Jesse Mu

muj@stanford.edu
cs.stanford.edu/~muj

Education

2018– **Ph.D. in Computer Science, Stanford University**
– Advisor: Noah Goodman

2017–2018 **MPhil in Advanced Computer Science,** *with distinction*, **University of Cambridge**
– Advisors: Ekaterina Shutova, Helen Yannakoudakis

2013–2017 **B.A. in Computer Science,** *summa cum laude*, **Boston College**
– Advisors: Joshua K. Hartshorne, Timothy J. O'Donnell

Experience

2023– **Member of Technical Staff, Anthropic**

2022 **Research Intern, DeepMind**
– Advisors: Joel Leibo and Jane Wang

2021 **Research Intern, FAIR, Meta**
– Advisors: Edward Grefenstette and Tim Rocktäschel

2020 **Visiting Researcher, Language and Intelligence Group (LINGO), MIT**
– Advisor: Jacob Andreas

2019–2020 **Course Consultant, Codecademy**
– Course advisor/designer for Deep Learning and Text Generation course

2017 **Applied Scientist Intern, Alexa AI, Amazon**

Preprints

2025 **Forecasting rare language model behaviors**
Erik Jones, Meg Tong, **Jesse Mu**, Mohammed Mahfoud, Jan Leike, Roger Grosse, Jared Kaplan, William Fithian, Ethan Perez, and Mrinank Sharma

2025 **Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming**
Mrinank Sharma, Meg Tong, **Jesse Mu**, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, Amanda Askell, Nathan Bailey, Joe Benton, Emma Bluemke, Samuel R. Bowman, Eric Christiansen, Hoagy Cunningham, Andy Dau, Anjali Gopal, Rob Gilson, *et al*

2024 **Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training**
Evan Hubinger, Carson Denison, **Jesse Mu**, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, *et al*

Publications

2023 **Learning to Compress Prompts with Gist Tokens**
Jesse Mu, Xiang Lisa Li, and Noah Goodman. In *Advances in Neural Information Processing Systems (NeurIPS)*

2023 **Characterizing tradeoffs between teaching via language and demonstrations in multi-agent systems**
Dhara Yu, Noah Goodman, and **Jesse Mu**. In *Proceedings of the 45th Annual Meeting of the Cognitive Science Society (CogSci)*

2022 **Improving Intrinsic Exploration with Language Abstractions**
Jesse Mu, Victor Zhong, Roberta Raileanu, Minqi Jiang, Noah Goodman, Tim Rocktäschel, and Edward Grefenstette. In *Advances in Neural Information Processing Systems (NeurIPS)*

2022 **Improving Policy Learning with Language Dynamics Distillation**
Victor Zhong, **Jesse Mu**, Luke Zettlemoyer, Edward Grefenstette, and Tim Rocktäschel. In *Advances in Neural Information Processing Systems (NeurIPS)*

2022 **STaR: Bootstrapping Reasoning with Reasoning**
Eric Zelikman, Yuhai Wu, **Jesse Mu**, and Noah Goodman. In *Advances in Neural Information Processing Systems (NeurIPS)*

2022 **Active Learning Helps Pretrained Models Learn the Intended Task**
Alex Tamkin, Dat Nguyen, Salil Deshpande, **Jesse Mu**, and Noah Goodman. In *Advances in Neural Information Processing Systems (NeurIPS)*

2022 **In the ZONE: Measuring difficulty and progression in curriculum generation**
Rose Wang, **Jesse Mu**, Dilip Arumugam, Natasha Jaques, and Noah Goodman. In *NeurIPS Deep RL Workshop*

2022 **Emergent Covert Signaling in Adversarial Reference Games**
Dhara Yu, **Jesse Mu**, and Noah Goodman. In *Proceedings of the 5th Workshop on Emergent Communication: New Frontiers*

2021 **Multi-party Referential Communication in Complex Strategic Games**
Jessica Mankewitz, Veronica Boyce, Brandon Waldon, Georgia Loukatou, Dhara Yu, **Jesse Mu**, Noah Goodman, and Michael Frank. In *NeurIPS Meaning in Context (MiC) Workshop*

2021 **Emergent Communication of Generalizations**
Jesse Mu and Noah Goodman. In *Advances in Neural Information Processing Systems (NeurIPS) (previously NAACL 2021 Workshop on Visually Grounded Interaction and Language)*

2021 **Calibrate Your Listeners! Robust Communication-based Training for Pragmatic Speakers**
Rose E. Wang, Julia White, **Jesse Mu**, and Noah Goodman. In *Findings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*

2020 **Compositional Explanations of Neurons**
Jesse Mu and Jacob Andreas. In *Advances in Neural Information Processing Systems (NeurIPS)* **[oral (top 1.1%)]**

2020 **Learning to Refer Informatively by Amortizing Pragmatic Reasoning**
Julia White, **Jesse Mu**, and Noah Goodman. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci)*

2020 **Shaping Visual Representations with Language for Few-shot Classification**

Jesse Mu, Percy Liang, and Noah Goodman. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) (previously NeurIPS 2019 Workshop on Visually Grounded Interaction and Language)*

2019 **Learning Outside the Box: Discourse-level Features Improve Metaphor Identification**
Jesse Mu, Helen Yannakoudakis, and Ekaterina Shutova. In *Proceedings of the 2019 North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*

2019 **Do we need natural language? Exploring "restricted" language interfaces for complex domains**
Jesse Mu and Advait Sarkar. In *CHI '19 Extended Abstracts on Human Factors in Computing Systems*

2019 **The meta-science of adult statistical word segmentation: Part 1**
Joshua K. Hartshorne, Lauren Skorb, Sven L. Dietz, Caitlin R. Garcia, Gina L. Iozzo, Katie E. Lamirato, James R. Ledoux, **Jesse Mu**, Kara N. Murdock, Jon Ravid, Alyssa A. Savery, James E. Spizzirro, Kelsey A. Trimm, Kendall D. van Horne, and Juliani Vidal. *Collabra* 5(1):1

2017 **Evaluating hierarchies of verb argument structure with hierarchical clustering**
Jesse Mu, Joshua K. Hartshorne, and Timothy J. O'Donnell. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*

2017 **Parkinson's disease subtypes identified from cluster analysis of motor and non-motor symptoms**
Jesse Mu, Kallol Ray Chaudhuri, Concha Bielza, Jesús de Pedro Cuesta, Pedro Larrañaga, and Pablo Martinez-Martin. *Frontiers in Aging Neuroscience* 9:301

Honors and awards

2021	Open Philanthropy AI Fellowship
2018	Finch Family Fellowship, Stanford School of Engineering
2018	NSF Graduate Research Fellowship
2017	John J. Neuhauser Award in Computer Science, Boston College
2017	Thomas I. Gasson, S.J. Award, Boston College
2017	Phi Beta Kappa
2017	Churchill Scholarship
2016	Barry M. Goldwater Scholarship
2013	Gabelli Presidential Scholarship, Boston College

Teaching

2023	Teaching Assistant, CS 224n Natural Language Processing with Deep Learning, Stanford
2022	Teaching Assistant, CS 221 Artificial Intelligence: Principles and Techniques, Stanford
2020	Guest Lecturer, Structure and Interpretation of Deep Networks, MIT IAP
2014–2016	Teaching Assistant, Computer Science I, Boston College

Leadership and service

2020–2021	Organizer, Stanford NLP Seminar
2014–2017	Co-president, Boston College Computer Science Society

Reviewing

2023 ICLR, TMLR (**expert reviewer**), ACL, NeurIPS

2022 ICLR (**highlighted reviewer**), ICML (**outstanding reviewer**), TMLR, ACL Learning with Natural Language Supervision (LNLS) Workshop, NeurIPS Language and Reinforcement Learning (LaReL) Workshop, NeurIPS Workshop on Interactive Learning for Natural Language Processing (InterNLP)

2021 NAACL, ACL, EMNLP, NeurIPS (**outstanding reviewer**), NeurIPS Meaning in Context Workshop