# Yuqing Zhou

✉ yzhou31@gmu.edu | yuqing-zhou.github.io

## Research Interests

LLM, Trustworthy AI, Robust Machine Learning, Causal Inference, Explainable AI, Agentic AI

## Education

**George Mason University**, Fairfax, VA, USA 08/2023 – Present
*Ph.D. Student, Computer Science*
**University of Michigan**, Ann Arbor, MI, USA 08/2019 – 04/2021
*M.S., Electrical and Computer Engineering (Computer Vision Track)*
**Southeast University**, Nanjing, Jiangsu, China 08/2015 – 06/2019
*B.Eng., Electronics Science and Technology*

## Work Experience

**Applied Scientist Intern**, Amazon Web Services (AWS), Seattle, WA, USA 05/2025 – 08/2025
Research on *Agentic Conversational AI Assistant*
**Graduate Research Assistant**, George Mason University, Fairfax, VA, USA 08/2023 – 08/2024
Research on *Mitigating Spurious Correlations in Machine Learning*
**Software Engineer**, Shanghai Huawei Technologies Co., Ltd, Shanghai, China 08/2021 – 05/2023

## Publications

[1] **Yuqing Zhou** and Ziwei Zhu. "Fighting Spurious Correlations in Text Classification via a Causal Learning Perspective". In: *NAACL* (2025).

[2] **Yuqing Zhou**, Ruixiang Tang, Ziyu Yao, and Ziwei Zhu. "Navigating the Shortcut Maze: A Comprehensive Analysis of Shortcut Learning in Text Classification by Language Models". In: *Findings of EMNLP* (2024).

[3] **Yuqing Zhou**, Tianshu Feng, Mingrui Liu, and Ziwei Zhu. "A Generalized Propensity Learning Framework for Unbiased Post-Click Conversion Rate Estimation". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM. 2023.

## Selected Projects

### Causally Calibrated Robust Classifier (CCR) for Text Classification 03/2024 – 10/2024

- Developed CCR to improve text classification robustness by mitigating reliance on spurious correlations, using causal feature selection and an inverse propensity weighting (IPW) loss.
- Achieved state-of-the-art performance on multiple across multiple tasks, with a 70.7% worst group accuracy on the CivilComments dataset that even outperforms methods Group-DRO and DFR that require group labels.

### Shortcut Learning Analysis in Language Models for Text Classification 09/2023 – 06/2024

- Proposed a systematic shortcut framework in text classification with three main shortcut types: occurrence, style, and concept.
- Generated datasets based on three public text classification datasets to construct a benchmark under the shortcut framework, such as using **Llama2-70b** to synthesize data with embedded style-based shortcuts.
- Finetuned multiple models on the synthetic datasets, including BERT, Llama, and state-of-the-art robust methods, uncovering their susceptibility to different types of shortcuts.
- **Finetuned large language models (LLMs)**, such as Llama2-7b, Llama2-13b, and Llama3-8b, on synthetic datasets, demonstrating that larger model sizes alone do not guarantee improved robustness.

### Generalized Propensity Learning (GPL) for Post-Click Conversion Rate Prediction 05/2023 – 06/2023

- Developed GPL framework to minimize bias and variance in CVR prediction for recommender systems, enhancing the performance and robustness of existing methods like IPS and DR-based estimators.
- Improved CVR prediction accuracy by 7% in DCG@2 and 6% in Recall@2 on the Yahoo dataset, demonstrating significant performance gains.

## Skills

**Languages**: Python, C/C++, Julia, MATLAB
**Frameworks**: PyTorch

## Teaching Experience

**GTA** CS 455: Computer Communications and Networking, GMU Spring 2025, Fall 2024