

WENQI JIANG

Incoming Assistant Professor
National University of Singapore
Last Update: January 2026

STF G222
Stampfenbachstrasse 114
8092 Zurich, Switzerland

+86 135 2016 8996
wenqi.jiang@inf.ethz.ch
<https://wenqijiang.github.io/>

EDUCATION

ETH Zurich <i>PhD in Computer Science</i> <i>Advisors: Gustavo Alonso and Torsten Hoefer</i> <i>Committee: Gustavo Alonso, Torsten Hoefer, Ana Klimovic, and Christos Kozyrakis</i>	2021~2025
Columbia University <i>Master in Electrical Engineering</i> <i>The Master's Award of Excellence (top 5%)</i>	2018~2020
Huazhong University of Science and Technology <i>Bachelor in Automation</i> <i>Outstanding Graduate</i>	2014~2018

PROFESSIONAL APPOINTMENTS

NVIDIA Research <i>Postdoctoral Researcher</i> <i>Mentor: Christos Kozyrakis</i>	Jan. 2026 ~ July 2026 Santa Clara, USA
Google Cloud <i>Student Researcher</i> <i>Mentors: Vidushi Dadu, Suvinay Subramanian, and Amir Yazdanbakhsh</i>	July 2024 ~ Dec. 2024 Sunnyvale, USA
Amazon Web Services <i>Applied Scientist Intern</i> <i>Mentors: Shuai Zhang and Boran Han</i>	Oct. 2023 ~ Jan. 2024 Santa Clara, USA
Alibaba Cloud <i>Software Engineering Intern</i>	Sep. 2019 ~ Dec. 2019 Shenzhen, China

AWARDS AND HONORS

VLDB'25 Best Paper Award (Scalable Data Science Track)	2025
ML and Systems Rising Stars Award	2024
AMD HACC Outstanding Researcher Award	2023

The Master's Award of Excellence (top 5%), Columbia University	2021
Outstanding Graduate, HUST	2018
Scholarship for Excellent Academic Performance, HUST	2015

CONFERENCE PAPERS

- [1] **Wenqi Jiang**, Suvinay Subramanian, Cat Graves, Gustavo Alonso, Amir Yazdanbakhsh, and Vidushi Dadu, "RAGO: Systematic Performance Optimization for Retrieval-Augmented Generation Serving." *Proceedings of 52nd Annual International Symposium on Computer Architecture (ISCA'25)*
- [2] **Wenqi Jiang**, Marco Zeller, Roger Waleffe, Torsten Hoefer, and Gustavo Alonso, "Chameleon: a Heterogeneous and Disaggregated Accelerator System for Retrieval-Augmented Language Models." *Proceedings of the VLDB Endowment (VLDB'25)*
 **Best Scalable Data Science Paper Award**
- [3] **Wenqi Jiang**, Hang Hu, Torsten Hoefer, and Gustavo Alonso, "Fast Graph Vector Search via Hardware Acceleration and Delayed-Synchronization Traversal." *Proceedings of the VLDB Endowment (VLDB'25)*
- [4] **Wenqi Jiang**, Oleh-Yevhen Khavrona, Martin Parvanov, and Gustavo Alonso, "SwiftSpatial: Spatial Joins on Modern Hardware." *International Conference on Management of Data (SIGMOD'25)*
- [5] **Wenqi Jiang**, Shuai Zhang, Boran Han, Jie Wang, Bernie Wang, and Tim Kraska, "PipeRAG: Fast Retrieval-Augmented Generation via Adaptive Pipeline Parallelism." *Proceedings of the 31th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'25)*
- [6] Qi Chen, Xiubo Geng, Corby Rosset, Carolyn Buractaon, Jingwen Lu, Tao Shen, Kun Zhou, Chenyan Xiong, Yeyun Gong, Paul Bennett, Nick Craswell, Xing Xie, Fan Yang, Bryan Tower, Nikhil Rao, Anlei Dong, **Wenqi Jiang**, Zheng Liu, Mingqin Li, Chuanjie Liu, Zengzhong Li, Rangan Majumder, Jennifer Neville, Andy Oakley, Knut Magne Risvik, Harsha Vardhan Simhadri, Manik Varma, Yujing Wang, Linjun Yang, Mao Yang, and Ce Zhang, "MS MARCO Web Search: A Large-scale Information-rich Web Dataset with Millions of Real Click Labels." *International World Wide Web Conference (WWW'24)*
- [7] Shuai Zhang and **Wenqi Jiang**, "Data-Informed Geometric Space Selection." *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS'23)*
- [8] **Wenqi Jiang**, Shigang Li, Yu Zhu, Johannes de Fine Licht, Zhenhao He, Runbin Shi, Cedric Renggli, Shuai Zhang, Theodoros Rekatsinas, Torsten Hoefer, and Gustavo Alonso, "Co-design Hardware and Algorithm for Vector Search." *The International Conference for High Performance Computing, Networking, Storage and Analysis (SC'23)*
- [9] **Wenqi Jiang***, Zhenhao He*, Shuai Zhang, Kai Zeng, Liang Feng, Jiansong Zhang, Tongxuan Liu, Yong Li, Jingren Zhou, Ce Zhang, and Gustavo Alonso, "FleetRec: Large-Scale Recommendation Inference on Hybrid GPU-FPGA Clusters." *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'21)*
- [10] Yu Zhu, Zhenhao He, **Wenqi Jiang**, Kai Zeng, Jingren Zhou, and Gustavo Alonso, "Distributed Recommendation Inference on FPGA Clusters." *31th International Conference on Field-Programmable Logic and Applications (FPL'21)*

- [11] **Wenqi Jiang**, Zhenhao He, Shuai Zhang, Thomas B. Preußer, Kai Zeng, Liang Feng, Jiansong Zhang, Tongxuan Liu, Yong Li, Jingren Zhou, Ce Zhang, and Gustavo Alonso, “MicroRec: Efficient Recommendation Inference by Hardware and Data Structure Solutions.” *4th Conference on Machine Learning and Systems (MLSys’21)*

JOURNAL PAPERS

- [1] Shaoxiong Ji, **Wenqi Jiang**, Anwar Walid, and Xue Li, “Dynamic Sampling and Selective Masking for Communication-Efficient Federated Learning.” *IEEE Intelligent Systems*, 2022

TUTORIALS

- [1] **Wenqi Jiang**, Dario Korolija, and Gustavo Alonso, “Data Processing with FPGAs on Modern Architectures.” *International Conference on Management of Data (SIGMOD’23)*

TEACHING

Guest Lecturer:

Cloud Computing Architecture	Spring 2025
Data Modelling and Databases	Spring 2024
Hardware Acceleration for Data Processing	Fall 2021, 2023

Teaching Assistant:

Big Data	Fall 2023
Big Data for Engineers	Spring 2022, 2023
Systems Programming and Computer Architecture	Fall 2021, 2022

PROFESSIONAL SERVICE

Program Committee: VLDB’27, EuroSys’27, ATC’26, HCDS@ASPLOS’26

REFERENCES

Available upon request