Weiqiu You

✉ ywq333@gmail.com 🌐 http://weiqiuyou.com/ ☎ +1 978-778-0875

Research Interests

I develop machine learning systems whose explanations are explicitly aligned with the way human experts reason in high-stakes domains such as science and medicine. My work focuses on using structured knowledge, including feature groups, validated premises, and expert-defined safety or scientific criteria, to build models that explain not just what they predict but why they predict it. I design self-attributing architectures that decompose predictions into expert-meaningful components, probabilistic frameworks for verifying the soundness of step-wise reasoning chains, and evaluation benchmarks that measure whether model explanations follow expert reasoning patterns in domains like cosmology and surgery. Ultimately, I aim to create AI systems whose explanations are verifiably faithful, logically sound, and aligned with expert knowledge, enabling safe, reliable, and interpretable deployment in real-world decision-making environments.

Education

2020 – 2026 **Ph.D. Computer and Information Science, University of Pennsylvania**,
Philadelphia, PA GPA: 3.97/4.00
Advisor: Eric Wong
Expected Graduation: May. 2026

2018 – 2020 **M.S. Computer Science, University of Massachusetts Amherst**,
Amherst, MA GPA: 3.90/4.00
Advisor: Mohit Iyyer

2014 – 2018 **B.S. Computer Science and Mathematics, Gordon College**,
Wenham, MA GPA: 3.87/4.00 summa cum laude
Advisor: Jonathan Senning, Russell Bjork
Double major. Honors Thesis title: *Predict Media Interestingness*.

Professional Experience

May-Aug 2025 **Meta (Bellevue, WA)** *Software Engineering Machine Learning Intern*
Working on creating an internal and external benchmarks for evaluating LLM agents' ability in assisting ML engineers in the Ads ML lifecycle. (Continuing collaboration)

May-Aug 2024 **Okinawa Institute of Science and Technology (Okinawa, Japan)** *Visiting Research Student*
Worked on theory in developing faster feature attribution methods that correlate with leave-one-out.

May-Aug 2022 **IBM Research (Yorktown Heights, NY)** *Research Intern*
Worked on developing a two-stage training pipeline to augment cyber threat intelligence attack models with auxiliary data.

May-Aug 2020 **University of Southern California, ISI (Los Angeles, CA | Remote)** *Research Assistant*
Worked on analyzing supervised and unsupervised neural machine translation.

Jun-Aug 2018 **Meituan-Dianping Inc, NLP Center (Beijing, China)** *Research Intern*
Worked on keyword extraction in delivery data.

Publications

Preprints *(* indicates equal contribution)*

1. Shreya Havaldar*, Helen Jin*, Chaehyeon Kim*, Anton Xue*, **Weiqiu You***, Marco Gatti, Bhuvnesh Jain, Helen Qu, Daniel A Hashimoto, Amin Madani, Rajat Deo, Sameed Ahmed M. Khatana, Gary E. Weissman, Lyle Ungar, and Eric Wong (2025). *T-FIX: Text-Based Explanations with Features Interpretable to eXperts*. arXiv: 2511.04070 [cs.CL]. 🔗 URL: https://arxiv.org/abs/2511.04070.

2. Delip Rao*, **Weiqiu You***, Eric Wong, and Chris Callison-Burch (2025). *NSF-SciFy: Mining the NSF Awards Database for Scientific Claims*. arXiv: 2503.08600 [cs.CL]. 🔗 URL: https://arxiv.org/abs/2503.08600.

3. **Weiqiu You**, Cassandra Goldberg, Amin Madani, Daniel A Hashimoto, and Eric wong (2026). *Sum-of-Cues: Structured Reasoning for Surgical Safety with Large Vision-Language Models.*

Selected Publications *(* indicates equal contribution)*

1. **Weiqiu You**, Anton Xue, Shreya Havaldar, Delip Rao, Helen Jin, Chris Callison-Burch, and Eric Wong (2025). “Probabilistic Soundness Guarantees in LLM Reasoning Chains”. In: *The 2025 Conference on Empirical Methods in Natural Language Processing.* 🔗 URL: https://openreview.net/forum?id=877mycbg81.

2. **Weiqiu You**, Siqi Zeng, Yao-Hung Hubert Tsai, Makoto Yamada, and Han Zhao (2025). “When LRP Diverges from Leave-One-Out in Transformers”. In: *The 8th BlackboxNLP Workshop.* 🔗 URL: https://openreview.net/forum?id=tyKwMl4nOT.

3. Helen Jin*, Anton Xue*, **Weiqiu You**, Surbhi Goel, and Eric Wong (2025). “Probabilistic Stability Guarantees for Feature Attributions”. In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems.* 🔗 URL: https://openreview.net/forum?id=pXoR0Sy4WQ.

4. **Weiqiu 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.

5. 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 (2025). “The FIX Benchmark: Extracting Features Interpretable to eXperts”. In: *Journal of Data-centric Machine Learning Research (DMLR).* 🔗 URL: https://openreview.net/forum?id=BJnusBahD3.

6. Chaehyeon Kim, **Weiqiu 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.

7. **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.

Teaching Experience

2021 ▸ **Computational Linguistics**
UPenn CIS530, Teaching Assistant, Spring 2021, Fall 2021

Spring 2020 ▸ **Advanced Natural Language Processing**
UMass COMPSCI685, Grader

Spring 2018 ▸ **Data Structures and Algorithms**
Gordon CPS222, Teaching Assistant

Teaching Experience (continued)

- 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

Invited Talks

- 2025 — **Speaker** *"Explaining and Verifying: Towards Trustworthy Machine Learning"*
  Machine Learning Seminar, University of Illinois Urbana-Champaign, Urbana, IL.
- 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 — **Gordon College Honors Thesis**.
- **Summa Cum Laude**.

Academic Services

- 2025 — **ICLR**.
  Reviewer.
- **ACL Rolling Review**.
  Reviewer.
- **ICML**.
  Reviewer.
- 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.