

# Weiqli You

✉ [weiqliuy@seas.upenn.edu](mailto:weiqliuy@seas.upenn.edu)

🌐 <http://fallcat.github.io/>

Last updated: Dec. 11, 2024

## Research Interests

I build trustworthy machine learning models with faithful and verifiable explanations.

## Education

- 2020 – ▶ **Ph.D. Computer and Information Science, University of Pennsylvania**, Philadelphia, PA  
Advisor: Prof. Eric Wong  
Expected Graduation: May. 2026
- 2018 – 2020 ▶ **M.S. Computer Science, University of Massachusetts Amherst**, Amherst, MA  
Advisor: Prof. Mohit Iyyer
- 2014 – 2018 ▶ **B.S. Computer Science and Mathematics, Gordon College**, Wenham, MA  
Advisor: Prof. Jonathan Senning, Prof. Russell Bjork  
Double major. Honors Thesis title: *Predict Media Interestingness*.

## Internship & Employment History

- 2024 ▶ **Research Intern** Okinawa Institute of Science and Technology. Okinawa, Japan.
- 2022 ▶ **Research Intern** IBM Research Yorktown Heights. Yorktown Heights, NY.
- 2020 ▶ **Research Assistant** University of Southern California, Information Sciences Institute. Los Angeles, CA.
- 2018 ▶ **Research Intern** NLP Center, Meituan-Dianping Inc. Beijing, China.

## Publications

### Conferences and Journals

- 1 **Weiqli You**, Helen Qu, Marco Gatti, Bhuvnesh Jain, and Eric Wong (2025). “Sum-of-Parts: Self-Attributing Neural Networks with End-to-End Learning of Feature Groups”. In: *International Conference on Machine Learning (ICML)*. 🔗 URL: <https://openreview.net/forum?id=r6y9TEdLMh>.
- 2 Chaehyeon Kim, **Weiqli You**, Shreya Havaldar, and Eric Wong (2024). “Evaluating Groups of Features via Consistency, Contiguity, and Stability”. In: *The Second Tiny Papers Track at ICLR 2024*. 🔗 URL: <https://openreview.net/forum?id=IP2etbIEuC>.
- 3 Shreya Havaldar\*, **Weiqli You\***, Lyle Ungar, and Eric Wong (2023). “Visual Topics via Visual Vocabularies”. In: *XAI in Action: Past, Present, and Future Applications*. 🔗 URL: <https://openreview.net/forum?id=h60T5pZrGc>.
- 4 Youngja Park and **Weiqli You** (Dec. 2023). “A Pretrained Language Model for Cyber Threat Intelligence”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Singapore: Association for Computational Linguistics, pp. 113–122. 🔗 DOI: [10.18653/v1/2023.emnlp-industry.12](https://doi.org/10.18653/v1/2023.emnlp-industry.12).
- 5 Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, **Weiqli You**, Manni Arora, and Chris Callison-Burch (May 2023). “Causal Reasoning of Entities and Events in Procedural Texts”. In: *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 415–431. 🔗 DOI: [10.18653/v1/2023.findings-eacl.31](https://doi.org/10.18653/v1/2023.findings-eacl.31).

- 6 Artemis Panagopoulou, Manni Arora, Li Zhang, Dimitri Cugini, **Weiqiu You**, Yue Yang, Liyang Zhou, Yuxuan Wang, Zhaoyi Hou, Alyssa Hwang, Lara Martin, Sherry Shi, Chris Callison-Burch, and Mark Yatskar (2022). “QuakerBot: A household dialog system powered by large language models”. In: *Alexa Prize TaskBot Challenge 1 Proceedings*. [URL: https://www.amazon.science/alexaprize/proceedings/quakerbot-a-household-dialog-system-powered-by-large-language-models](https://www.amazon.science/alexaprize/proceedings/quakerbot-a-household-dialog-system-powered-by-large-language-models).
- 7 Thamme Gowda, **Weiqiu You**, Constantine Lignos, and Jonathan May (June 2021). “Macro-Average: Rare Types Are Important Too”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 1138–1157. [DOI: 10.18653/v1/2021.naacl-main.90](https://doi.org/10.18653/v1/2021.naacl-main.90).
- 8 **Weiqiu You\***, Simeng Sun\*, and Mohit Iyyer (July 2020). “Hard-Coded Gaussian Attention for Neural Machine Translation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7689–7700. [DOI: 10.18653/v1/2020.acl-main.687](https://doi.org/10.18653/v1/2020.acl-main.687).

## Preprints

- 1 Helen Jin\*, Anton Xue\*, **Weiqiu You**, Surbhi Goel, and Eric Wong (2025). *Probabilistic Stability Guarantees for Feature Attributions*. [URL: https://antonxue.github.io/files/papers/jin2025probabilistic.pdf](https://antonxue.github.io/files/papers/jin2025probabilistic.pdf).
- 2 Siqi Zeng, Yifei He, **Weiqiu You**, Yifan Hao, Yao-Hung Hubert Tsai, Makoto Yamada, and Han Zhao (2025). *Efficient Model Editing with Task Vector Bases: A Theoretical Framework and Scalable Approach*. arXiv: 2502.01015 [cs.LG]. [URL: https://arxiv.org/abs/2502.01015](https://arxiv.org/abs/2502.01015).
- 3 Helen Jin, Shreya Havaldar, Chaehyeon Kim, Anton Xue, **Weiqiu You**, Helen Qu, Marco Gatti, Daniel A Hashimoto, Bhuvnesh Jain, Amin Madani, Masao Sako, Lyle Ungar, and Eric Wong (2024). *The FLX Benchmark: Extracting Features Interpretable to eXperts*. arXiv: 2409.13684 [cs.LG]. [URL: https://arxiv.org/abs/2409.13684](https://arxiv.org/abs/2409.13684).
- 4 **Weiqiu You** and Youngja Park (2024). *Cyber-Attack Technique Classification Using Two-Stage Trained Large Language Models*. arXiv: 2411.18755 [cs.LG]. [URL: https://arxiv.org/abs/2411.18755](https://arxiv.org/abs/2411.18755).

## Ongoing Works

- 1 “Claim Verifications with Reasoning Attributions” (2024). Ongoing.
- 2 “Fast Leave-One-Out Feature Attribution” (2024). Ongoing.
- 3 Helen Jin, **Weiqiu You**, and Eric Wong (2024). “Certifiably Robust Evaluation of Feature Attributions via Boolean Influences”. Ongoing.
- 4 “Laparoscopic Cholecystectomy Safe-Guarded with Explanations” (2024). Ongoing.
- 5 “T-FIX: Textual Features Interpretable to eXperts” (2024). Ongoing.

## Teaching Experience

2021	<ul style="list-style-type: none"> <li>► <b>Computational Linguistics</b> UPenn CIS530, Teaching Assistant, Spring 2021, Fall 2021</li> </ul>
Spring 2020	<ul style="list-style-type: none"> <li>► <b>Advanced Natural Language Processing</b> UMass COMPSCI685, Grader</li> </ul>

---

\* Equal contribution.

## Teaching Experience (continued)

---

- Spring 2018    ▶    **Data Structures and Algorithms**  
Gordon CPS222, Teaching Assistant
- Spring 2017    ▶    **Calculus II**  
Gordon MAT122, Teaching Assistant
- Fall 2016      ▶    **Differential Equations**  
Gordon MAT225, Teaching Assistant
- 2016 – 2018   ▶    **Biostatistics**  
Gordon, SPSS Help Session Tutor
- ▶    **Calculus**  
Gordon, Tutor

## Invitations

---

- 2024    ▶    **Panalist**  
Women in CS Panel, Computers and Society class. Gordon College, MA.
- ▶    **Speaker**  
Artificial Intelligence Week Alumni Forum. High School Affiliated to Renmin University of China, Beijing, China.
- 2022    ▶    **Panalist**  
Women in CS Panel, Computers and Society class. Gordon College, MA.

## Awards

---

- 2024    ▶    **AWS-AI ASSET Fellow.**
- 2018    ▶

## Academic Services

---

- 2024    ▶    **ICLR.**  
Reviewer.
- 2022 – 2023   ▶    **ACL Rolling Review.**  
Reviewer.
- 2023    ▶    **ACL.**  
Reviewer.
- 2022    ▶    **CLunch, a weekly NLP research seminar run by PennNLP.**  
Organizer
- 2021 – 2023   ▶    **EMNLP.**  
Reviewer.