---

### Institute of Software, Chinese Academy of Sciences

Sep 2019 - Jun 2025

PhD of Computer Science and Technology (transfer from Masters to PhD)

Beijing, China

- Research Intrest: LLM for Data Intelligence (Entity Resolution, Text2SQL, and Structured Text Retrieval) .

- GPA: 3.73/4 (2rd), Honors: Outstanding Student, First-Class Academic Scholarship

### Beijing Jiaotong University

Sep 2015 - Jun 2019

Bachelor of Computer Science and Technology

Beijing, China

- GPA: 91.2/100 (Top 10%), CET-6: 513

## Research Experience

---

Tianshu Wang *et al.*, “Bridging the Gap between Reality and Ideality of Entity Matching: A Revisting and Benchmark ReConstrction”, in *Proc. of IJCAI 2022*, pp. 3978–3984. doi: [10.24963/ijcai.2022/552](https://doi.org/10.24963/ijcai.2022/552).

- Analyze the erroneous assumptions implied in entity matching benchmarks
- Construct the first open entity, imbalanced label, and multimodal benchmarks to mitigate evaluation bias

Tianshu Wang *et al.*, “DBCopilot: Scaling Natural Language Querying to Massive Databases”, *CoRR*, 2023, doi: [10.48550/arXiv.2312.03463](https://doi.org/10.48550/arXiv.2312.03463). Submit to KDD 2024

- Present a LLM-Copilot collaboration framework for scaling natural language querying to massive databases
- Propose a relation-aware, generative joint retrieval approach for schema routing

Tianshu Wang *et al.*, “Match, Compare, or Select? An Investigation of Large Language Models for Entity Matching”, *CoRR*, 2024, doi: [10.48550/arXiv.2405.16884](https://doi.org/10.48550/arXiv.2405.16884). Submit to ACL ARR 2024 February

- Investigate different (proposed) strategies for entity matching with LLMs
- Propose a compositional framework for more effective and efficient than single strategies

Tianshu Wang *et al.*, “Towards Universal Dense Blocking for Entity Resolution”, *CoRR*, 2024, doi: [10.48550/arXiv.2404.14831](https://doi.org/10.48550/arXiv.2404.14831).

- Pretrain a universal structured record representation model for dense blocking without domain-specific training

## Work Experience

---

### AI Lab Intern | ByteDance

Feb 2019 - Aug 2019

Contributed to “Precision Answering” for [Toutiao Search](#) by developing and optimizing key components

- Mined top frequent unanswered queries daily using *Spark & Hive* to support forward mining
- Filtered CQA data based on rules, question intent, and answer quality/types using *MapReduce*
- Optimized the answer extraction module’s efficiency by over 50% through *dynamic programming*

## Honors & Awards

---

### International Collegiate Programming Contest (ICPC) Asia Regional Contest

2016 - 2018

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

### China Collegiate Programming Contest (CCPC)

2017 - 2018

- Third place × 1: The 4th National Invitational Contest Hunan Site
- Gold Medal × 3: The 3rd Qinhuangdao Site, the 4th Guilin Site, and Jilin Site

## Skills

---

- Research Interests: LLM & Agent, Entity Resolution, Text-to-SQL, Information Retrieval
- Expertise: Software development, Data Structures & Algorithms, DevOps, Data Analytics
- Programming Languages: Python, C++, C, Bash | SQL, Java, Lua, TypeScript
- Software: macOS/Linux, Emacs, Git, Docker, Ray, Spark,  $\LaTeX$
- Open Source Contributions: contributed to [Pytorch Lightning](#) (10 commits) and [other 45](#) repositories.

Last Updated on Jun 29, 2024