

# Danrui Qi

📍 Simon Fraser University, Burnaby, BC, Canada    ✉️ dq@sfu.ca    ☎️ (+1)236-966-8109

🔗 qidanrui.github.io    in LinkedIn    🌐 qidanrui

## Research Interests

My research interest lies in developing intelligent systems based on large language models (LLMs) and multi-agent frameworks. Specifically, I focus on building agentic architectures for reasoning, planning, and collaboration among LLMs, as well as exploring their applications in data-centric AI, database systems, and automation of complex real-world tasks.

## Education

<b>Ph.D. Simon Fraser University</b> , School of Computer Science	Sept. 2020 – Present
• Ph.D. Candidate, supervised by <i>Prof. Jiannan Wang</i> and worked closely with <i>Prof. Zhengjie Miao</i>	
<b>M.S.E. Tsinghua University</b> , School of Software	Aug. 2017 – Jul. 2020
• Master of Software Engineering, supervised by <i>Prof. Shaoxu Song</i>	
<b>B.E. Tsinghua University</b> , School of Software	Aug. 2013 – Jul. 2017
• Bachelor of Engineering of Software Engineering	

## Experience

<b>Microsoft Research</b> , Research Intern (Database System Group)	Redmond, WA
<i>Supervised by Dr. Yeye He</i>	May 2024 – Aug. 2024
• Fine-tuned a GPT-4 filter model to include only question-relevant table content in prompts, enhancing reasoning accuracy in NL2SQL and TableQA tasks	
• Used GPT-4 via the Azure API to generate synthetic training data, achieving a 4% average accuracy gain on two TableQA datasets and 3.9% on five NL2SQL datasets, compared to unfiltered and vanilla GPT-4 filtering baselines	
• Reduced input prompt length by up to 2× across all datasets, while improving accuracy by as much as 9% when integrated as a preprocessing layer for existing NL2SQL and TableQA methods	
• Authored paper <i>TABLE-REDUCER: General-Purpose Models for Table Context Reduction in Diverse Table Task</i>	

## Publications

### Published Papers

<b>1.</b> Arijit Khan1, Yuyu Luo, M. Tamer Ozsu, <b>Danrui Qi</b> , Jiannan Wang. Second International Workshop on LLM+Vector Data: Agentic RAG Edition	ICDE 2026
<b>2.</b> <b>Danrui Qi</b> , Jiannan Wang. CleanAgent: Automating Data Standardization with LLM-based Agents.	DataAI@VLDB 2025
<b>3.</b> <b>Danrui Qi</b> , Weiling Zheng, Jiannan Wang. FeatAug: Automatic Feature Augmentation From One-to-Many Relationship Tables.	ICDE 2024
<b>4.</b> <b>Danrui Qi</b> , Jinglin Peng, Yongjun He, Jiannan Wang. Auto-FP: An Experimental Study of Automated Feature Preprocessing for Tabular Data.	EDBT 2024
<b>5.</b> Siqiao Xue, <b>Danrui Qi</b> , Caigao Jiang, Wenhui Shi, Fangyin Cheng, Keting Chen, Hongjun Yang et al. Demonstration of DB-GPT: Next Generation Data Interaction System Empowered by Large Language Models.	Demonstration@VLDB 2024

- |   |                          |
|---|--------------------------|
| 6. Jinglin Peng, Weiyuan Wu, Jing Nathan Yan, <b>Danrui Qi</b> , Jeffrey M. Rzeszotarski, Jian-nan Wang. User Interfaces for Exploratory Data Analysis: A Survey of Open-Source and Commercial Tools. | IEEE Data Eng. Bull 2022 |
| 7. <b>Danrui Qi</b> . On concise explanations of non-answers over big data.   | SRC@SIGMOD 2017          |

## Preprints

- |   |            |
|---|------------|
| 1. Aaron Xuxiang Tian, Ruofan Zhang, Jiayao Tang, Young Min Cho, Xueqian Li, Qiang Yi, Ji Wang, Zhunping Zhang, <b>Danrui Qi</b> , Sharath Chandra Guntuku, Lyle Ungar, Tianyu Shi and Chi Wang. Beyond the Strongest LLM: Multi-Turn Multi-Agent Orchestration vs. Single LLMs on Benchmarks.        | arXiv 2025 |
| 2. Fan Zhou, Siqiao Xue, <b>Danrui Qi</b> , Wenhui Shi, Wang Zhao, Ganglin Wei, Hongyang Zhang et al. DB-GPT-Hub: Towards Open Benchmarking Text-to-SQL Empowered by Large Language Models.   | arXiv 2024 |
| 3. Siqiao Xue, Caigao Jiang, Wenhui Shi, Fangyin Cheng, Keting Chen, Hongjun Yang, Zhiping Zhang, Jianshan He, Hongyang Zhang, Ganglin Wei, Wang Zhao, Fan Zhou, <b>Danrui Qi</b> , Hong Yi, Shaodong Liu, Faqiang Chen. DB-GPT: Empowering Database Interactions with Private Large Language Models. | arXiv 2023 |

## Research Experience

---

- |  |                        |
|--|------------------------|
| <b>Agentic Data Preparation</b> <ul style="list-style-type: none"> <li>Reframed data preparation as a code-generation problem, leveraging a multi-agent framework to translate natural language (NL) user requests into executable data preparation workflows</li> <li>Designed CleanAgent with a task-specific clean API abstraction, enabling LLMs to reason and generate table standardization code using column type annotations and API calls</li> </ul>  | Feb. 2024 - Present    |
| <b>Automatic Feature Augmentation</b> <ul style="list-style-type: none"> <li>Defined and formalized the automatic feature augmentation problem as a novel SQL query generation challenge</li> <li>Developed FeatAug, leveraging Bayesian Optimization to address the problem, resulting in up to 10.74% AUC improvement on various datasets</li> <li>Authored a paper detailing these findings, published in top-tier database conference ICDE 2024</li> </ul>   | Sept. 2022 - Jan. 2024 |
| <b>Automatic Feature Preprocessing</b> <ul style="list-style-type: none"> <li>Defined the automatic feature preprocessing problem, illustrating its practical challenges and potential solutions</li> <li>Translated the problem into Hyperparameter Optimization and Neural Architecture Search contexts, evaluating 15 algorithms and identifying the superiority of PBT</li> <li>Showed 0.7%-3.5% average accuracy improvement of Auto-FP compared to existing AutoML's feature processing modules</li> <li>Documented these results and published a paper in top-tier database conference EDBT 2024</li> </ul> | Sept. 2020 - Nov. 2022 |

## Open-Source Projects

---

- |  |                     |
|--|---------------------|
| <b>Dataprep (2.2K Stars on Github) <a href="#">↗</a></b> <ul style="list-style-type: none"> <li>Designed the architecture of Dataprep.Clean, a Python library that simplifies and accelerates column type-based data cleaning, featuring 140+ utility functions and comprehensive documentation</li> </ul> | Sep. 2020 - Present |
|--|---------------------|

- Developed Clean GUI, a code-free data cleaning tool in Jupyter, which streamlined user workflows and boosted performance by 20% through integration with Dask

**DB-GPT (17.3K Stars on Github)** [🔗](#)

Sep. 2023 - Present

- Engineered the workflow and API to harness LLMs for data science applications
- Implemented the multi-agent framework to automate database QA with fully NL-based interactions

**CleanAgent (69 Stars on Github)** [🔗](#)

Feb. 2024 - Present

- Extended Dataprep.Clean with a multi-agent framework based on popular LLMs, automating data standardization processes through natural language interactions
- Developed a web-based GUI that enabled efficient and high-quality data cleaning, supported by a real-time backend processing server, ensuring a seamless user experience

**MassGen (569 Stars on Github)** [🔗](#)

Jul. 2025 - Present

*Goal: Multi-Agent Scaling System for GenAI - Leveraging collaborative agents to solve complex tasks*

- Developed orchestration mechanisms for parallel multi-agent collaboration, enabling simultaneous task processing across diverse LLMs with real-time synchronization
- Architected cross-model synergy supporting 15+ providers (OpenAI, Anthropic, Google, xAI) with MCP integration and custom tool support

**Honors & Awards**

---

<b>Westak International Sales Inc. Scholarship</b>	May 2024
<b>PhD Research Scholarship at Simon Fraser University</b>	Sep. 2023, Jan. 2024
<b>Graduate Dean's Entrance (GDES) at Simon Fraser University</b>	Sep. 2020
<b>Outstanding Graduate Thesis Award at Tsinghua University</b>	Jun. 2017
<b>National Inspirational Scholarship at Tsinghua University</b>	Oct. 2016

**Services**

---

**Workshop Co-Chair:** LLM + Vector Data @ ICDE 2026  
**Shadow Program Committee:** VLDB 2026  
**Program Committee Member:** WWW 2026, SurveyTrack@IJCAI 2025, CIKM 2025, CIKM 2024, ICDE 2022  
**Reviewer:** TKDE, ICLR 2026, KDD 2025, IJCAI 2025, WACV 2025, ICLR 2025, DASFAA 2024, CIKM 2024, CIKM 2023, ICDE 2022

**Skills**

---

**Languages:** C, C++, Java, Python, Javascript, HTML  
**Frameworks:** Scikit-Learn, Pandas, Dask, Spark, Cassandra, Hadoop