

Sunny Chunhua Yu

+1-626-265-2016 | syu03@princeton.edu

 Sunny Yu |  sunnyych |  Google Scholar |  sunnyyuych | [personal website](#)

Stanford, California - 94305, US

EDUCATION

- **Princeton University** September 2026 - June 2031
Computer Science, PhD Princeton, New Jersey
 - research in natural and artificial intelligence
- **Stanford University** September 2022 - June 2026
Symbolic Systems (with Honors and Distinction), BS Stanford, California
 - GPA: 4.023/4.00
 - Relevant Coursework: Computer Organization and Systems; Curiosity in AI; Topics in Natural and Artificial Intelligence; Data Mining and Analysis; Linear Algebra; Mind, Matter, Meaning; How To Make A Moral Agent; Introduction to Cognitive Neuroscience
 - Activities: Symbolic Systems Society, Symbolic Systems Advising Fellow, Embodied Journal, The Stanford Daily
 - Advised by Dan Jurafsky and Robert Hawkins
- **Stanford University** September 2023 - June 2026
Computer Science, MS Stanford, California
 - Relevant Coursework: Design and Analysis of Algorithms; Deep Learning For Computer Vision; Deep Reinforcement Learning; Continuous Mathematical Methods with an Emphasis on Machine Learning; Human Centered NLP; Human-Computer Interaction: Foundations and Frontiers; Cross-platform Mobile App Development; Social Computing; Natural Language Processing with Deep Learning; Operating Systems Principles
- **Vivian Webb School** August 2018 - June 2022
Secondary Education Claremont, California
 - GPA: 4.7/4.0
 - Activities and Honors: Webb Canyon Chronicle Editor-in-Chief (Crown Award, Columbia Scholastic Press Association), Debate Team Captain, Varsity Cross Country (Captain and Most Inspirational Runner), Founder and Copy Editor of Breakfast Literary Magazine, Cum Laude Society Member, Dodd Fischer 61' Community Service Award, E.T Price Award in Humanities

RESEARCH EXPERIENCE

- **Stanford NLP Group** September 2023 - June 2026
Researcher Stanford, California
 - **HumT DumT**: Used NLP methods to measure and decrease human-likeness in large language models (LLMs).
 - **Social Sycophancy**: Constructed ELEPHANT, an evaluation framework for benchmarking social sycophancy in LLMs (press coverage by [Venture Beast](#), [MIT Technology Review](#), and [NPR](#)).
 - **Sycophancy's Effects**: Conducted human experiments to examine the behavioral outcomes of users who interact with sycophantic models; published in Science. (press coverage by [The Guardian](#), [Time](#), and [US News](#)).
 - **GSSBench**: Define and benchmark LLMs' generation space size for open-ended tasks and explore applications in human-AI conversational grounding, reasoning, and diversity optimization.
- **Stanford HCI Group** March 2024 - December 2024
Researcher Stanford, California
 - **LLM and Ontologies**: Propose ontology as a lens to analyze the pluriversality of language models.
- **Cognitive Science Research** March 2023 - present
Researcher Stanford, California
 - **The efficiency-gain illusions**: A large-scale human study identifying cognitive biases underlying human-AI interactions.
 - **Children's Book and Generics**: A large-scale corpus analysis of the use of generic and habitual statements to describe different object kinds in children's books.
 - **Dev-Bench**: Create benchmarks for evaluating 10+ multimodal models on linguistic tasks to analyze their developmental trajectories as compared with children and adults.
 - Decision Prediction (on-going): Collect an ecological dataset of humans' everyday decisions and benchmark how well LLMs can simulate and predict everyday decisions.

SELECTED PUBLICATIONS AND PREPRINTS

C=CONFERENCE, S=IN SUBMISSION, J=JOURNAL, W=WORKSHOP

- [J.1] Yu, C. (2021). [Rising Scholar: An Examination of the Institutionally Oppressive White Savior Complex in Uganda Through Western Documentaries](#). *International Social Science Review*, 97(2), 1-28.
- [J.2] Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). [Sycophantic AI decreases prosocial intentions and promotes dependence](#). *Science*.
- [S.1] Yu, S., Cheng, M., Jabbar, A., Sucholutsky, I., Collins, K., Jurafsky, D., & Hawkins, R. (2026). [The efficiency-gain illusion: People underestimate the rate of AI use and overestimate its benefits on simple tasks](#). Arxiv preprint.
- [S.2] Yu, S., Jabbar, A., Hawkins, R., Jurafsky, D., & Cheng, M. (2025). [Generation Space Size: Understanding and Calibrating Open-Endedness of LLM Generations](#). Arxiv preprint.
- [S.3] Tan, A., Cardinal, D., Lorido-Botran, T., Bravo-Sanchez, L., Yu, S., & Frank, M. (2026). [LEVANTE-bench: Multi-Scale Comparison of VLMs to Children Using Cognitive Tasks \(or, "Is Your VLM Smarter Than a 5th Grader?"\)](#). Arxiv preprint.
- [S.4] Ziems, C., Zhao, D., Wang R., Jorke, M., Rushdi, A., Deepak, A., Yu, S., ... Landay, J., & Yang, D. (2026). [Reflections and New Directions for Human-Centered Large Language Models](#). Arxiv preprint.
- [W.1] Yu, S., Cheng, M., Jabbar, A., Hawkins, R., & Jurafsky, D. (2025). [Measuring LLM Generation Spaces with EigenScore](#). *CogInterp: Interpreting Cognition in Deep Learning Models @ NeurIPS 2025*.
- [C.1] Cheng, M.*, Yu, S.*, Lee, C., Khadpe, P., Ibrahim, L., & Jurafsky, D. (2025). [ELEPHANT: Measuring and understanding social sycophancy in LLMs](#). *ICLR 2026*.
- [C.2] Tan, A., Yu, C., Long, B., Ma, W., Murray, T., Silverman, R., ... & Frank, M. C. (2024). [Devbench: A multimodal developmental benchmark for language learning](#). *NeurIPS 2025 (oral)*.
- [C.3] Yu, S., Tan, A., Zhang, S., Miao, P., Carlson, R., Gerstenberg, T., & Rose, D. (2025). [Generics revisited: Analyzing generalizations in children's books and caregivers' speech](#). *The 47th Annual Meeting of the Cognitive Science Society*.
- [C.4] Yu, S., Cheng, M., Jabbar, A., Sucholutsky, I., Collins, K., Jurafsky, D., & Hawkins, R. (2026). [Cognitive offloading and the speedup illusion in human-AI interaction](#) *The 48th Annual Meeting of the Cognitive Science Society*.
- [C.5] Cheng, M., Yu, S., & Jurafsky, D. (2025). [HumT DumT: Measuring and controlling human-like language in LLMs](#). *ACL 2025*.
- [C.6] Cheng, M., Sieh, I., Zope, H., Yu, S., Ibrahim, L., Arora, A., Moore, J., Ong, D., Jurafsky, D., & Yang, D. (2026). [Verbalizing LLMs' assumptions to explain and control sycophancy](#). *COLM 2026*.

SKILLS

- **Languages:** Python, C, C++, Javascript, R
- **Web:** React, React Native, CSS, HTML
- **ML:** Pandas, PyTorch, TensorFlow, NumPy

AWARDS AND HONORS

- **Phi Beta Kappa** June 2026
Stanford University
- **Firestone Medals for Excellence in Undergraduate Research** June 2026
Stanford University

LEADERSHIP EXPERIENCE

- **President** September 2023 - June 2024
Symbolic Systems Society
 - Founded *Machina*, Stanford's inaugural undergraduate journal in Symbolic Systems
- **Advising Fellow** September 2024 - present
Symbolic Systems Program
 - Peer Advisor for the Symbolic Systems Program; organize community events; hold office hours to advise students

VOLUTEERING AND TEACHING EXPERIENCE

- **Organizer** September 2024 - June 2025
Stanford NLP Seminar
- **Reviewer**
Reviewed for ICLR 2025, CogInterp 2025 (NeurIPS workshop), WiML 2025 (NeurIPS workshop), NeurIPS 2026
- **Course Assistant** January 2026 - March 2026
CS124: From Languages to Information (Stanford)
- **Course Assistant** September 2025 - December 2025
CS329X: Human-Centered Large Language Models (Stanford)
- **Course Assistant** June 2024 - August 2024
DATASCI112: Principles of Data Science (Stanford)

ADDITIONAL INFORMATION

Languages: English (Native), Spanish (Proficient), Chinese (Native)