# Yao Qiang

☎ +1 (313) 329-3094 | ✉ yaocsphd@gmail.com | Google Scholar | in: LinkedIn | 🌐: Website

### RESEARCH INTERESTS

- Natural Language Processing (NLP) & Large Language Model (LLM)
- Trustworthy AI: Fairness, Explainability, Robustness
- Machine Learning Theory & Applications

### EDUCATION

**Wayne State University**, Detroit, Michigan, USA 09/2019 – Expected 05/2024

- Doctor of Philosophy in Computer Science
- Advisor: Dr. Dongxiao Zhu

**Wayne State University**, Detroit, Michigan, USA 09/2018 – 12/2019

- Master of Science in Computer Science

**Xidian University**, Xi'an, China 09/2006 – 07/2010

- Bachelor of Science in Computer Science

### WORK EXPERIENCE

**Trustworthy AI Lab, Wayne State University** 09/2019 – Present
Graduate Research Assistant

**Robust and Modeling Team, Alexa, Amazon** 05/2023 – 08/2023
Applied Scientist Intern

**Mike Ilitch School of Business, WSU** 08/2018 – 08/2019
Student Research Assistant, Part-time

**Xi'an Microelectronics Technology Institute** 08/2010 – 12/2017
Computer Hardware Designer

### TEACHING EXPERIENCE

- Instructor for CSC 2111 Computer Science: Lab 2020
  - Topic: C++ Programming: From Problem Analysis to Program Design
  - Tools: Visual Studio C++
  - Lectures: 24 labs
  - Enrollment: 30 students
- Instructor for CSC 3101 Computer Architecture and Organization: Lab 2021
  - Topic: Digital Design and Computer Architecture
  - Tools: Logicly, Minecraft Educational Edition, x86 Assembly
  - Lectures: 12 labs
  - Enrollment: 30 students
- Invited Lecturer for CSC 5825 Machine Learning&Apps (Graduate Level) 2020 – 2023
  - Topic: Generative Model Theory and Application, Machine Learning System Design
  - Lectures: 2 lectures
  - Enrollment: 40 students
- Invited Lecturer for CSC 7825 Machine Learning (Graduate Level) 2020 – 2022
  - Topic: Deep Learning Frameworks Introduction and Application
  - Lectures: 2 lectures
  - Enrollment: 30 students
- Teaching Assistant for CSC 2111 Computer Science 2020
- Teaching Assistant for CSC 3101 Computer Architecture and Organization 2021
- Teaching Assistant for CSC 5825 Machine Learning&Apps (Graduate Level) 2019, 2020, 2022
- Teaching Assistant for CSC 6580 Design and Analysis of Algorithms (Graduate Level) 2020
- Teaching Assistant for CSC 7825 Machine Learning (Graduate Level) 2019 – 2020

**PUBLICATIONS**

**Google Scholar**: https://scholar.google.com/citations?user=8ADcg38AAAAJ&hl=en

**Publications**

- "Prompt Perturbation Consistency Learning (PPCL) for Robust Language Models"

  **Yao Qiang**, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky, Aram Galstyan

  In 18th Conference of the European Chapter of the Association for Computational Linguistics, **EACL** 2024.

- "Attcat: Explaining transformers via attentive class activation tokens"

  **Yao Qiang**, Deng Pan, Chengyin Li, Xin Li, Rhongho Jang, and Dongxiao Zhu

  Advances in Neural Information Processing Systems 35: 5052-5064, **NeurIPS** 2022.

- "Counterfactual interpolation augmentation (CIA): A unified approach to enhance fairness and explainability of DNN"

  **Yao Qiang**, Chengyin Li, Marco Brocanelli, and Dongxiao Zhu

  In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, pp. 732-739, **IJCAI** 2022.

- "Tiny rnn model with certified robustness for text classification"

  **Yao Qiang**, Supriya Tumkur Suresh Kumar, Marco Brocanelli, and Dongxiao Zhu

  In 2022 International Joint Conference on Neural Networks, pp. 1-8. IEEE, **IJCNN** 2022.

- "Toward tag-free aspect based sentiment analysis: A multiple attention network approach"

  **Yao Qiang**, Xin Li, and Dongxiao Zhu

  In 2020 International Joint Conference on Neural Networks, pp. 1-8. IEEE, **IJCNN** 2020.

- "Benchmark and Neural Architecture for Conversational Entity Retrieval from a Knowledge Graph"

  Zamiri, M, **Yao Qiang**, Nikolaev, F, Zhu, D, and Kotov, A

  In the proceedings of the 2024 ACM Web Conference.

- "Learning compact features via in-training representation alignment"

  Xin Li, Xiangrui Li, Deng Pan, **Yao Qiang**, and Dongxiao Zhu

  In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 7, pp. 8675-8683. **AAAI**, 2023.

- "Negative Flux Aggregation to Estimate Feature Attributions"

  Xin Li, Deng Pan, Chengyin Li, **Yao Qiang**, and Dongxiao Zhu

  In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, **IJCAI**, 2023.

- "FocalUNETR: A Focal Transformer for Boundary-Aware Prostate Segmentation Using CT Images"

  Chengyin Li, **Yao Qiang**, Rafi Ibn Sultan, Hassan Bagher-Ebadian, Prashant Khanduri, Indrin J. Chetty, and Dongxiao Zhu

  In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 592-602. **MICCAI**, 2023.

- "Saliency guided adversarial training for learning generalizable features with applications to medical imaging classification system"

  Xin Li, **Yao Qiang**, Chengyin Li, Sijia Liu, and Dongxiao Zhu

  In The First Workshop on New Frontiers in Adversarial Machine Learning. **ICML** workshop, 2022.

- "Proximal Compositional Optimization for Distributionally Robust Learning"

  Prashant Khanduri, Chengyin Li, Rafi Ibn Sultan, **Yao Qiang**, Joerg Kliewer, and Dongxiao Zhu

  In The Second Workshop on New Frontiers in Adversarial Machine Learning. **ICML** workshop, 2023.

### Pre-prints

- "Learning to Poison Large Language Models During Instruction Tuning"  
  **Yao Qiang**, Zhou, X, Zare Zade, S, Rosani A, Zytko, D, and Zhu, D  
  arXiv:2402.13459 [cs.LG], 2024.
- "Hijacking Large Language Models via Adversarial In-Context Learning"  
  **Yao Qiang**, Xiangyu Zhou, and Dongxiao Zhu  
  arXiv:2311.09948 [cs.LG], 2023.
- "Fairness-aware Vision Transformer via Debiased Self-Attention"  
  **Yao Qiang**, Chengyin Li, Prashant Khanduri, and Dongxiao Zhu  
  arXiv preprint arXiv:2301.13803, 2023.
- "Interpretability-Aware Vision Transformer"  
  **Yao Qiang**, Chengyin Li, Prashant Khanduri, and Dongxiao Zhu  
  arXiv preprint arXiv:2309.08035, 2023.
- "Adversarially Robust and Explainable Model Compression with On-Device Personalization for Text Classification"  
  **Yao Qiang**, Supriya Tumkur Suresh Kumar, Marco Brocanelli, and Dongxiao Zhu  
  arXiv preprint arXiv:2101.05624, 2021.
- "Auto-Prompting SAM for Mobile Friendly 3D Medical Image Segmentation"  
  Chengyin Li, Prashant Khanduri, **Yao Qiang**, Rafi Ibn Sultan, Indrin Chetty, and Dongxiao Zhu  
  arXiv preprint arXiv:2308.14936, 2023.

## HONORS&AWARDS

| | |
|---|---|
| Michael E. Conrad Award (Highest Honor at WSU CS Department) | 2023 |
| AAAI 2023 Student Scholarship | 2022 |
| NeurIPS 2022 Scholar Award | 2022 |
| Department Travel Award for Outstanding Conference Publications | 2022 |
| Graduate Student Professional Travel Award | 2022 |
| IEEE CIS Conference Participation and Travel Grants | 2022 |
| IJCAI 2022 Travel and Accessibility Grant | 2022 |
| Department Oustanding GTA Award | 2020 |
| Graduate School Master's Scholarship Award | 2019 |

## SERVICES

### Program Committee Member

| | |
|---|---|
| SIAM International Conference on Data Mining (SDM) | 2023 |
| ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) | 2023 |
| AAAI Conference on Artificial Intelligence (AAAI) | 2022 – 2023 |
| Adversarial Machine Learning Frontiers (ICML Workshop) | 2022 – 2023 |

### Conference Reviewer

| | |
|---|---|
| SIAM International Conference on Data Mining (SDM) | 2023 |
| IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR) | 2023 |
| International Conference on Machine Learning (ICML) | 2022 – 2024 |
| International Joint Conferences on Artificial Intelligence (IJCAI) | 2021 – 2024 |
| ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) | 2023 –2024 |
| AAAI Conference on Artificial Intelligence (AAAI) | 2020 – 2023 |
| Conference on Neural Information Processing Systems (NeurIPS) | 2020 – 2023 |
| International Conference on Learning Representations (ICLR) | 2022 – 2023 |
| Medical Image Computing and Computer Assisted Intervention (MICCAI) | 2022 – 2023 |
| Adversarial Machine Learning Frontiers (ICML Workshop) | 2022 – 2023 |

### Journal Reviewer

| | |
|---|---|
| ACM Transactions on Internet of Things (TIOT) | 2021 |
| Artificial Intelligence (AI) | 2022 |
| ACM Transactions on Knowledge Discovery from Data (TKDD) | 2023 |

### Conference Student Volunteering

| | |
|---|---|
| AAAI Conference on Artificial Intelligence (AAAI) | 2023 |
| Conference on Neural Information Processing Systems (NeurIPS) | 2022 |
| International Joint Conferences on Artificial Intelligence (IJCAI) | 2022 |
| International Joint Conference on Neural Networks (IJCNN) | 2022 |