Eve Fleisig

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

Education

University of California, Berkeley 2021-2026
PhD, Computer Science
Advisor: Dan Klein, Berkeley Natural Language Processing

Princeton University 2018-2021
Bachelor of Science in Engineering in Computer Science (summa cum laude), minor in Linguistics
Advisor: Christiane Fellbaum

Languages: Fluent in Spanish (bilingual); fluent in French; proficient in Portuguese and Italian

Awards and Honors

Outstanding Paper Award, NAACL 2025 2025
Outstanding Paper Award, EMNLP 2023 2023
2nd place, LSA Summer Institute Posters 2025
2nd place, LSA Summer Institute Three-Minute Thesis 2025
Outstanding Graduate Student Instructor Award, UC Berkeley 2024
Berkeley ICBS Interdisciplinary Grant 2024
NSF Graduate Research Fellowship Award 2022
Chancellor's Fellowship, UC Berkeley 2021
Outstanding Senior Thesis Award, Princeton Computer Science 2021
Sigma Xi Book Award for Outstanding Undergraduate Research 2021
Outstanding Undergraduate Researcher Award honorable mention, Computing Research Association 2020
Distinguished Scholar, Hispanic Alliance for Education 2018

Representative Papers

When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks
Eve Fleisig, Rediet Abebe, Dan Klein (EMNLP 2023).

Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination
Eve Fleisig*, Genevieve Smith*, Madeline Bossi*, Ishita Rustagi*, Xavier Yin*, Dan Klein (EMNLP 2024).

Incorporating Worker Perspectives into MTurk Annotation Practices for NLP
Olivia Huang, Eve Fleisig, Dan Klein (EMNLP 2023 - **Outstanding Paper Award**).

Additional Papers

GRACE: A Granular Benchmark for Evaluating Model Calibration against Human Calibration
Yoo Yeon Sung*, Eve Fleisig*, Yu Hou, Ishan Upadhyay, Jordan Boyd-Graber (ACL 2025 - **Oral**).

Is your benchmark truly adversarial? AdvScore: Evaluating Human-Grounded Adversarialness
Yoo Yeon Sung, Maharshi Gor, Eve Fleisig, Ishani Mondal, Jordan Boyd-Graber (NAACL 2025 - **Outstanding Paper Award**).

Mapping Social Choice Theory to RLHF
Jessica Dai, Eve Fleisig (R2FM workshop at ICLR 2024).

Ghostbuster: Detecting Text Ghostwritten by Large Language Models
Vivek Verma, Eve Fleisig, Nicholas Tomlin, Dan Klein (NAACL 2024).

The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels
Eve Fleisig, Su Lin Blodgett, Dan Klein, Zeerak Talat (NAACL 2024).

First Tragedy, then Parse: History Repeats Itself in the New Era of Large Language Models
Naomi Saphra, Eve Fleisig, Kyunghyun Cho, Adam Lopez (NAACL 2024).

Accurate and Data-Efficient Toxicity Prediction when Annotators Disagree
Harbani Jaggi, Kashyap Murali, Eve Fleisig, Erdem Bıyık (EMNLP 2024).

Hedges and apologies in ChatGPT responses to African-American English
Eve Fleisig (NWAV51, 2023).

FairPrism: Evaluating fairness-related harms in text generation
Eve Fleisig**,** Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann, Hanna Wallach (ACL 2023).

Centering the Margins: Outlier-Based Identification of Harmed Populations in Toxicity Detection
Vyoma Raman*, Eve Fleisig*, Dan Klein (EMNLP 2023).

Hedges and Apologies in ChatGPT Responses to African-American English
Eve Fleisig (NWAV 2023).

Mitigating Gender Bias in Machine Translation through Adversarial Learning
Eve Fleisig and Christiane Fellbaum (arXiv, 2020).

Bilingual Lexical Access and Cognate Idiom Comprehension
Eve Fleisig. *Proceedings of the Workshop on Cognitive Aspects of the Lexicon* (CogALex-VI workshop at COLING 2020).

VEMOS: A Visual Explorer for Similarity Metrics on Complex Data Sets
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.
Advisor: Hanna Wallach

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 by modifying 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

Instructor, CS 188 - Introduction to Artificial Intelligence (UC Berkeley) Summer 2024
Lectured to 300+ person class, designed course content, organized exams, managed 15+ course staff with co-instructor, and held instructor office hours.

Teaching Assistant, CS 189/289 - Introduction to Machine Learning (UC Berkeley) Spring 2023
Taught weekly course sections, designed exam content, and held weekly office hours.

Teaching Assistant, Independent Work Seminar in Natural Language Processing (Princeton) Fall 2020
As primary TA, guided students on independent research and established weekly course tutorials in NLP.

Service

Mentor, SUPERB/BAIR-HBCU REU Summer 2024
Mentored student one-on-one for research experience for HBCU undergraduates.

Student Research Workshop Chair, Association for Computational Linguistics (ACL) 2024
Organized review of 200+ workshop submissions, meta-reviewed 50+ papers, and co-organized workshop program.

Berkeley PhD EECS Admissions 2023
Reviewed applications to Berkeley EECS for the Berkeley NLP Group.

Berkeley Student Committee for Faculty Hiring 2022-2024
Participated in student interviews for EECS faculty candidates.
Underrepresented Undergraduates Mentor, Berkeley AI Research 2023
Guided student on approaches to NLP research.

Reviewing

Area Chair, ACL Rolling Review 2025
ACL Rolling Review 2021-2025
FAccT 2024
Program Chair, ACL Student Research Workshop 2024
NeurIPS 2024 Workshop EvalEval 2024
NeurIPS 2023 Workshop ATTRIB 2023

Invited Talks

Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination
Berkeley Responsible AI Workshop 2025
Berkeley Language & AI Forum 2025
Berkeley Sociolinguistics 2024
Stanford NLP 2024

When the Majority is Wrong: Modeling Annotator Disagreement for Language Tasks
Università Bocconi (MilaNLP) 2024
Columbia NLP 2023
DIMACS Workshop on Foundation Models, Large Language Models, and Game Theory 2023

Others
Berkeley Research, Teaching, and Learning Center: "GSI Perspectives on Generative AI." Panelist. 2024
Berkeley CS288: The Future of Natural Language Processing. Panelist. 2022
Natl. Institute of Standards and Technology: "VEMOS: A Visual Explorer for Similarity Metrics on Complex Datasets." 2019

Guest Lectures

CS 288 – Natural Language Processing: "Misuse, Risks, and Harms of NLP" 2023
CS 294 – Vision and Language: "Ethical Concerns of Large-Scale Models" 2023
LIN 255 – Advanced Sociolinguistics: "Linguistic Bias in ChatGPT" 2025
UGBA 192 – Responsible AI Innovation & Mgmt: "Addressing Equity & Fairness Issues in LLMs" 2024, 2025
EWMBA295 – AI for Business Leaders: "Addressing Equity & Fairness Issues in LLMs" 2025
CS 10 – The Beauty and Joy of Computing: "Introduction to Generative AI" 2024
CS 288 – Natural Language Processing: "Ethics of NLP" 2022

Advising

Vyoma Raman, UC Berkeley BS (→ Stanford MS, Cornell Tech PhD) 2022-2025
Kayla Lee, UC Berkeley BS/MS (→ YC startup founder) 2023-2025
Harbani Jaggi, UC Berkeley BS 2022-2025
Samuel Ghezae, Howard University BS 2024
Vivek Verma, UC Berkeley BS/MS (→ OpenAI) 2023-2024
Xavier Yin, UC Berkeley BS (→ CMU PhD) 2022-2024
Olivia Huang, UC Berkeley BS 2021-2023
Kashyap Murali, UC Berkeley BS (→ Anthropic) 2022-2023
Zaina Shaik, UC Berkeley BS 2023
Mahathi Ryali, UC Berkeley BS 2022