SHERIDAN FEUCHT

Computer Science PhD Candidate at Northeastern University
linkedin.com/in/sheridan-feucht ⋄ sfeucht.github.io

EDUCATION

PhD in Computer Science, Northeastern University, GPA: 4.0 Expected Graduation 2028

Selected Coursework: AI as Archival Science, Foundations and Applications of Information Theory, Seminar in Human-Computer Interaction

Bachelor of Science in Computer Science, Brown University, GPA: 4.0 Graduated May 2023

Selected Coursework: Logic in Language and Mind, Language Processing in Humans and Machines, Syntax, Psycholinguistics, Pattern Theory, Information Theory, Systems (C, x86 Assembly)

RESEARCH EXPERIENCE

Goodfire - Research Fellow Jan 2026 - April 2026

- Working full-time with Atticus Geiger, focusing on project that connects causal abstraction to representation theory and feature geometry.

Northeastern University Bau Lab - PhD Candidate Sep 2023 - Present

- Published two first-author papers on multi-token semantic representations in LLMs (**Publications 1 and 4**), along with one NeurIPS Mechanistic Interpretability workshop paper (**Publication 2**).
- Advised student Adrian Chang for a NeurIPS Mechanistic Interpretability workshop paper on abstract letter representations in text-to-image diffusion models (**Publication 3**).
- Advised master's student Kerem Sahin on independent project on training dynamics of induction heads (under review).
- Assisted labmate Rohit Gandikota in writing paper on concept erasure in language models (**Publication 5**).
- Currently working on project on diffusion models, as well as advising on project on implicit grammatical constructions in LLMs.

Brown University LUNAR Lab - Undergraduate Researcher Feb 2022 - September 2023

- Evaluating CNN, ResNet, and ViT performance classifying same-different relations between two shapes in an image, determining whether model can generalize to unseen shapes (See **Publication 6**).

Brown University Health-NLP Lab - Undergraduate Researcher Feb - Dec 2021

- Used LDA topic modeling to sample news articles from CNN/DailyMail; collected and validated human annotations on Amazon MTurk for 3000 sampled news articles; fine-tuned language models on corpus of newly-collected data. Presented this work to Brown AI and ML Labs and as a Findings paper at ACL 2022 (see **Publication 7**)

Brown University Sloman Lab - Research Assistant Mar 2020 - Jul 2021

- Developed a new manual for syntactic and discourse-level annotation of documents. Annotated human- and computer-generated documents to create a corpus of online discourse (see **Publication 8**)

PUBLICATIONS

1. **Sheridan Feucht**, Eric Todd, Byron Wallace, and David Bau. The Dual-Route Model of Induction. *2nd Conference on Language Modeling (COLM)*. 2025. https://dualroute.baulab.info
2. **Sheridan Feucht**, Byron Wallace, and David Bau. Vector Arithmetic in Concept and Token Subspaces. *2nd Mechanistic Interpretability Workshop at NeurIPS*. 2025. https://arithmetic.baulab.info
3. Adrian Chang*, **Sheridan Feucht***, Byron Wallace, and David Bau. Does FLUX Know What It's Writing? *2nd Mechanistic Interpretability Workshop at NeurIPS*. 2025. (*Equal contribution.)

4. **Sheridan Feucht**, David Atkinson, Byron Wallace, and David Bau. Token Erasure as a Footprint of Implicit Vocabulary Items in LLMs. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2024. https://footprints.baulab.info

5. Rohit Gandikota, **Sheridan Feucht**, Samuel Marks, and David Bau. Erasing Conceptual Knowledge from Language Models. *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*. 2025.

6. Alexa R. Tartaglini*, **Sheridan Feucht***, Michael A. Lepori, Wai Keen Vong, Charles Lovering, Brenden M. Lake, and Ellie Pavlick. Deep Neural Networks Can Learn Generalizable Same-Different Visual Relations. *8th Annual Conference on Cognitive Computational Neuroscience*. 2025. (*Equal contribution.)

7. Seyed Ali Bahrainian*, **Sheridan Feucht***, and Carsten Eickhoff. NEWTS: A Corpus for News Topic-Focused Summarization. *Findings of the Association for Computational Linguistics*. 2022. (*Equal contribution.)

8. Babak Hemmatian, **Sheridan Feucht**, Rachel Avram, Alexander Wey, Muskaan Garg, Kate Spitalnic, Carsten Eickhoff, Ellie Pavlick, Bjorn Sandstede, Steven Sloman. A Novel Corpus of Discourse Structure in Humans and Computers. *The 2nd Workshop on Computational Approaches to Discourse at EMNLP*. 2021.

PRESENTATIONS

1. "LLMs Represent Words, Not Just Tokens." *UT Austin Linguistics*, Austin, TX. September 2025. (Virtual, Invited Talk)

2. "LLMs Represent Words, Not Just Tokens." *Brown ANCOR Talks*, Providence, RI. October 2025.

3. "Equifinality: There's Often Multiple Explanations." *Mechanistic Interpretability Workshop at NeurIPS*, San Diego, CA. December 2025. (Invited).

4. "Concept Heads for Vector Arithmetic." *Simplex AI Safety*, Emeryville, CA. January 2026. (Invited).

WORK EXPERIENCE

Northeastern University - Graduate Teaching Assistant (Interpretability) Sep - Dec 2024

- Developed course "Structure and Interpretation of Deep Networks" alongside Prof. David Bau. Led in-class discussion on papers in AI interpretability. Wrote material and tutorials on course website: https://sidn.baulab.info/

Brown University - Undergraduate Teaching Assistant (Ghanaian Drumming) Jan - May 2022

- Assisted Prof. Kwaku Kwaakye (Martin) Obeng in weekly classes; led weekly rehearsals, teaching students Ghanaian drumming, dancing, and singing.

Brown University - Undergraduate Teaching Assistant (Computational Linguistics) Sep - Dec 2022

- Developed assignment on machine translation, making students put together their own Transformer model. Assisted in developing assignments on topic modeling, BERT finetuning, and dependency parsing. Held office hours to explain NLP concepts to students and help them debug their assignments.

Brown University - Undergraduate Teaching Assistant (Introduction to Music Theory) Sep - Dec 2022

- Assisted lecture section for MUSC 0400A (Introduction to Music Theory) with Professor Andrew Welch. Held office hours to help students with sight-singing, musical notation, and composition.

Brown University - Undergraduate Teaching Assistant (Intro. to Computer Systems) Sep - Dec 2021

- Held conceptual office hours to answer student questions on course concepts (e.g. procedure calls and stack frames, memory/heap management); held code-based office hours (C, x86 Assembly).

Shaw Communications - Data Strategy Summer Student May - Aug 2020

- Queried, validated, and investigated company data to increase understanding of consumer behavior; wrote stored procedures and views in SQL to help implement foundational pipelines.