

---

# YUE YU

E-mail: [yueyu@gatech.edu](mailto:yueyu@gatech.edu) | Webpage: [yueyu1030.github.io](https://yueyu1030.github.io)

Address: CODA Building 1312, 756 W Peachtree St NW, Atlanta, GA 30308, USA

## EDUCATION

School of Computational Science and Engineering, Georgia Institute of Technology  
Ph.D. in Computational Science and Engineering

Atlanta, GA, USA  
Aug. 2019 - Present

- Ph.D. Advisor: [Dr. Chao Zhang](#);
- **Research Interest:** Pretrained Language Models, Data-centric AI (e.g. Active/Interactive Learning, Weak Supervision).
- **Thesis Topic:** Towards Efficiently and Effectively Harnessing Large Pre-trained Models via Data-centric Lens.

Department of Electronic Engineering, Tsinghua University  
B.Eng. in Electronic Engineering

Beijing, China  
Aug. 2015 - July 2019

- Research Assistant in the Future Internet & Communication Lab advised by [Dr. Yong Li](#);

## INDUSTRY EXPERIENCE

News Understanding Team, Google Research

New York City, NY, USA  
May 2023 -

Research Intern, Host: [Jiaming Shen](#), Co-host: [Tianqi Liu](#), Manager: [Jialu Liu](#)

**Topic:** Empowering Large Language Model In-context Learning with Free-text Rationales.

Productivity and Intelligence Group, Microsoft Research

Redmond, WA, USA  
May 2021 - Aug. 2021

Research Intern, Mentor: [Chenyan Xiong](#), Manager: [Arnold Overwijk](#)

**Topic:** Zero-shot Learning for Generalizable Dense Text Retrieval.

**Publication:** One conference paper in [EMNLP 2022](#).

Analytics Center of Excellence, IQVIA

Boston, MA, USA  
May 2020 - Aug. 2020

Machine Learning Research Intern, Mentor: [Cao \(Danica\) Xiao](#)

**Topic:** Multi-typed Drug Interaction Prediction via Knowledge Graph Summarization.

**Publication:** One journal paper in [Bioinformatics 2021](#).

## SELECTED PUBLICATIONS

(The full publication list can be found in [this link](#), \* stands for equal contribution):

1. **Yue Yu\***, Yuchen Zhuang\*, Jieyu Zhang\*, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, Chao Zhang. "Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias". In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks (NeurIPS)*, 2023.
2. Yuchen Zhuang\*, **Yue Yu\***, Kuan Wang\*, Haotian Sun, Chao Zhang. "ToolQA: A Dataset for LLM Question Answering with External Tools". In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks (NeurIPS)*, 2023.
3. **Yue Yu**, Rongzhi Zhang, Ran Xu, Jieyu Zhang, Jiaming Shen and Chao Zhang. "[Cold-Start Data Selection for Few-shot Language Model Fine-tuning: A Prompt-Based Uncertainty Propagation Approach](#)." In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.
4. **Yue Yu**, Yuchen Zhuang, Rongzhi Zhang, Yu Meng, Jiaming Shen, and Chao Zhang. "[REGEN: Zero-Shot Text Classification via Training Data Generation with Progressive Dense Retrieval](#)." In *Findings of the Association for Computational Linguistics: ACL 2023 (Findings of ACL)*, 2023.
5. **Yue Yu**, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. "[COCO-DR: Combating the Distribution Shift in Zero-Shot Dense Retrieval with Contrastive and Distributionally Robust Learning](#)." In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.
6. **Yue Yu**, Ling kai Kong, Jieyu Zhang, Rongzhi Zhang, and Chao Zhang. "[AcTune: Uncertainty-Based Active Self-Training for Active Fine-Tuning of Pretrained Language Models](#)." In *Proceedings of the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2022.
7. **Yue Yu\***, Simiao Zuo\*, Haoming Jiang, Wendi Ren, Tuo Zhao and Chao Zhang, "[Fine-Tuning Pre-trained Language Model with Weak Supervision: A Contrastive-Regularized Self-Training Approach](#)", In *Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2021.

8. **Yue Yu**, Yinghao Li, Jiaming Shen, Hao Feng, Jimeng Sun and Chao Zhang, “[STEAM: Self-Supervised Taxonomy Expansion via Path-Based Multi-View Co-Training](#)”, In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2020.
9. Chen Liang\*, **Yue Yu\***, Haoming Jiang\*, Siawpeng Er, Ruijia Wang, Tuo Zhao and Chao Zhang, “[BOND: Bert-Assisted Open-Domain Named Entity Recognition with Distant Supervision](#)”, In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2020.
10. Jieyu Zhang, **Yue Yu**, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang and Alexander Ratner, “[WRENCH: A Comprehensive Benchmark for Weak Supervision](#)” In *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 2021.

## RESEARCH EXPERIENCE

### Data Mining and Machine Learning Group, Georgia Tech

Atlanta, GA, USA

Advisor: Dr. Chao Zhang

- **Efficiently and Effectively Harnessing Large Pre-trained Models via Data-centric Lens** Aug. 2019 - Now
  - **Language Model Fine-tuning with Weak Labels**: Adopted self-training with contrastive regularization on sample pairs to improve the robustness of self-training for fine-tuning Language Models; Leveraged prompts to design additional labeling rules for improving the performance with human feedbacks.
  - **Active Fine-tuning of Language Model**: Designed active self-training framework to enhance the performance of fine-tuning pretrained language models with limited budgets; Proposed techniques to strategically select training examples to improve the performance of few-shot language model fine-tuning with prompts.
  - **Large Language Models for**

### Future Internet & Communication Lab, Tsinghua University

Beijing, China

Advisor: Dr. Yong Li

- **Spatio-temporal Data Mining and Recommender Systems** Dec. 2017 - July 2019
  - **Urban Dynamics Modeling**: Designed a novel urban dynamic revealing system based on state-sharing HMM to identify the typical dynamic patterns on various regions of the city with different urban functions.
  - **Privacy-preserving Recommendation**: Presented a new framework for privacy-preserving cross-domain recommendation. Designed confidence-enhanced collective matrix factorization (CCMF) to balance the effect of two domains.
  - **App Usage Representation Learning**: Built a heterogeneous App usage graph regarding App, time, and location units as nodes and their co-occurrence relations as edges. Developed a Graph Convolutional Network with meta path-based objective function to learn the semantic-aware representations.

## HONORS AND AWARDS

- Best Paper Award at Machine Learning for Health 2022 Nov. 2022
- ACM SIGKDD Student Registration Award Aug. 2020
- Excellent Graduate, Tsinghua University & Beijing City (Top 2% over 3292 graduate students) July 2019
- Comprehensive Scholarship, Tsinghua University (Top 1%) Oct. 2018
- Award from Tsinghua University Initiative Scientific Research Program (5000 USD) May 2018
- Comprehensive Scholarship, Tsinghua University (Top 5%) Oct. 2016, Oct. 2017

## PROFESSIONAL SKILLS

- Programming language: C++ , Python, MATLAB, Latex.
- Deep learning frameworks: Keras, Pytorch.

## SERVICES

- **Teaching Experience**: Teaching Assistant for CX4240: Introduction to Computational Data Analysis. Spring 2020, 2021
- **Conference Program Committee**: ICLR 2024; ACL 2023; KDD 2023; IJCAI 2023; NeurIPS 2022, 2023; EMNLP 2022, 2023; LOG 2022, 2023.
- **Reviewing Experience**: NeurIPS 2022, 2023; EMNLP 2022, 2023; LOG 2022, 2023; ACL 2023, IJCAI 2023, ICML 2022; ACL Rolling Review (ARR) 2021, 2022, 2023; KDD 2021; TKDE 2020; AAAI 2020; CIKM 2019.