

# Wenqian Ye

85 Engineer's Way, Rice Hall  
Charlottesville, VA  
571-519-9957  
✉ wenqian@virginia.edu  
↗ Google Scholar  
🔗 Github   ↳ LinkedIn

## Research Statement

My research focuses on robustness and alignment in machine learning systems. I have worked extensively on out-of-domain (OOD) generalization, uncertainty quantification, and the impact of spurious correlations in classical machine learning. More recently, my work has shifted toward AI alignment, with a particular focus on understanding and mitigating reward hacking behaviors in Large Language Models (LLMs) and Agentic AI.

## Education

- 2023 – Now **PhD in Computer Science**, *School of Engineering and Applied Science*, University of Virginia.  
Advisor: Aidong Zhang
- 2020 – 2022 **MS in Computer Science**, *Courant Institute of Mathematical Sciences*, New York University.  
Concentration: Machine Learning
- 2017 – 2020 **BS in Mathematics**, University of Illinois Urbana-Champaign, High Distinction.  
Thesis Advisor: Sanjay Patel  
*Minor in Computer Science and Electrical Engineering*

## Selected Publications († denotes co-first authors)

- 2025 **Wenqian Ye, Guangtao Zheng, Aidong Zhang**, *Rectifying Shortcut Behaviors in Preference-based Reward Learning*, Advances in Neural Information Processing Systems (NeurIPS).
- 2025 **Wenqian Ye, Guangtao Zheng, Aidong Zhang**, *Improving Group Robustness on Spurious Correlation via Evidential Alignment*, ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).  
*Best Paper Award*
- 2025 **Guangtao Zheng, Wenqian Ye, Aidong Zhang**, *NeuronTune: Towards Self-Guided Spurious Bias Mitigation*, International Conference on Machine Learning (ICML).
- 2025 **Guangtao Zheng, Wenqian Ye, Aidong Zhang**, *ShortcutProbe: Probing Prediction Shortcuts for Learning Robust Models*, International Joint Conference on Artificial Intelligence (IJCAI).
- 2025 **Wei Qian, Chenxu Zhao, Yangyi Li, Wenqian Ye, Mengdi Huai**, *Towards Unveiling Predictive Uncertainty Vulnerabilities in the Context of the Right to Be Forgotten*, Conference on Information and Knowledge Management (CIKM).
- 2025 **Xu Cao, Yifan Shen, Bolin Lai, Wenqian Ye, Yunsheng Ma, Joerg Heintz, Jintai Chen, Jianguo Cao, James M Rehg**, *What is the Visual Cognition Gap between Humans and Multimodal LLMs?*, Conference on Language Modeling (COLM).
- 2024 **Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James Rehg, Aidong Zhang**, *MM-SpuBench: Towards Better Understanding of Spurious Biases in Multimodal LLMs*, NeurIPS Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models.  
*Oral Presentation*
- 2024 **Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, Aidong Zhang**, *Spurious Correlations in Machine Learning: A Survey*, ICML Workshop on Data-Centric Machine Learning Research.

- 2024 **Guangtao Zheng, Wenqian Ye, Aidong Zhang**, *Benchmarking Spurious Bias in Few-Shot Image Classifiers*, European Conference on Computer Vision (ECCV).
- 2024 **Guangtao Zheng, Wenqian Ye, Aidong Zhang**, *Spuriousness-Aware Meta-Learning for Learning Robust Classifiers*, ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).
- 2024 **Guangtao Zheng, Wenqian Ye, Aidong Zhang**, *Learning Robust Classifiers with Self-Guided Spurious Correlation Mitigation*, International Joint Conference on Artificial Intelligence (IJCAI).
- 2024 **Xu Cao<sup>†</sup>, Wenqian Ye<sup>†</sup>, Kenny Moise, Megan Coffee**, *MpoxVLM: A Vision-Language Model for Diagnosing Skin Lesions from Mpox Virus Infection*, Machine Learning for Health Symposium (ML4H).
- 2024 **Yunsheng Ma, Xu Cao, Wenqian Ye, Can Cui, Kai Mei, Ziran Wang**, *Learning Autonomous Driving Tasks via Human Feedbacks with Large Language Models*, Findings in Conference on Empirical Methods in Natural Language Processing (EMNLP Findings).
- 2024 **Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James Rehg, Chao Zheng**, *MAPLM: A Real-World Large-Scale Vision-Language Dataset for Map and Traffic Scene Understanding*, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- 2024 **Yunsheng Ma, Can Cui, Xu Cao, Wenqian Ye, Peiran Liu, Juanwu Lu, Amr Abdelraouf, Rohit Gupta, Kyungtae Han, Aniket Bera, James Rehg, Ziran Wang**, *LaMPilot: An Open Benchmark Dataset for Autonomous Driving with Language Model Programs*, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- 2023 **Wenqian Ye, Yunsheng Ma, Xu Cao, Kun Tang**, *Mitigating Transformer Overconfidence via Lipschitz Regularization*, Conference on Uncertainty in Artificial Intelligence (UAI).
- 2023 **Xu Cao<sup>†</sup>, Wenqian Ye<sup>†</sup>, Elena Sizikova, Xue Bai, Megan Coffee, Hongwu Zeng, Jianguo Cao**, *ViTASD: Robust ViT Baselines for Autism Spectrum Disorder Facial Detection*, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).
- 2023 **Yunsheng Ma, Wenqian Ye, Xu Cao, Amr Abdelraouf, Kyungtae Han, Rohit Gupta, Ziran Wang**, *CEMFormer: Learning to Predict Driver Intentions from In-Cabin and External Cameras via Spatial-Temporal Transformers*, IEEE Intelligent Transportation Systems Conference (ITSC).
- 2023 **Wenqian Ye<sup>†</sup>, Yunsheng Ma<sup>†</sup>, Xu Cao**, *Uncertainty Estimation in Deterministic Vision Transformer*, AAAI Workshop on Uncertainty Reasoning and Quantification in Decision Making (UDM-AAAI).

## Professional Experience

- 2025 **Applied Scientist Intern**, AWS AI Fundamental Research.  
Research on improving agentic alignment with reward reasoning models. Analyze this problem with a Bayesian inference and reward maximization perspective.
- 2023 – Now **Graduate Research Assistant**, University of Virginia.  
Conduct research on improving robustness and alignment of machine learning models. Publish research in top AI and data mining conferences, and develop open-source tools.
- 2022 – Now **Adjunct Researcher**, NYU Langone Health, New York University.  
Conduct part-time research on Artificial Intelligence-enabled diagnosis of Tuberculosis and COVID-19 using radiologic imaging in resource-constrained environments. Lead the development of AI algorithmic frameworks (e.g., VLMs and AI agents) for screening Monkeypox using dermatologic images.
- 2022 – 2023 **Software Engineer**, Cirrus Logic Inc.  
Performed comprehensive validation and testing of embedded software for audio and haptics applications, focusing on automation and analysis. Contributed to both internal and customer-facing UI design, while executing system-level testing across device drivers, firmware, and UI software. Developed prototypes of DSP algorithms using Python/Matlab and implemented fixed-point firmware using C/C++.

## Fellowships & Grants

- 2025 NeurIPS 2025 Scholar Award
- 2025 KDD Best Paper Award
- 2024 OpenAI Researcher Access Program
- 2023 UAI Travel Award
- 2023 AAAI Travel Award
- 2023 UVA Computer Science Fellowship

## Teaching Experience

- Fall 2025 **CS 4774: Machine Learning**, Prof. Hadi Daneshmand, University of Virginia.
  - Design projects and manage course logistics.
  - Lead weekly office hours and actively supported students on Canvas.
- Spring 2025 **CS 4501/6501: Analyzing Online Behavior for Public Health**, Prof. Henry Kautz, University of Virginia.
  - Graded assignments, projects and provided detailed feedback.
  - Led weekly office hours and actively supported students on Canvas.
- Fall 2024 **CS 6316: Machine Learning**, Prof. Aidong Zhang, University of Virginia.  
Guest Lecture on the topic of Spurious Correlations in Machine Learning.
- Fall 2024 **CS 4501: Natural Language Processing**, Prof. Yu Meng, University of Virginia.
  - Designed coding/conceptual assignments for the course contents.
  - Graded assignments and provided detailed feedback.
  - Led weekly office hours and actively supported students on Canvas.
- Fall 2021 **CSCI-GA 2590: Natural Language Processing**, Prof. He He, New York University.
  - Graded assignments, exams, and final projects.
  - Developed the autograder for coding assignments.
  - Led office hours and supported students on CampusWire.

## Services

- Organizer **Co-organizer/Program Chair.**  
WDFM-AD Workshop (CVPR 2025; ICCV 2025);  
LLVM-AD Workshop (WACV 2024; ITSC 2024; WACV 2025);  
AI4CHL Workshop (ICLR 2025);
- Program **Journals.**  
Committee ACM TIST; IEEE TPAMI; IEEE IoT-J; IJHCI; IEEE T-IV; IEEE VTM; IEEE Internet Computing
- Conferences **Conferences.**  
ICML; ICLR; NeurIPS; KDD; CVPR; ECCV; ICCV; AAAI; IJCAI; AISTATS; ICASSP; MICCAI; ISBI; ACML
- Workshops **Workshops.**  
DMLR(ICML); MLSP; VTNA(NeurIPS); NIVIT(ICCV); UDM(AAAI, KDD)
- Membership **Member.**  
IEEE; ACM; IEEE SPS
- Mentorship **Mentor.**  
ML4H(2023, 2024)

## Technical Skills

- Languages **Python**, C/C++, R, MATLAB, Golang, SystemVerilog, L<sup>A</sup>T<sub>E</sub>X

Packages PyTorch, TensorFlow, AG2, LangChain/LangGraph, Huggingface, Scikit-learn

Others AWS Bedrock, CUDA, SQL, Git, Jenkins