Myra Cheng

myra@cs.stanford.edu • myracheng.github.io

Education

Stanford University 2022 – present
PhD Candidate in Computer Science, 4.0/4.0 GPA. Advised by Dan Jurafsky.

California Institute of Technology (Caltech) 2018 – 2022
BS in Computer Science and History, 4.2/4.0 GPA. Advised by Yisong Yue.

Awards

Caltech EAS Trailblazers Symposium	2026
ACL Senior Area Chair Highlights Award	2025
ACL Social Impact Award	2023
Stanford Knight-Hennessy Scholarship	2022
NSF Graduate Research Fellowship	2022
Stanford EDGE Doctoral Fellowship	2022
Caltech Senior Undergraduate Thesis Prize	2022
Caltech Eleanor Searle Prize in Law, Politics & Institutions	2021
Barry M. Goldwater Scholarship	2020

Papers

* denotes co-first authorship.

Warning labels shift perceptions of sycophantic AI, but not its influence.
Under journal review
L Ibrahim*, M Cheng*, C Lee, P Khadpe, D Ong, D Jurafsky, D Yang

Sycophantic AI makes human interaction feel more effortful and less satisfying over time.
Under journal review
L Ibrahim, FS Hafner, M Cheng, C Lee, R Anselmetti, R Willer, L Rocher, D Yang

The efficiency-gain illusion: People underestimate the rate of AI use and overestimate its benefits on simple tasks.
Under journal review; also peer-reviewed and presented at CogSci 2026
S Yu, M Cheng, A Jabbar, I Sucholutsky, KM Collins, D Jurafsky, RD Hawkins

Verbalizing LLMs' assumptions to explain and control sycophancy.
COLM 2026; also peer-reviewed and presented in CHI 2026 Extended Abstracts, ICLR 2026 Re-Align Workshop
M Cheng, I Sieh, H Zope, S Yu, L Ibrahim, A Arora, J Moore, D Ong, D Jurafsky, D Yang

Sycophantic AI decreases prosocial intentions and promotes dependence.
Science, 2026 **Cover story for Science magazine**
M Cheng, C Lee, P Khadpe, S Yu, D Han, D Jurafsky
Press coverage by the New York Times, Associated Press, Scientific American, NPR, and 200+ other outlets.

Accommodation and Epistemic Vigilance: A Pragmatic Account of Why LLMs Fail to Challenge Harmful Beliefs.
ACL 2026 (Oral)
M Cheng, RD Hawkins, D Jurafsky
Press coverage by IEEE Spectrum.

Thinking beyond the anthropomorphic paradigm benefits LLM research.
ACL 2026 (Oral)
L Ibrahim*, M Cheng*

ELEPHANT: Measuring and Understanding Social Sycophancy in LLMs.

ICLR 2026
M Cheng*, S Yu*, C Lee, P Khadpe, L Ibrahim, D Jurafsky
Press coverage by MIT Technology Review, NPR, IEEE Spectrum, and VentureBeat.

Metaphors of AI indicate that people increasingly perceive AI as warm and human-like.
Communications Psychology 2026; also peer-reviewed and presented at FAccT 2025
M Cheng*, AY Lee*, K Rapuano, K Niederhoffer, A Liebscher, J Hancock
Press coverage by Fortune, Forbes, and New Scientist.

Attention to Non-Adopters.
Findings of ACL 2026
K Zhou, K Gligoric, M Cheng, MS Lam, V Raman, B Aminu, C Woo, M Brockman, H Cha, D Jurafsky

Characterizing Delusional Spirals through Human-LLM Chat Logs.
FAccT 2026
J Moore, A Mehta, W Agnew, JR Anthis, R Louie, Y Mai, P Yin, M Cheng, SJ Paech, K Klyman, S Chancellor, E Lin, N Haber, D Ong
Press coverage by the Financial Times.

AI Automatons: AI Systems Intended to Imitate Humans.
FAccT 2026
A Olteanu, S Barocas, SL Blodgett, L Egede, A DeVrio, M Cheng

Computer-vision research powers surveillance technology.
Nature 2025
P Kalluri*, W Agnew*, M Cheng*, K Owens*, L Soldaini*, A Birhane*
Press coverage by 404 Media.

HumT DumT: Measuring and controlling human-like language in LLMs.
ACL 2025
M Cheng, S Yu, D Jurafsky

Dehumanizing Machines: Mitigating Anthropomorphic Behaviors in Text Generation Systems.
ACL 2025 (Oral) **SAC Highlights Award**
M Cheng, SL Blodgett, A DeVrio, L Egede, A Olteanu

A Taxonomy of Linguistic Attributes That Contribute To Anthropomorphism of Language Technologies.
CHI 2025
A DeVrio, M Cheng, L Egede, A Olteanu, SL Blodgett

"I Am the One and Only, Your Cyber BFF": Understanding the Impact of GenAI Requires Understanding the Impact of Anthropomorphic AI.
ICLR 2025, Blogposts Track
M Cheng, A DeVrio, L Egede, SL Blodgett, A Olteanu

NLP Systems That Can't Tell Use from Mention Censor Counterspeech, but Teaching the Distinction Helps.
NAACL 2024
K Gligoric, M Cheng, L Zheng, E Durmus, D Jurafsky

AnthroScore: A Computational Linguistic Measure of Anthropomorphism.
EACL 2024 (Oral); also presented at IC2S2 2024 (Oral)
M Cheng, K Gligoric, T Piccardi, D Jurafsky
Press coverage by New Scientist and Scientific American.

CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations.
EMNLP 2023
M Cheng, T Piccardi, D Yang

Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models.
ACL 2023 **ACL Social Impact Award**
M Cheng, E Durmus, D Jurafsky

Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale.
FAccT 2023
F Bianchi*, P Kalluri*, E Durmus*, F Ladhak*, M Cheng*, D Nozza, T Hashimoto, D Jurafsky, J Zou, A Caliskan
Press coverage by Nature, CBS News Prime Time, the Washington Post and MIT Technology Review.

Social Norm Bias: Residual Harms of Fairness-Aware Algorithms.
Data Mining and Knowledge Discovery 2023
M Cheng, M De-Arteaga, L Mackey, AT Kalai

Taxonomy of Risks Posed by Language Models.
FAccT 2022
L Weidinger, J Mellor, M Rauh, C Griffin, J Uesato, P Huang, M Cheng et al.

Human Preference-Based Learning for High-dimensional Optimization of Exoskeleton Walking Gaits.
IROS 2020
M Tucker, M Cheng, E Novoseller, R Cheng, Y Yue, JW Burdick, AD Ames

Invited Talks

University of Illinois Urbana-Champaign, CS 565: Ethics in AI 2026
UC Berkeley EECS 2026
Johns Hopkins University, Ethics of AI and Automation 2026
Microsoft, Policy Working Group 2025
Stanford University, Politics and Social Change Lab Seminar Series 2025
Stanford University, CS 329X: Human-centered LLMs 2025
Stanford University, COMM 324: Language and Technology 2025
Carnegie Mellon University, Sap Group Seminar 2024
Johns Hopkins University, Center for Digital Humanities Seminar 2024
EACL, Personalization of Generative AI Workshop Panel 2024
Cornell University, CS 5382: Practical Principles for Designing Fair Algorithms 2024
UT Austin, LIN 393: Social Applications and Impact of NLP 2023, 2024
Stanford University, Algorithmic Fairness Seminar 2023

Industry Experience

Student Researcher, Google Research 2025
Led research project developing a taxonomy and automatic evaluation pipeline for LLMs' adherence to cultural norms in open-ended use contexts. Supervised by Sunipa Dev and Vinodkumar Prabhakaran.

Research Intern, Microsoft Research Montreal 2024
Led research project on alternatives to anthropomorphism in LLM outputs. Published at ACL and CHI. Supervised by Alexandra Olteanu and Su Lin Blodgett (FATE team).

Research Engineer Intern, DeepMind 2021
Designed methods for benchmarking and detecting microaggressions in LLMs; contributed to FAccT paper on social and ethical risks of LLMs. Supervised by Lisa Anne Hendricks and John Mellor.

Research Intern, Microsoft Research Cambridge 2021
Investigated gender bias in automated recruiting; developed fairness framework published in Data Mining & Knowledge Discovery. Supervised by Adam Kalai.

Teaching Experience

CS 224N: Natural Language Processing with Deep Learning, Stanford (TA) 2025
CS 329R: Race and Natural Language Processing, Stanford (TA) 2024
CS 12: Algorithmic Fairness & Justice, Caltech (Student Lecturer) 2022
CS 144: Networks: Structures & Economics, Caltech (Head TA) 2022
EE 111: Signal-Processing Systems & Transforms, Caltech (TA) 2020
Peer Writing Fellow, Caltech 2019–2022
AddisCoder Teaching Assistant (data structures & algorithms, Addis Ababa) 2018

Mentoring

Research mentees, Stanford University
Pre-doctoral students: Yuewen Yang
Master's students: Isabel Sieh, Sunny Yu, Hannah Cha, Humishka Zope, Poonam Sahoo
Undergraduates: Dyllan Han, Neil Rathi, Bolu Aminu, Michael Brockman, Sofia Kim, Caeley Woo

Grants

Stanford HAI Seed Grant, "Measuring and Mitigating Social Sycophancy" ($75K) 2026
NSF Discover ACCESS Grant, "Advancing Trustworthy AI by Understanding and Mitigating Social Sycophancy in Large Language Models" ($37K) 2026
Stanford HAI-Google Research Fund, "Human-like language in LM generations and its effect on LM-human reliance" ($105K) 2024
HAI Seed Grant, "Advancing AI for Interpreting Implicit Language in the Courtroom" ($75K) 2024
NAIRR Pilot Grant, "Advancing AI for Understanding Implicit Language in the Courtroom" ($32K) 2024

Community Service

Area Chair, Reviewer 2020–present
Area Chair for COLM; Reviewer for CHI, ARR, FAccT, ICLR, IC2S2, CogSci, NeurIPS ML4D Workshop, NeurIPS Workshop on Regulatable ML, NeurIPS Ethics Committee.

Stanford Computer Science Faculty Hiring Committee 2024–2025
Reviewed faculty applications for the Stanford CS department (committee chaired by Kunle Olukotun).

Stanford Social NLP Reading Group Organizer 2023–2024
Organized weekly reading group on social and ethical aspects of NLP.

Caltech TechReach Cofounder and President 2018–2022
Developed initiative to explore societal impacts of technology; student teams built technical projects for local nonprofits.

Caltech COMPASS Mentor 2020–2022
Mentored students on academic, professional, and personal development in the "Women Mentoring Women" program.