Eve Fleisig

efleisig@berkeley.edu • linkedin.com/in/eve-fleisig • github.com/efleisig • Twitter @enfleisig

Education

University of California, Berkeley 2021-2026

PhD student in computer science, advised by Rediet Abebe and Dan Klein

Courses: Natural Language Processing, Statistical Learning Theory, Computer Vision & NLP, Deep Learning, Pragmatics, Sociolinguistics

Princeton University 2018-2021

Bachelor of Science in Engineering in Computer Science (summa cum laude), minor in Linguistics GPA: 3.96

Graduate courses: Deep Learning for Natural Language Processing, Limits to Prediction

Machine Translation, Theory of Algorithms, Theory of Computation, Syntax, Phonetics & Phonology

Mathematics: Linear algebra, multivariable calculus, real analysis, graph theory

Languages: Fluent in Spanish (bilingual), highly proficient in French and Portuguese, proficient in Italian

Programming experience: Python, C++, Java, C; PyTorch and TensorFlow

Research

FairPrism: Evaluating fairness-related harms in text generation 2023

We introduced a dataset and methodology for fine-grained measurement of harms in text generation.

Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann, Hanna Wallach. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023).* [ACL 2023 paper] [dataset]

When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks 2023

To refine hate speech detection, we modeled individual annotators to predict the opinions of the groups being targeted by hate speech.

Eve Fleisig, Rediet Abebe, Dan Klein [arXiv]

Centering the Margins: Outlier-Based Identification of Harmed Populations in Toxicity Detection 2023

We introduced a method of operationalizing marginalized groups through outlier analysis to identify harmed populations.

Vyoma Raman*, Eve Fleisig*, Dan Klein [arXiv]

Ghostbuster: Detecting Text Ghostwritten by Large Language Models 2023

We developed a model that detects AI-generated text that outperforms previous approaches and has average 99.1 F1 across all datasets tested.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, Dan Klein [arXiv]

Mitigating Gender Bias in Machine Translation through Adversarial Learning 2020-2021

Developed an adversarial neural network that mitigates machine translation gender bias in seq2seq translation.

Eve Fleisig and Christiane Fellbaum [arXiv]

Bilingual Lexical Access and Cognate Idiom Comprehension 2020

Investigated the effects of figurative language transfer on bilingual lexical processing.

Eve Fleisig. *Proceedings of the Workshop on Cognitive Aspects of the Lexicon* (CogALex-VI), *COLING*, 2020.

Independent Research in Deep Learning for Natural Language Processing 2019-2021

Advised by Christiane Fellbaum

Cognate identification through transfer learning from a character-level convolutional neural network 2020

Automatically identifying semantic shift using unsupervised learning 2019

Sentiment Analysis for Reinforcement Learning 2020-2021

We optimized reinforcement learning rewards for text-based games with BERT-based sentiment analysis, tackling the problem of sparse rewards and potentially permitting reinforcement learning without rewards.

Eve Fleisig* and Ameet Deshpande*. [arXiv]

VEMOS: A Visual Explorer for Similarity Metrics on Complex Data Sets 2020

Eve Fleisig and Gunay Dogan. NIST Technical Report (2020).

Work Experience

Research Intern, Microsoft Research Summer 2022
Led FairPrism project, a dataset and methodology for measuring harms in text generation.

Software Engineering Intern, Google Summer 2021
Contributed to natural language processing research for new product development.

Software Engineering Intern, Duolingo Summer 2020
Contributed to machine learning research on personalized learning through adjustments to Duolingo's BirdBrain model.

Research Assistant, National Institute of Standards and Technology (NIST) 2015-2019
Created VEMOS, a Python user interface to assess fairness and reliability of computer vision models.

Teaching and Service

Teaching Assistant, CS 189/289 - Introduction to Machine Learning (UC Berkeley) 2023
Taught section, ran office hours, designed exam problems, and assisted with graduate student projects

Teaching Assistant, Independent Work Seminar in Natural Language Processing (Princeton University) 2020
Assisted students with approaches to natural language processing research

Student Chair, Association for Computational Linguistics (ACL) 2024

Guest Lectures and Invited Talks

CS 288 – Natural Language Processing (UC Berkeley): "Misuse, Risks, and Harms of NLP" 2023
CS 294 – Vision and Language (UC Berkeley): "Ethical Concerns of Large-Scale Models" 2023
CS 288 – Natural Language Processing (UC Berkeley): "Ethics of NLP" 2022
Natl. Institute of Standards and Technology: "VEMOS: A Visual Explorer for Similarity Metrics on Complex Datasets" 2019

Undergraduate Mentorship

Olivia Huang, UC Berkeley undergraduate 2021-2023
Harbani Jaggi, UC Berkeley undergraduate 2022
Kashyap Murali, UC Berkeley undergraduate 2022
Mahathi Ryali, UC Berkeley undergraduate 2022
Vyoma Raman, UC Berkeley undergraduate, now at Stanford 2022-2023
Zaina Shaik, UC Berkeley undergraduate 2023-
Vivek Verma, UC Berkeley undergraduate 2023-
Xavier Yin, UC Berkeley undergraduate 2022-
Berkeley AI Research: Underrepresented Undergraduates Mentor 2023

Selected Awards and Activities

NSF Graduate Research Fellowship Awardee 2022
Outstanding Senior Thesis Award, Princeton Computer Science 2021
Sigma Xi Book Award for Outstanding Undergraduate Research 2021
Elected to Phi Beta Kappa Honors Society and Tau Beta Pi Engineering Honors Society 2021
Outstanding Undergraduate Researcher Award honorable mention, Computing Research Association 2020
Distinguished Hispanic Scholar, Hispanic Alliance for Education 2018
Founder and President, Princeton Computational Linguistics Society 2019-2021