Hannah Cyberey

Charlottesville, VA hannahcyberey@virginia.edu hannahxchen.github.io hannahxchen

Research Interests

- Bias and Fairness
- Trustworthy Machine Learning
- Adversarial Machine Learning
- Natural Language Processing (NLP)

Education

Ph.D. in Computer Science Aug 2019 – Aug 2025 (Expected)
University of Virginia, Charlottesville, VA
GPA: 4.0/4.0, Advisors: David Evans, Yangfeng Ji

B.S. in Information Management Sep 2014 – Jun 2018
Chang Gung University, Taoyuan, Taiwan
GPA: 3.53/4.0 (Last 60 GPA: 4.0/4.0)

Research Experience

Microsoft, Research Intern Feb 2022 – May 2022
Cryptography and Privacy Group, Mentors: Wei Dai, Kim Laine Redmond, WA

- Led research on investigating privacy leakage in large language models (LLMs) for code generation
- Implemented membership inference and training data reconstruction attacks
- Proposed several mitigation methods to improve the current pipeline

Institute for Information Industry, Machine Learning Intern Jun 2018 – Dec 2018
Cybersecurity Technology Institute, Mentor: Yu-De Lin, Manager: Ching-Hao Mao Taipei, Taiwan

- Exploratory data analysis of trends in security vulnerabilities and exposures on Twitter
- Built binary classifiers for the Secbuzzer System to identify security-related Tweets
- Developed Sec2Vec embedding method. sec2vec

Chang Gung University, Undergraduate Research Assistant Jul 2017 – Jun 2018
Lab of Ubiquitous Security and Applications, Advisor: Chien-Lung Hsu Taoyuan, Taiwan

- Assisted in IoT security research project
- Implemented device authentication using NTRU encryption in Java
- Programmed Raspberry Pi sensors to capture environmental data

Publications

Hannah Cyberey, Yangfeng Ji, David Evans. Sensing and Steering Stereotypes: Extracting and Applying Gender Representation Vectors in LLMs. In *Arxiv Preprint*, Feb 2025.

Hannah Cyberey, Yangfeng Ji, David Evans. The Mismeasure of Man and Models: Evaluating Allocational Harms in Large Language Models. In *Arxiv Preprint*, Aug 2024.

Hannah Cyberey, Yangfeng Ji, David Evans. Addressing Both Statistical and Causal Gender Fairness in NLP Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Jun 2024.

Hannah Cyberey, Yangfeng Ji, David Evans. Balanced Adversarial Training: Balancing Tradeoffs Between Oversensitivity and Undersensitivity in NLP Models. In *Proceedings of the 2022 Conference on Empirical Meth-*

ods in Natural Language Processing (EMNLP). Oct 2022.

Hannah Cyberey, Yangfeng Ji, David Evans. Finding Friends and Flipping Frenemies: Automatic Paraphrase Dataset Augmentation Using Graph Theory. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Nov 2022.

Hannah Cyberey, Yangfeng Ji, David Evans. Pointwise Paraphrase Appraisal Is Potentially Problematic. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Jul 2020.

Projects

Improving LLM Reliability through Representation Engineering 2024 - Present

- Proposed a novel method for extracting "steering vectors" from LLM internals to manipulate model outputs related to a specific concept
- Identified gender concept in LLMs using steering vectors and showed its effectiveness in mitigating gender bias in model predictions gender-bias-steering
- Applied steering vectors for LLM safety and censorship.

Replication of *Refusal in Language Models Is Mediated by a Single Direction* 2024

- Replicated the experiments and verified the claims made in the paper by Arditi et al.
- Conducted extended analysis and evaluation and showed potential limitations of the proposed method
- refusal-direction-replication

Bias and Fairness in NLP Models 2023 - 2024

- Proposed mitigation that addresses statistical and causal fairness criteria composed-debiasing
- Demonstrated prevalent bias metrics do not effectively indicate potential allocational harms from LLMs
- Proposed a new metric that shows a high correlation with group disparities in allocation decision outcomes allocational-harm-eval

Adversarial Robustness in NLP Models 2021 - 2022

- Demonstrated common adversarial training methods can lead to robustness tradeoffs in NLP models
- Proposed a new adversarial training method with minimal tradeoffs in model robustness balanced-adversarial-training

Improving Paraphrase Identification and Evaluation 2019 - 2020

- Demonstrated the current paraphrase evaluation setup can falsely indicate the model performance
- Developed a method for automatic dataset augmentation and labeling error detection, which improves model performance on paraphrase identification automatic-paraphrase-dataset-augmentation

Mentoring Experience

Varun Vejalla (UVA Undergrad) May 2023 – Nov 2023
Project: Evaluating Large Language Models for Bias

Jason Briegel (UVA Undergrad) May 2023 – Aug 2023
Project: Adjectives Can Reveal Gender Biases Within NLP Models (Blog Post)

Pragun Ananda (UVA Undergrad) May 2020 – Sep 2020
Project: Data Augmentation with Graph Theory

Teaching Experience

Data Privacy (UVA CS6501/CS4501)	Fall 2022
AI for Social Good (UVA CS6501)	Fall 2021
Natural Language Processing (UVA CS6501)	Spring 2021
Exploratory Text Analytics (UVA DS5001)	Fall 2020
Python Programming (CGU)	Spring 2018

Awards & Honors

- UVA Engineering Dean's Scholar Fellowship (2019 – 2024)
- Student member of IEEE HKN Gamma Pi Chapter at UVA (2021)
- Three times Presidential Awards (Top 3% of class): 2016 Fall, 2017 Spring, and 2017 Fall
- First runner-up of 2014 Chang Gung University English Speech Contest

Service

- President of Taiwanese Graduate Student Association (TGSA) at UVa (2022-2023)
- Reveiwer: NLPCC 2021, IJCNLP-AACL 2023, NeurIPS 2023 SoLaR Workshop, NAACL Insights Workshop 2025, ACL Rolling Review 2023-now

Skills

- **Programming Languages**: Python, HTML/CSS, JavaScript
- **Frameworks/Tools**: Pytorch, HuggingFace, Scikit-Learn, Plotly, Matplotlib, Pandas
- **Language**: English, Mandarin Chinese