

Hannah Cyberey

📍 Charlottesville, VA 📩 hannahcyberey@virginia.edu 🌐 hannahxchen.github.io 💬 [hannahxchen](https://www.linkedin.com/in/hannahxchen/)

Research Interests

- Trustworthy & Responsible AI
- Natural Language Processing (NLP)
- AI Ethics & Safety
- Large Language Models (LLMs)

Education

University of Virginia

Ph.D. in Computer Science

Advisors: [David Evans](#), [Yangfeng Ji](#)

Charlottesville, VA

2019 – 2025

Chang Gung University

B.S. in Information Management

Taoyuan, Taiwan

2014 – 2018

Experience

University of Virginia

Graduate Research Assistant

Charlottesville, VA

Apr 2019 – Aug 2025

- Conducted research on trustworthy NLP—robustness, bias & fairness, and censorship.
- Published five peer-reviewed papers as first author, and presented research at conferences
- Mentored five undergraduate students, providing guidance on their research projects

Microsoft

Research Intern – Cryptography and Privacy Group

Redmond, WA

Feb 2022 – May 2022

- Worked with a multidisciplinary research team to investigate privacy leakage and copyright implications of code generation models
- Implemented membership inference and training data reconstruction attacks
- Proposed several mitigation methods to improve the current pipeline

Institute for Information Industry

Machine Learning Intern – Cybersecurity Technology Institute

Taipei, Taiwan

Jun 2018 – Dec 2018

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

Chang Gung University

Undergraduate Research Assistant

Taoyuan, Taiwan

Jul 2017 – Jun 2018

- Assisted in IoT security research project
- Programmed device authentication and Raspberry Pi sensors for capturing environmental data

Awards & Honors

UVA Engineering Dean's Scholar Fellowship (2019 – 2024)

Student member of IEEE HKN Gamma Pi Chapter at UVA (2021)

Chang Gung University Presidential Awards (Top 3% of class): 2016 Fall, 2017 Spring, and 2017 Fall

First runner-up of Chang Gung University English Speech Contest (2014)

Publications

- Hannah Cyberez**, Yangfeng Ji, David Evans. [Unsupervised Concept Vector Extraction for Bias Control in LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Nov 2025.
- Hannah Cyberez**, David Evans. [Steering the CensorShip: Uncovering Representation Vectors for LLM “Thought” Control](#). In *Conference on Language Modeling (COLM)*. Oct 2025.
- Hannah Cyberez**, Yangfeng Ji, David Evans. [Do Prevalent Bias Metrics Capture Allocational Harms from LLMs?](#) In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*. May 2025
- Hannah Cyberez**, 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 Cyberez**, Yangfeng Ji, David Evans. [Balanced Adversarial Training: Balancing Tradeoffs Between Oversensitivity and Undersensitivity in NLP Models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Oct 2022.
- Hannah Cyberez**, 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 2020.
- Hannah Cyberez**, 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.

Talks

- Invited Talk: “Debiasing Can Be Complementary,” *The AI and Machine Learning Seminar at UVA*, Nov 2023
Guest Lecture: “Adversarial Attacks and Defenses for NLP Models,” *UVA CS6501/CS4501 Data Privacy*, Sep 2022

Selected Projects

- Countering Censorship in Instruction and Reasoning LLMs** [GitHub](#) 2025
• Examined LLM censorship mechanisms through internal representations and uncovered a distinct type of censorship in reasoning LLMs (e.g., DeepSeek-R1)
• Developed a method for detecting and controlling the level of censorship in LLM outputs
• Built a demo app with Gradio and deployed models with Docker on HuggingFace endpoints
- Bias Mitigation using Representation Engineering** [GitHub](#) 2024
• Proposed a method that finds “steering vectors” to control model outputs related to a specific concept
• Demonstrated the effectiveness of our method in reducing gender bias in model predictions
- Improving Adversarial Robustness in NLP Models** [GitHub](#) 2021 – 2022
• Demonstrated robustness tradeoffs in NLP models arise from common adversarial training methods
• Proposed a new adversarial training method with minimal tradeoffs in model robustness

Leadership Experience

- Co-Lead – Causal Learning Reading Group at UVA Summer 2023
President – Taiwanese Graduate Student Association at UVA 2022 - 2023

Teaching Experience

- Graduate Teaching Assistant – CS6501/4501 Data Privacy (Fall 2022), CS6501 AI for Social Good (Fall 2021), CS6501 Natural Language Processing (Spring 2021), DS5001 Exploratory Text Analytics (Fall 2020)
Undergraduate Teaching Assistant – Python Programming (Spring 2018)

Professional & Public Service

Reviewer: NLPCC 2021, IJCNLP-AACL 2023, NeurIPS SoLaR Workshop 2023, NAACL Insights Workshop 2025, ACL Rolling Review 2023-now

Volunteer: IEEE-HKN High School Outreach Program (2021)

Skills

Programming: Python, HTML/CSS, JavaScript

Frameworks/Tools: Pytorch, HuggingFace, Git, Jupyter Notebook, FastAPI, Docker

Libraries: Transformers, Pandas, Scikit-Learn, NumPy, SciPy, Plotly, Matplotlib, Gradio

Languages: English, Mandarin Chinese