Lucy Li

Natural language processing (NLP), computational social science, digital humanities, AI fairness

✉ lucy3_li@berkeley.edu | ⌂ lucy3.github.io

Education

University of California, Berkeley   *Berkeley, CA*

PhD Information Science   *Aug 2019 - present*

- Advisor: David Bamman
- Committee members: Isaac Bleaman, Niloufar Salehi, Dan Jurafsky
- Berkeley AI Research (BAIR), Stanford SPARQ Research Affiliate

Stanford University   *Stanford, CA*

MS Computer Science, BS Symbolic Systems   *Sept 2014 - June 2019*

- Coterminal degrees, w/ a concentration in language and depth in artificial intelligence.

Awards, Fellowships, & Grants

Best Resource Paper, for Dolma, Association for Computational Linguistics (ACL)	*2024*
Learning Engineering Tools Competition Catalyst Award, $50k, The Learning Agency	*2024*
Rising Star in Data Science, University of Chicago & University of San Diego	*2023*
Rising Star in EECS, Rising Stars Academic Career Workshop	*2023*
Meta Research PhD Fellowship Finalist, Meta	*2023*
AI2 Outstanding Intern of the Year Award, Allen Institute for Artificial Intelligence	*2022*
Human-Centered Artificial Intelligence Seed Grant, Stanford HAI (PI: Patricia Bromley)	*2021*
Google PhD Fellowship Nominee, one of four students selected by UC Berkeley	*2021*
Best Paper Award, Conference on Educational Data Science	*2020*
Graduate Research Fellowship, National Science Foundation	*2019*
Top 5% of Course Assistants, Stanford Computer Science	*2019*
K. Jon Barwise Award for Distinguished Contributions, Stanford Symbolic Systems	*2018*
Undergraduate Advising & Research (UAR) Small Grant, $1500, Stanford University	*2018*
Grants for Education and Research, $1145, Stanford Symbolic Systems	*2017*
Phi Beta Kappa, Stanford University	*2017*

Experience

Allen Institute for Artificial Intelligence   *Seattle, WA*

Research Intern   *June 2023 - Present*

- Mentors: Jesse Dodge (June 2023 - Dec 2023), Kyle Lo (Jan 2024-Present)
- Summer / Fall 2023: Analyzed large language models' data curation practices on the AllenNLP team.
- Spring / Summer 2024: Led new AI for education research initiatives at AI2, funded by the Bill & Melinda Gates Foundation and the Learning Agency.

Allen Institute for Artificial Intelligence   *Seattle, WA*

Research Intern   *May 2022 - Dec 2022*

- Mentors: Katie Keith (Semantic Scholar team), Jesse Dodge (AllenNLP team)
- Mapped discipline-specific language across scientific domains.

Microsoft Research   *Montreal, Canada*

Research Intern   *May 2021 - Aug 2021*

- Mentors: Alexandra Olteanu, Su Lin Blodgett
- Evaluated natural language generation systems on the Fairness, Accountability, Transparency, and Ethics (FATE) team.

Stanford Computer Science   *Stanford, CA*

Research Assistant   *Jan 2019 – Dec 2019*

- Advisors: Dan Jurafsky (Stanford NLP), Patricia Bromley (School of Education)
- Investigated the framing and representation of underrepresented groups in history textbooks with linguistics PhD student Dora Demszky (later Stanford Assistant Professor).

École Polytechnique Fédérale de Lausanne   *Lausanne, Switzerland*

Research Intern   *July 2018 – Sept 2018*

- Advisor: Robert West (Data Science Lab)
- Operationalized and analyzed behavioral trends in a political quote dataset.

Stanford Computer Science   *Stanford, CA*

Research Assistant   *April 2017 – June 2018*

- Advisors: David Jurgens, Jure Leskovec (Stanford Network Analysis Project), Dan Jurafsky (Stanford NLP)
- Leveraged language and social network features to classify fictional and real relationships.

Papers

*indicates equal contribution.

Note on disciplinary audiences: *ACL*, *EMNLP*, *NAACL*, and *TACL* are natural language processing venues, and *AERA Open* and *Educational Researcher* are education journals.

Journals & Conferences

David Bamman, Kent K. Chang, **Li Lucy**, Naitian Zhou. On Classification with Large Language Models in Cultural Analytics. *Computational Humanities Research (CHR) Conference*, 2024.

Li Lucy, Tal August, Rose E. Wang, Luca Soldaini, Courtney Allison, Kyle Lo. Evaluating Language Model Math Reasoning via Grounding in Educational Curricula. *Findings of Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Minju Choi*, **Li Lucy***, Patricia Bromley, David Bamman. "Othering" through War: Depiction of Asians/Asian Americans in U.S. History Textbooks from California and Texas. *Educational Researcher*, forthcoming.

Li Lucy, Suchin Gururangan, Luca Soldaini, Emma Strubell, David Bamman, Lauren Klein, Jesse Dodge. AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters. *Association of Computational Linguistics (ACL)*, 2024.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, **Li Lucy**, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, Kyle Lo. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *Association of Computational Linguistics (ACL)*, 2024. [Best Resource Paper]

Li Lucy, Su Lin Blodgett, Milad Shokouhi, Hanna Wallach, Alexandra Olteanu. "One-Size-Fits-All"? Examining Expectations around What Constitute "Fair" or "Good" NLG System Behaviors. *North American Association for Computational Linguistics (NAACL)*, 2024.

Li Lucy, Jesse Dodge, David Bamman, Katherine A. Keith. Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications. *Findings of the Association of Computational Linguistics (ACL)*, 2023.

Li Lucy, Divya Tadimeti, David Bamman. Discovering Differences in the Representation of People using Contextualized Semantic Axes. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022

Li Lucy, David Bamman. Characterizing English variation across social media communities with BERT. *Transactions of the Association of Computational Linguistics (TACL)*, 2021.

Li Lucy*, Dora Demszky*, Patricia Bromley, Dan Jurafsky. Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks. *AERA Open*, 2020. [Best Paper Award at American Educational Research Association (AERA) Educational Data Science Conference.]

Workshops & Non-Archival Venues

Sami Baral*, **Li Lucy***, Ryan Knight, Alice Ng, Luca Soldaini, Neil Heffernan, Kyle Lo. DrawEduMath: An Expert-Annotated Dataset of Students' Math Images. *NeurIPS Workshop on Mathematical Reasoning and AI*, 2024. [non-archival]

Li Lucy, David Bamman. Gender and Representation Bias in GPT-3 Generated Stories. *Workshop on Narrative Understanding (WNU) at the North American Association for Computational Linguistics (NAACL)*, 2021.

Emma Lurie, **Li Lucy**, Masha Belyi, Sofia Dewar, Daniel Rincón, John Baldwin, Rajvardhan Oak. Investigating Causal Effects of Instructions in Crowdsourced Claim Matching. *Computation + Journalism Symposium (C+J)*, 2020. [non-archival]

Li Lucy, Julia Mendelsohn. Using sentiment induction to understand variation in gendered online communities. *Society for Computation in Linguistics (SCiL)*, 2019.

Li Lucy, Jon Gauthier. Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning. *Language Grounding for Robotics (RoboNLP) Workshop at the Association for Computational Linguistics (ACL)*, 2017.

Blog Posts

Maria Antoniak, **Li Lucy**, Maarten Sap, Luca Soldaini. Using Large Language Models With Care. 2023. [Link.]

David Jurgens, **Li Lucy**. A Look inside the Pedagogy of Natural Language Processing. 2018. [Link.]

Invited Talks

Large-Scale Language Patterns Across Social Groups
 Sept 2024. "Next 10 in AI" talk series, Max Planck Institute for Software Systems.
 Oct 2023. USC NLP Group, University of Southern California.

AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters
 Sept 2024. NLP Group, University of Massachusetts, Amherst.
 July 2024. MilaNLP Seminar Series, Bocconi University.
 March 2024. Computing and Society Seminar, Stanford University.
 Feb 2024. Computational Linguistics group, University of Toronto.

Measuring Depictions and Expressions of Social Groups with NLP
 Jan 2024. Jantina Tammes School of Digital Society, Technology and AI. University of Groningen, Netherlands.
 May 2023. NLP Seminar, Stanford University.

Reader Response to Characters of Color in Literature Taught in U.S. Schools
 Jan 2024. Center for Interdisciplinary Research. Bielefeld University, Germany.

Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications
 Nov 2023. New Directions in Analyzing Text as Data (TADA). Amherst, MA.
 June 2023. Computational Culture Lab, Stanford Graduate School of Business & Berkeley Haas.

"One-size-fits-all"? Observations and Expectations of NLG Systems across Identity-related Language Features.
 Aug 2023. AI, Ethics and Effects in Engineering and Research (Aether), Microsoft.

Context-Dependent Depictions of People Across Three Domains.
 March 2023. "NLP for Social Science: From Language Models to Social Structures" talk series, Columbia University.

What big data and big models bring to the table.
 February 2023. "Roundtable Series: Learning How to Play with the Machines," University of California, Berkeley.

Characterizing English variation across social media communities with BERT.
 Oct 2022. Guest lecture for "Practical Approaches to Data Science with Text" course, Emory University.
 June 2021. Guest lecture for "Computational Text Analysis" course, Barnard College.
 Nov 2021. Guest lecture for "Practical Approaches to Data Science with Text" course, Emory University.

Content Analysis of Textbooks via Natural Language Processing.
 Sept 2022. Narrative and Society Conference, Department of English, McGill University.

Oct 2021. 103rd Anniversary of the School of Information, Berkeley.
Feb 2021. Guest lecture for "Doing Digital History" course, Stanford University.
Feb 2021. Human-Computer Interaction Lunch Seminar, Stanford University.
May 2021. Guest lecture for "Using Data to Describe the World" course, Stanford University.
May 2020. Guest lecture for "Using Data to Describe the World" course, Stanford University.
Oct 2019. New Directions in Analyzing Text as Data (TADA). Stanford, CA.

Teaching and Mentoring Experience

Stanford CS 224U, Natural Language Understanding, Course Assistant (top 5% in CS) *Spring 2019*
Symbolic Systems Program, Advising Fellow *2016 - 2017, 2019*
Stanford EE/CME 103, Introduction to Matrix Methods, Course Assistant *Fall 2017*

Undergraduate advisees: Dani Roytburg (Emory Computer Science & Quantitative Sciences, 2024-Present), Hadley Schachter (UC Berkeley Computer Science & Linguistics, 2024-Present), Claire Ying Wang (UC Berkeley Data Science & Media Studies, 2022-Present), JJ Kim-Ebio (UC Berkeley Data Science, 2022-Present), Iris Wu (Emory Women, Gender, and Sexualities Studies & Quantitative Sciences, 2024), Odelia Larbi-Amoah (Emory Women, Gender, and Sexualities Studies & Quantitative Sciences, 2023-2024), Ethan N. Elasky (UC Berkeley Data Science, 2023-2024), Carolyn Wang (UC Berkeley Computer Science, 2023-2024), Sabrina Baur (UC Berkeley Cognitive Science, 2023 → Deloitte), Tiffany Liang (UC Berkeley Data Science, 2023 → Kohl's), Aryia Dattamajumdar (UC Berkeley Computer Science, 2023 → Meta), Sebastian Orozco (UC Berkeley Ethnic Studies & Political Science, 2022), Divya Tadimeti (UC Berkeley Computer Science, 2021-2022 → Columbia MS), Nikhil Mandava (UC Berkeley Computer Science & Business, 2021 → Glean).

Master's student advisees: Vyoma Raman (Stanford Computer Science, 2023-Present).

Service

Professional

Reviewer: ACL Rolling Review (2021-Present), ACL (2021-Present), EMNLP (2021-Present), Computational Humanities Research (2024), ACL Student Research Workshop (2024), NLP+Computational Social Science Workshop (2024), Workshop on Reference, Framing, and Perspective (2024), FAccT (2024), Journal of Big Data (2024), PLOS One (2024), LREC (2024), NeurIPS (2023-2024), AI & HCI Workshop at ICML (2023), The Web Conference (2023), NLP for Positive Impact Workshop (2022), COLING (2022), NAACL Student Research Workshop (2022), NAACL Workshop on Understanding Implicit and Underspecified Language (2022), SCiL (2022), CHI (2022), CSCW (2022), ACL Workshop on NLP for Positive Impact (2021), AERA Open (2020), NeurIPS Human and Machine in-the-Loop Evaluation and Learning Strategies Workshop (2020).

Organizing committee: Teaching NLP (2021) at NAACL.

Advisory board: AI for Humanists (2020-Present), UC Berkeley Human Rights Center's Assessment of Large Language Models (2023-Present).

Community

General: Diaries of Social Data Research (Podcast Host, 2021-Present), Sociologists of Digital Things (Admin, 2021), NLP+CSS PhD Summer Reading Group (2020).

Students: Berkeley Undergraduate Research Apprentice Program (2021-Present), BAIR Mentoring Program (2020-2021), CS Kickstart (Speaker; 2020), UC Berkeley Girls in Engineering (Leader; 2020), Berkeley AI4ALL (Mentor; 2019), Stanford AI4All (Mentor; 2019), Girls Teaching Girls to Code (Mentor/Lead; 2018, 2019).