Danrui Qi

Simon Fraser University, Burnaby, BC, Canada dqi@sfu.ca (+1)236-966-8109
qidanrui.github.io LinkedIn qidanrui

Research Interests

My research focuses on agentic AI systems that integrate reasoning, planning, and collaboration to automate complex, real-world, data-centric tasks. I build agent-powered data systems and a self-improving multi-agent scaling system to redefine how data workflows and data systems operate at scale. Earlier, my work centered on automating tabular data preparation using Bayesian Optimization and AutoML, achieving notable gains in downstream machine learning performance.

Education

Ph.D. **Simon Fraser University**, School of Computer Science — Sept. 2020 – Present
- Ph.D. Candidate, supervised by *Prof. Jiannan Wang* and worked closely with *Prof. Zhengjie Miao*

M.S.E. **Tsinghua University**, School of Software — Aug. 2017 – Jul. 2020
- Master of Software Engineering, supervised by *Prof. Shaoxu Song*

B.E. **Tsinghua University**, School of Software — Aug. 2013 – Jul. 2017
- Bachelor of Engineering of Software Engineering

Experience

Microsoft Research, Research Intern (Database System Group) — Redmond, WA
May 2024 – Aug. 2024

Supervised by Dr. Yeye He

- 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 *TABLE-REDUCER: General-Purpose Models for Table Context Reduction in Diverse Table Task*

Publications

Published Papers

1. Arijit Khan1, Yuyu Luo, M. Tamer Ozsu, **Danrui Qi**, Jiannan Wang. Second International Workshop on LLM+Vector Data: Agentic RAG Edition — ICDE 2026

2. Danrui Qi, Jiannan Wang. CleanAgent: Automating Data Standardization with LLM-based Agents. — DataAI@VLDB 2025

3. Danrui Qi, Weiling Zheng, Jiannan Wang. FeatAug: Automatic Feature Augmentation From One-to-Many Relationship Tables. — ICDE 2024

4. Danrui Qi, Jinglin Peng, Yongjun He, Jiannan Wang. Auto-FP: An Experimental Study of Automated Feature Preprocessing for Tabular Data. — EDBT 2024

5. Siqiao Xue, **Danrui Qi**, 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. — Demostration@VLDB 2024

6. Jinglin Peng, Weiyuan Wu, Jing Nathan Yan, **Danrui Qi**, Jeffrey M. Rzeszotarski, Jiannan Wang. User Interfaces for Exploratory Data Analysis: A Survey of Open-Source and Commercial Tools. IEEE Data Eng. Bull 2022

7. Danrui Qi. On concise explanations of non-answers over big data. SRC@SIGMOD 2017

Preprints & Draft

1. Aaron Xuxiang Tian, Ruofan Zhang, Jiayao Tang, Young Min Cho, Xueqian Li, Qiang Yi, Ji Wang, Zhunping Zhang, **Danrui Qi**, 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. Danrui Qi with Microsoft Research collaborators. TABLE-REDUCER: General-Purpose Models for Table Context Reduction in Diverse Table Task. draft 2025

3. Fan Zhou, Siqiao Xue, **Danrui Qi**, 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

4. Siqiao Xue, Caigao Jiang, Wenhui Shi, Fangyin Cheng, Keting Chen, Hongjun Yang, Zhiping Zhang, Jianshan He, Hongyang Zhang, Ganglin Wei, Wang Zhao, Fan Zhou, **Danrui Qi**, Hong Yi, Shaodong Liu, Faqiang Chen. DB-GPT: Empowering Database Interactions with Private Large Language Models. arXiv 2023

Research Experience

Agentic Data Preparation Feb. 2024 - Present

- 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
- 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

Automatic Feature Augmentation Sept. 2022 - Jan. 2024

- Defined and formalized the automatic feature augmentation problem as a novel SQL query generation challenge
- Developed FeatAug, leveraging Bayesian Optimization to address the problem, resulting in up to 10.74% AUC improvement on various datasets
- Authored a paper detailing these findings, published in top-tier database conference ICDE 2024

Automatic Feature Preprocessing Sept. 2020 - Nov. 2022

- Defined the automatic feature preprocessing problem, illustrating its practical challenges and potential solutions
- Translated the problem into Hyperparameter Optimization and Neural Architecture Search contexts, evaluating 15 algorithms and identifying the superiority of PBT
- Showed 0.7%-3.5% average accuracy improvement of Auto-FP compared to existing AutoML's feature processing modules
- Documented these results and published a paper in top-tier database conference EDBT 2024

Open-Source Projects

Dataprep (2.2K Stars on Github) *Role:* ***Founder Member*** *and* ***Maintainer*** Sep. 2020 - Present

- 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

- 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 (17.5K Stars on Github) *Role:* ***Main Contributor*** 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 (70 Stars on Github) *Role:* ***Founder*** *and* ***Maintainer*** 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 (585 Stars on Github) *Role:* ***Founder Member*** *and* ***Maintainer*** 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

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