Naiqing (Neil) Guan

naiqing.guan@mail.utoronto.ca · https://www.linkedin.com/in/gnaiqing/ · (437)-684-8926

Executive Summary

- PhD candidate in the Department of Computer Science at the University of Toronto, specializing in data-centric machine learning and weak supervision.
- Published researcher with strong track record in top-tier venues (SIGMOD, VLDB, AAAI, ISCA) and industry research experience at Microsoft Research Asia.
- Core expertise in efficient and reliable data annotation in machine learning pipelines, with recent focus on LLM applications and active learning.
- Demonstrated ability to reduce computational costs (4000X reduction) while maintaining model performance in labeling function design through novel algorithmic approaches.

Education

Department of Computer Science, University of Toronto Toronto, Canada
PhD student of Professor Nick Koudas Sep 2020 – present (expected to graduate in Dec 2025)

- Research Interests: data-centric machine learning, LLM agents, weak supervision, active learning
- CGPA: 4.0/4.0

Department of Electronic Engineering and Computer Science, Peking University Beijing, China
Undergraduate Sep 2016 – Jul 2020

- Major: Computer Science and Technology Minor: Philosophy
- Major GPA: 3.72/4.00 (rank top 10% in Department of Computer Science and Technology)

Publications

1. **Naiqing Guan**, and Nick Koudas. "WeShap: Weak Supervision Source Evaluation with Shapley Values." arXiv preprint arXiv:2406.11010 (2024). (accepted by VLDB 2025)
2. **Naiqing Guan**, Kaiwen Chen, and Nick Koudas. "DataSculpt: Cost-Efficient Label Function Design via Prompting Large Language Models." (2025). (accepted by EDBT 2025)
3. **Naiqing Guan**, and Nick Koudas. "ActiveDP: Bridging Active Learning and Data Programming." arXiv preprint arXiv:2402.06056 (2024). (accepted by EDBT 2024)
4. Ali Harakeh, Jordan Sir Kwang Hu, **Naiqing Guan**, Steven Waslander, and Liam Paull. "Estimating regression predictive distributions with sample networks." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 6, pp. 7830-7838. 2023.
5. **Naiqing Guan**, and Nick Koudas. "FILA: Online Auditing of Machine Learning Model Accuracy under Finite Labelling Budget." In Proceedings of the 2022 International Conference on Management of Data (SIGMOD), pp. 1784-1794. 2022.
6. Liqiang Lu*, **Naiqing Guan***, Yuyue Wang, Liancheng Jia, Zizhang Luo, Jieming Yin, Jason Cong, and Yun Liang. "Tenet: A framework for modeling tensor dataflow based on relation-centric notation." In 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), pp. 720-733. IEEE, 2021.
7. Lin Hu, **Naiqing Guan**, and Lei Zou. "Triangle counting on GPU using fine-grained task distribution." In 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW), pp. 225-232. IEEE, 2019.

Research Experience

University of Toronto (Data Systems Group) Toronto, Canada
PhD Student of Professor Nick Koudas Sep 2020 – present

WeShap: Weak Supervision Source Evaluation with Shapley Values (accepted by VLDB 2025)

- Proposed WeShap, a comprehensive and efficient evaluation metric for labeling functions (noisy rules for automatic data labeling in weak supervision) based on Shapley values, leveraging techniques from the game theory to measure the contribution of every labeling function in the data labeling pipeline.
- WeShap has versatile applications, including identifying harmful labeling functions, improving downstream model performance, and understanding model behaviors. Notably, WeShap leads to a 5.0 points downstream model accuracy improvement on average compared to SOTA works when revising labeling functions.

DataSculpt: Designing Accurate Labeling Functions using Large Language Models (accepted by EDBT 2025)

- Developed DataSculpt, a novel framework for developing labeling functions automatically and cost-efficiently with prompting techniques, employing advanced large language models like GPT-4.
- Compared to SOTA works in labeling function design using LLMs, DataSculpt is highly cost-efficient, achieving comparable downstream model accuracy while reducing token cost by over 4000X.

ActiveDP: Bridging Active Learning with Programmatic Weak Supervision (accepted by EDBT 2024)

- Introduced ActiveDP, an interactive framework that integrates weak supervision with active learning for effective large dataset annotation with high label accuracy and coverage.

- ActiveDP exceeds previous weak supervision and active learning methods by 2.6-13.5 points in downstream model accuracy, delivering consistent performance under varying labeling budgets.

Estimating regression predictive distributions with sample networks (accepted by AAAI 2023, collaborated with TRAIL lab at the University of Toronto)

- Developed SampleNet, a flexible and scalable architecture for modeling uncertainty in deep neural network predictions that avoids specifying a parametric form on the output distribution.
- SampleNet rivals or outperforms prior SOTA works on uncertainty estimation in all 15 evaluated datasets.

FILA: Online Auditing of Machine Learning Model Accuracy (accepted by SIGMOD 2022)

- Proposed FILA, an innovative stratified sampling technique for auditing machine learning models post-deployment that minimizes estimation error in class imbalanced settings within limited labeling budget.
- Formally proved that FILA is asymptotically optimal for sample allocation, and experimentally verified its effectiveness. Experiments show that FILA reduces over 20% estimation error compared to SOTA works.

Peking University (Center for Energy-Efficient Computing and Applications) Beijing, China
Intern Student of Professor Yun(Eric) Liang Dec 2019 – Jul 2020

TENET: Dataflow Analysis of DNN Operators on Spatial Hardware (accepted by ISCA 2021)

- Innovated a new relation-centric notation to model and analyze dataflows in spatial architectures, surpassing existing notations in expressiveness, facilitating the identification of complex dataflows and unlocking the full capabilities of spatial architectures.
- TENET achieves 37.4% and 51.4% latency reduction for CONV and GEMM kernels compared with the SOTA data-centric notation by identifying more sophisticated hardware dataflows.

Peking University (Laboratory of Management of Data) Beijing, China
Intern Student of Professor Lei Zou Mar 2018 – Jul 2019

Large Scale Triangle Counting on GPU (accepted by ICDE Workshop 2019)

- Developed a fine-grained task distribution strategy for the triangle counting task in graph data processing, addressing the challenges of GPU-based algorithm design due to data irregularity.
- Achieved 2x to 7x acceleration on large graphs, outpacing SOTA algorithms.

WORK EXPERIENCE

Teaching assistant at the University of Toronto Toronto, Canada
Department of Computer Science, University of Toronto Sep 2022 – Apr 2024

- Served as a Teaching Assistant for multiple courses at the University of Toronto, encompassing Introduction to Databases, Introduction to Programming Language, and Advanced Data Systems.
- Conducted office hours to provide academic support, clarified course material, and facilitated learning for students. Assessed and graded a range of student work, including homework and examinations.

Research Internship Beijing, China
Microsoft Research Asia Sep 2019 – Dec 2019

- Collaborated with Fan Yang in the systems group at Microsoft Research Asia (MSRA) to develop a DNN instrumentation framework, abstracting DNN analysis and optimization tasks as instrumentation tools.
- Conducted comprehensive literature reviews to support the development process and contributed to the creation of a research prototype for the DNN instrumentation framework.

SELECTED AWARDS AND HONORS

- Outstanding Graduates of Peking University 2020
- Meritorious Winner of 2019 ICM/MCM competition 2019
- Jinlongyu Scholarship (top 2 in class comprehensive assessment) 2017
- Panasonic Scholarship (top 2 in class comprehensive assessment) 2016
- Silver medal of National Olympiad in Informatics 2015

ADDITIONAL INFORMATION

Extracurricular Experiences

- Division Leader of Chinese Students and Scholars Association at the University of Toronto 2022-2023
- Mentor of Public Welfare Summer Camp of FutureChina NGO (林荫公益) 2022-2023

Programing Languages and Frameworks

- Python, C, C++, Pascal, MATLAB
- Pytorch, Pandas, Scikit-learn