

EDUCATION

Massachusetts Institute of Technology

2026–2030 (*expected*)

Ph.D. student in Computer Science

Research areas: Interpretability, Robust ML, Science of AI

Advisor: Marzyeh Ghassemi

Brown University

2024–2026

M.S. in Computer Science; GPA: 4.0/4.0

Mentors: Ellie Pavlick and Carsten Eickhoff

Wellesley College

2020–2024

B.A. in Computer Science and History; GPA: 3.94/4.0

Summa cum laude

PUBLICATIONS AND PAPERS

- [1] **Jennifer Meng Lu**, Ruochen Zhang, Isabelle Lee, David Alvarez-Melis, Ellie Pavlick, and Naomi Saphra. **Adversarial Concept Search: Predicting Compositional Errors from Feature Geometry.** *ICML Mechanistic Interpretability Workshop*.
- [2] **Meng Lu***, Catherine Chen*, and Carsten Eickhoff. (2025). **Pathway to Relevance: How Cross-Encoders Implement a Semantic Variant of BM25.** *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- [3] **Meng Lu***, Ruochen Zhang*, Carsten Eickhoff, and Ellie Pavlick. (2025). **Paths Not Taken: Understanding and Mending the Multilingual Factual Recall Pipeline.** *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- [4] **Meng Lu**, Catherine Arnett, and Ruochen Zhang. (2025). **Mechanisms of In-Context Syntactic Generalization in Language Models.** *BlackboxNLP 2025, extended abstract*.
- [5] **Meng Lu***, Abinitha Gourabathina*, Ruochen Zhang, Naomi Saphra, and Marzyeh Ghassemi. **Can Latent Manifold Geometry Guide the Generation of Challenging Compositional Examples?** *Work in progress*.
- [6] Etha Hua, **Meng Lu**, Zhuonan Yang, Michael Lepori, and Ellie Pavlick. **How Training Choices Shape Representation Manifolds.** *Work in progress*.

AWARDS AND HONORS

MIT Presidential Fellowship

2026

Summa cum laude, Wellesley College

2024

NSF-REU Fellowship

2023

INVITED TALKS

VSAONLINE, Hyperdimensional Computing community (IEEE Task Force)

Dec. 2026 (upcoming)

Google Multilingual Group

Sept. 2025

Cross-Institution Mechanistic Interpretability Reading Group

June 2026

(Brown University, Northeastern University, and University of Maryland)

RESEARCH EXPERIENCE

Massachusetts Institute of Technology

2026–Present

Graduate Researcher, EECS / LIDS | Advisor: Marzyeh Ghassemi

- **Representation geometry and dynamic monitoring:** Study how concept entanglement and latent manifold structure contribute to generalization failures. Use geometric diagnostics at training checkpoints to generate synthetic examples targeting predicted model weaknesses and guide continual learning [5].
- **Self-improving models and robust continual learning:** Develop adaptive training strategies that target models’ evolving weaknesses with synthetic examples to improve out-of-distribution robustness, transfer, and generalization while accelerating convergence.

Harvard Kempner Institute

2025–2026

Research Collaboration | Collaborators: Naomi Saphra and David Alvarez-Melis

Developed Adversarial Concept Search, a framework that predicts compositional failures from interference among latent concept representations without evaluating the composed inputs. Demonstrated its predictive power on multilingual factual recall using only English fact representations and language subspace geometry, reducing the need for exhaustive multilingual evaluations [1].

Brown University

July 2024–Present

Research Assistant, LUNAR Lab / HealthNLP Lab | Advisors: Ellie Pavlick and Carsten Eickhoff

- **Multilingual factual recall:** Investigated why models fail to reliably express internally represented knowledge across languages; developed mechanistically informed interventions that improved cross-lingual factual recall (EMNLP 2025) [3].
- **Circuits in ranking models:** First to uncover a complete ranking circuit in neural models, revealing how it implements a semantic variant of BM25 (EMNLP 2025) [2].
- **Syntactic representations:** Examine structural priming and identify causal syntactic representations underlying it, demonstrating how mechanistic analysis can inform psycholinguistic hypotheses about language production [4].

NSF-REU at Brown University

June–Aug. 2023

Research Assistant, Machine Learning, Computer Vision | Advisor: James Tompkin

Developed an ENeRF-based VR teleoperation pipeline, calibrated Spot robot cameras, and improved mixed-reality robot-arm control; presented at the Leadership Alliance National Conference and Tanner Conference.

MIT UROP

2022–2024

Undergraduate Researcher | Advisors: Cynthia Breazeal, Kamal Youcef-Toumi, Fotini Christia

- **Personal Robotics Group** Sept. 2023–Aug. 2024
Developed a multimodal pipeline for parent–child affect analysis; fine-tuned YOLOv5 and DeepFace and incorporated memory augmentation to identify parents and children and estimate valence and attention.
- **Mechatronics Research Laboratory** Sept. 2023–Jan. 2024
Refined augmented-reality displays of robot trajectories and human motion; evaluated efficiency and safety in human–robot assembly tasks.
- **Institute for Data, Systems, and Society (IDSS)** Aug. 2022–June 2023
Created digital twins and virtual-reality archives of seven Afghan heritage sites, developed an augmented-reality game for displaced Afghan children, and presented at the MIT Media Lab’s “Ways of Seeing” conference.