Swabha Swayamdipta

Gabilan Assistant Professor of Computer Science, University of Southern California

swabhas@usc.edu | https://swabhs.com | swabhs | @swabhz

Appointments

University of Southern California — *Los Angeles, CA, USA*

Gabilan Assistant Professor of Computer Science — *Aug 2022 - PRESENT*

- Co-Associate Director of the Center for AI and Society

Amazon AGI — *Los Angeles, CA, USA*

Amazon Scholar — *Jan 2025 - PRESENT*

- Team: Responsible AI

Allen Institute for AI — *Seattle, WA, USA*

Postdoctoral Investigator — *Aug 2019 - June 2022*

- Team: MOSAIC led by Yejin Choi

Oracle Server Technologies — *Bangalore, India*

Member Technical Staff — *2010 - 2011*

- Team: Unified messaging for Fusion Middleware

Education

Carnegie Mellon University — *Pittsburgh, PA, USA*

PhD in Language and Information Technologies — *2013 - 2019*

- **Advisors:** Noah A. Smith, Chris Dyer
- **Thesis:** Syntactic Inductive Biases in NLP

Columbia University — *City of New York, NY, USA*

Masters in Computer Science — *2011 - 2012*

- **Advisors:** Owen Rambow, Michael Collins

National Institute of Technology — *Calicut, Kerala, India*

Bachelors of Technology in Computer Science and Engineering — *2006 - 2010*

Awards and Grants

2025	**Simons Institute Research Fellowship**, Special Program on LLMs and Transformers, Spring 2025	***UC Berkeley***
2024	**Outstanding Paper Award**, OATH-Frames: Characterizing Online Attitudes Towards Homelessness with LLM Assistants	***EMNLP***
2024	**NSF CISE Core Medium**, RII: Hard Data to the Model	***NSF***
2024	**Zumberge Fellowship**, DEI in Research	***USC OORI***
2023	**Intel Rising Stars Award**, LLM Inference for Scaling Data Creation	***Intel Labs***
2022	**AI2 Young Investigators Award**, Synthesizing Data Towards Commonsense and Causal Reasoning	***Allen Institute for AI***
2022	**Gabilan Assistant Professor Fellowship**,	***USC WISE***
2022	**Outstanding Paper Award**, Understanding Dataset Difficulty with V-Usable Information	***ICML***
2021	**Outstanding Paper Award**, MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers	***NeurIPS***
2020	**Best Paper: Honorable Mention**, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks	***ACL***
2019	**Rising Star in EECS**, UIUC	
2014	**Best Student Presentation**, Diversity in Dependency Parsing	***CMU LTI Student Research Symposium***
2009	**Sun Campus Ambassador Scholarship**, Bangalore, India	***Sun Microsystems***
2003	**National Talent Search Scholarship**, Rourkela, India	***NCERT, India***

Publications

Conference and Journal Papers

Better Language Model Inversion by Compactly Representing Next-Token Distributions — *NeurIPS 2025*
- M. Nazir, M. Finlayson, J. X. Morris, X. Ren, **S. Swayamdipta**

Teaching Models to Understand (but not Generate) High-risk Data — *CoLM 2025*
- R. Wang, M. Finlayson, L. Soldaini, **S. Swayamdipta**, R. Jia

Improving LLM Personas via Rationalization with Psychological Scaffolds — *EMNLP 2025*
- B. Joshi, X. Ren, **S. Swayamdipta**, R. Koncel-Kedziorski, T. Paek

Evaluating Evaluation Metrics – The Mirage of Hallucination Detection — *EMNLP-Findings 2025*
- A. Kulkarni, Y. Zhang, J. R. A. Moniz, X. Ge, B.-H. Tseng, D. Piraviperumal, **S. Swayamdipta**, H. Yu

Robust Data Watermarking in Language Models by Injecting Fictitious Knowledge — *ACL-Findings 2025*
- X. Cui, J. Wei, **S. Swayamdipta**, R. Jia,

ELI-Why: Evaluating the Pedagogical Utility of LLM Explanations — *ACL-Findings 2025*
- B. Joshi , K. He , S. Ramnath , S. Sabouri , K. Zhou , S. Chattopadhyay, **S. Swayamdipta**, X. Ren,

Compare without Despair: Reliable Preference Evaluation with Generation Separability — *EMNLP 2024*
- S. Ghosh, T. Srinivasan, **S. Swayamdipta**,

Out-of-Distribution Detection through Soft Clustering with Non-Negative Kernel Regression — *EMNLP 2024*
- A. Gulati, X. Dong, C. Hurtado, S. Shekkizhar, **S. Swayamdipta**, A. Ortega

OATH-Frames: Characterizing Online Attitudes towards Homelessness via LLM Assistants — *EMNLP 2024*
- J. Ranjit, B. Joshi, R. Dorn, L. Petry, O. Koumoundouros, J. Bottarini, P. Liu, E. Rice, and **S. Swayamdipta**
- **Outstanding Paper Award @ EMNLP 2024 & Best Poster @ ShowCAIS 2024**

Logits of API-Protected LLMs Leak Proprietary Information — *COLM 2024*
- M. Finlayson, X. Ren, **S. Swayamdipta**,

Crowd-Calibrator: Can Annotator Disagreement Inform Calibration in Subjective Tasks? — *COLM 2024*
- U. Khurana, E. Nalisnick, A. Fokkens, **S. Swayamdipta**,

Annotating FrameNet via Structure-Conditioned Language Generation — *ACL 2024*
- X. Cui, and **S. Swayamdipta**

Closing the Curious Case of Neural Text Degeneration — *ICLR 2024*
- M. Finlayson, J. Hewitt, A. Koller, **S. Swayamdipta**, A. Sabharwal

Does Video Summarization Require Videos? Quantifying the Effectiveness of Language in Video Summarization — *ICASSP 2024*
- Y. Nam, A. Lehavi, D. Yang, D. Bose, **S. Swayamdipta**, and S. Narayanan

NeuroComparatives: Neuro-Symbolic Distillation of Comparative Knowledge — *NAACL-Findings 2024*
- P. Howard, J. Wang, V. Lal, G. Singer, Y. Choi, and **S. Swayamdipta**

Generative Explanations for Program Synthesizers — *ICSE + VL/HCC 2024*
- A. Nazari, S. Chattopadhyay, **S. Swayamdipta**, M. Raghothaman

We're Afraid Language Models Aren't Modeling Ambiguity — *EMNLP 2023*
- A. Liu, Z. Wu, J. Michael, A. Suhr, P. West, A. Koller, **S. Swayamdipta**, N. A. Smith, and Y. Choi

MAUVE Scores for Generative Models: Theory and Practice — *JMLR 2023*
- K. Pillutla, L. Liu, J. Thickstun, S. Welleck, **S. Swayamdipta**, R. Zellers, S. Oh, Y. Choi, and Z. Harchaoui

I2D2: Inductive Knowledge Distillation with NeuroLogic and Self-Imitation — *ACL 2023*
- C. Bhagavatula, J. D. Hwang, D. Downey, R. Le Bras, X. Lu, K. Sakaguchi, **S. Swayamdipta**, P. West, and Y. Choi

REV: Information-Theoretic Evaluation of Free-Text Rationales — *ACL 2023*
- H. Chen, F. Brahman, X. Ren, Y. Ji, Y. Choi, and **S. Swayamdipta**

COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements — *ACL-Findings 2023*
- X. Zhou, H. Zhu, A. Yerukola, T. Davidson, J. D. Hwang, **S. Swayamdipta**, and M. Sap

Investigating the Benefits of Free-Form Rationales *EMNLP-Findings 2022*
J. Sun, **S. Swayamdipta**, J. May, and X. Ma

NeuroCounterfactuals: Beyond Minimal-Edit Counterfactuals for Richer Data Augmentation *EMNLP-Findings 2022*
P. Howard, G. Singer, V. Lal, Y. Choi, and **S. Swayamdipta**

WaNLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation *EMNLP-Findings 2022*
A. Liu, **S. Swayamdipta**, N. A. Smith, and Y. Choi

Reframing Human-AI Collaboration for Generating Free-Text Explanations *NAACL 2022*
S. Wiegreffe, J. Hessel, **S. Swayamdipta**, M. Riedl, and Y. Choi

Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection *NAACL 2022*
M. Sap, **S. Swayamdipta**, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith

Understanding Dataset Difficulty with V-Usable Information *ICML 2022*
- K. Ethayarajh, Y. Choi, and **S. Swayamdipta**
- **Outstanding Paper Award**

MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers *NeurIPS 2021*
- K. Pillutla, **S. Swayamdipta**, R. Zellers, J. Thickstun, S. Wellecks, Y. Choi, and Z. Harchaoui
- **Outstanding Paper Award**

Contrastive Explanations for Model Interpretability *EMNLP 2021*
A. Jacovi, **S. Swayamdipta**, S. Ravfogel, Y. Elazar, Y. Choi, and Y. Goldberg

On-the-Fly Controlled Text Generation with Experts and Anti-Experts *ACL 2021*
A. Liu, M. Sap, X. Lu, **S. Swayamdipta**, C. Bhagavatula, N. A. Smith, and Y. Choi

Challenges in Automated Debiasing for Toxic Language Detection *EACL 2021*
- X. Zhou, M. Sap, **S. Swayamdipta**, N. A. Smith, and Y. Choi

Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics *EMNLP 2020*
- **S. Swayamdipta**, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, Y. Choi

Generative Data Augmentation for Commonsense Reasoning *EMNLP-Findings 2020*
- Y. Yang, C. Malaviya, J. Fernandez, **S. Swayamdipta**, R. LeBras, J. Wang, C. Bhagavatula, Y. Choi, and D. Downey

Adversarial Filters of Dataset Biases *ICML 2020*
- R. LeBras, **S. Swayamdipta**, C. Bhagavatula, R. Zellers, M. E. Peters, A. Sabharwal, and Y. Choi

The Right Tool for the Job: Matching Model and Instance Complexities *ACL 2020*
- R. Schwartz, G. Stanovsky, **S. Swayamdipta**, J. Dodge, and N. A. Smith

Don't Stop Pretraining: Adapt Language Models to Domains and Tasks *ACL 2020*
- S, Gururangan, A. Marasović, **S. Swayamdipta**, K. Lo, I. Beltagy, D. Downey, and N. A. Smith
- **Best Paper Honorable Mention**

Syntactic Scaffolds for Semantic Structures *EMNLP 2018*
- **S. Swayamdipta**, S. Thomson, K. Lee, L. Zettlemoyer, C. Dyer, N. A. Smith.

Learning Joint Semantic Parsers from Disjoint Data *NAACL 2018*
- H. Peng, S. Thomson, **S. Swayamdipta**, N. A. Smith

Annotation Artifacts in Natural Language Inference Data *NAACL 2018*
- S. Gururangan*, **S. Swayamdipta***, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith
- * equal contribution

Polyglot Semantic Role Labeling *ACL 2018*
- P. Mulcaire, **S. Swayamdipta**, and N. A. Smith

Multi-Mention Learning for Reading Comprehension with Neural Cascades *ICLR 2018*
- **S. Swayamdipta**, A. Parikh, T. Kwaitkowski

Greedy, Joint Syntactic and Semantic Parsing with Stack LSTMs *CoNLL 2016*
- **S. Swayamdipta**, M. Ballesteros, C. Dyer, N. A. Smith

A Dependency Parser for Tweets *EMNLP 2014*
- L. Kong, N. Schneider, **S. Swayamdipta**, A. Bhatia, C. Dyer, N. A. Smith

The Pursuit of Power and its Manifestation in Written Dialog — *ICSC 2012*
- **S. Swayamdipta**, O. Rambow

Workshop Papers

Sample, Align, Synthesize: Graph-Based Response Synthesis with CONGRS — *ScaLR@COLM / ER@NeurIPS 2025*
- S. Ghosh, S. S. Warraich, D. Tarsadiya, G. Yauney, **S. Swayamdipta**

Evaluation Under Imperfect Benchmarks and Ratings: A Case Study in Text Simplification — *LLM-Evals@NeurIPS 2025*
- J. Liu, Y. Nam, X. Cui, **S. Swayamdipta**

Uncovering Intervention Opportunities for Suicide Prevention with Language Model Assistants — *EAAMO / GenAI for Health @NeurIPS 2025*
- J. Ranjit, H. J. Cho, C. J. Smerdon, Y. Nam, M. Phung, J. May, J. R. Blosnich, **S. Swayamdipta**

ChEmREF: Evaluating Language Model Readiness for Chemical Emergency Response Assistance — *LLM-Evals@NeurIPS 2025*
- R. Surana, Q. Ye, **S. Swayamdipta**

Sister Help: Data Augmentation for Frame-Semantic Role Labeling — *LAW-DMR @ EMNLP 2021*
- A. Pancholy, **S. Swayamdipta**, M. R. L. Petruck

Multi-Task Learning for Incremental Parsing using Stack LSTMs — *WiML @ NeurIPS 2016*
- **S. Swayamdipta**, M. Ballesteros, C. Dyer, N. A. Smith

CMU: Arc-Factored, Discriminative Semantic Dependency Parsing — *SemEval 2014*
- S. Thomson, D. Bamman, J. Dodge, **S. Swayamdipta**, N. Schneider, C. Dyer, N. A. Smith

The CMU Machine Translation Systems — *WMT 2014*
- A. Matthews, C. Dyer, A. Lavie, G. Hanneman, W. Ammar, A. Bhatia, **S. Swayamdipta**, E. Schlinger, Y. Tsvetkov

Working Papers

Are We Automating the Joy Out of Work? Designing AI to Augment Work, Not Meaning — *Under Review at CHI 2026*
- J. Ranjit, K. Zhou, **S. Swayamdipta**, D. Quercia

Believing without Seeing: Quality Scores for Contextualizing Vision-Language Model Explanations — *Under Review at ARR 2026*
- K. He, T. Srinivasan, B. Joshi, X. Ren, J. Thomason, **S. Swayamdipta**

How Reliable is Language Model Micro-Benchmarking? — *Under Review at ICLR 2026*
- G. Yauney, S. S. Warraich, **S. Swayamdipta**

Every Language Model Has a Forgery-Resistant Signature — *Under Review at ICLR 2026*
- M. Finlayson, X. Ren, **S. Swayamdipta**

Uncovering and Mitigating Covert Dialect Bias in LLMs — *Under Review at FaccT 2026*
- K. Kondapally, C. Smerdon, P. Patel, O. Akoni, J. Torres, J. Ranjit, M. Finlayson, **S. Swayamdipta**

BenchBrowser: Retrieving Evidence of Intent Representation from Evaluation Benchmarks — *Under prep for ICML 2026*
- H. Diddee, G. Yauney, **S. Swayamdipta**, D. Ippolito

Mark My Words: Toward Unforgeable Language Model Signatures via Token Rankings — *Under prep for ICML 2026*
- M. Finlayson, A. Grivas, X. Ren, **S. Swayamdipta**

Representation Collapse in Language Models — *Under prep for ICML 2026*
- A. Kulkarni, A. Balasubramonian, J. Springer, **S. Swayamdipta**

On the Trustworthiness of Generative Foundation Models: Guideline, Assessment, and Perspective — *Under Review at TMLR 2026*
- Y. Huang, C. Gao, S. Wu, H. Wang, X. Wang, Y. Zhou, Y. Wang, J. Ye, J. Shi, and 57 more authors

Political-LLM: Large Language Models in Political Science — *Under Review at TMLR 2026*
- L. Li, J. Li, C. Chen, F. Gui, H. Yang, C. Yu, Z. Wang, J. Cai, J. A. Zhou, and 38 more authors

Shallow Syntax in Deep Water — *arXiv:1908.11047*
- **S. Swayamdipta**, M. Peters, B. Roof, C. Dyer and N. A. Smith

Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold *arXiv:1706.09528*

- **S. Swayamdipta**, S. Thomson, C. Dyer and N. A. Smith

DyNet: The Dynamic Neural Network Toolkit *arXiv:1701.03980*

- G. Neubig, C. Dyer, Y. Goldberg, A. Matthews, W. Ammar, A. Anastasopoulos, M. Ballesteros, D. Chiang, D. Clothiaux, T. Cohn, K. Duh, M. Faruqui, C. Gan, D. Garrette, Y. Ji, L. Kong, A. Kuncoro, G. Kumar, C. Malaviya, P. Michel, Y. Oda, M. Richardson, N. Saphra, **S. Swayamdipta**, P. Yin

Tutorial and Organization Papers

Proceedings of the 3rd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2022) *NAACL 2022*

- C. Cherry, A. Fan, G. Foster, G. Haffari, S. Khadivi, N. Peng, X. Ren, E. Shareghi, **S. Swayamdipta**

Transfer Learning in Natural Language Processing *NAACL 2019*

- S. Ruder, M. E. Peters, **S. Swayamdipta**, T. Wolf

Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019) *EMNLP 2019*

- C. Cherry, G. Durrett, G. Foster, R. Haffari, S. Khadivi, N. Peng, X. Ren, **S. Swayamdipta**

Frame Semantics across Languages: Towards a Multilingual FrameNet *CoLING 2018*

- C. F. Baker, M. Ellsworth, M. R. L. Petruck, **S. Swayamdipta**

Mentorship

Ph.D. students

Fall 2024-Present	**Xinyue Cui**, Ph.D.	
Fall 2024-Present	**Muru Zhang**, Ph.D.	*Co-Advisor: Robin Jia*
Fall 2024-Present	**Atharva Kulkarni**, Ph.D.	
Spring 2024-Present	**Brihi Joshi**, Ph.D., passed Proposal	*Co-Advisor: Xiang Ren*
Fall 2023-Present	**Matthew Finlayson**, Ph.D., passed Quals	*Co-Advisor: Xiang Ren*
Fall 2022-Present	**Jaspreet Ranjit**, Ph.D., passed Quals	

Masters Students

Summer 2025-Present	**Harshavardhan Alimi**, Masters
Spring 2025-Present	**Dhruv Tarsadiya**, Masters
Spring 2024-Present	**Shahzaib Saqib Warraich**, Masters
Spring 2024-Present	**Xingjian Dong**, Masters

Undergrads

Fall 2023-Present	**Risha Surana**, BS / PDP
Fall 2025-Present	**Naysa Bhargava**, BS
Fall 2025-Present	**Thor Christoffersen Hochman**, BS
Fall 2025-Present	**Brennen Ho**, BS
Fall 2025-Present	**Mike Gee**, BS

Alumni

Fall 2024-Summer 2025	**Gregory Yauney**, PostDoc	
Spring 2024-2025	**Ryan Wang**, BS	*Now at UC Berkeley*
Spring 2024-2025	**Joseph Liu**, BS	*Now at CMU*
Spring 2024-2025	**Keyu He**, BS	*Now at CMU*
Summer 2023-2025	**Aryan Gulati**, BS	*Now at Bloomberg*
Summer 2023-2024	**Catherine He**, BS	
Spring 2023-Spring 2024	**Yoonsoo Nam**, Masters	*Now at RenderWolf*
Summer 2023	**Ruyuan Zuo**, Masters	*Now at Google*
Spring 2022	**Hanjie Chen / Junlin Wang / Abdallah Bashir**, Ph.D., UVA / Masters, UCI / Masters, Saarland Universit	*AI2 Intern*
Fall 2021	**Jillian Fisher / Liwei Jiang**, Ph.D., UW	
Summer 2021	**Kawin Ethayarajh**, Ph.D., Stanford University	*AI2 Intern*
Summer 2021	**Sarah Wiegreffe**, Ph.D., George Institute of Technology	*AI2 Intern*
Summer 2021	**Ximing Liu**, BS, UW CSE	*AI2 Intern*
Fall 2020-2023	**Alisa Liu**, Ph.D., UW CSE	
Fall 2020	**Alon Jacovi**, Ph.D., Bar Ilan University	*AI2 Intern*
Summer 2020-1	**Jenny Liang**, BS, UW CSE	*AI2 Intern*
Spring 2020-2023	**Xuhui Zhou**, CLMS, UW Linguistics	
Fall 2019 - Spring 2020	**Yiben Yang**, Ph.D., Northwestern University	
Fall 2019 - Spring 2020	**Chaitanya Malaviya**, PYI, Allen Institute for AI	

Professional Service

External Service

Senior Area Chair

- EMNLP 2025: Generalizability and Transfer
- EMNLP 2024: Machine Learning in NLP
- ACL 2024: Sentence-level Semantics

Area Chair

- ICLR 2026, COLM 2025, ICLR 2025, COLM 2024, ICLR 2024, EMNLP 2023: Machine Learning for NLP, ACL 2023: Interpretability, EMNLP 2022: Language Models, EMNLP 2021: Machine Learning for NLP, NAACL 2021: Sentence-Level Semantics, EACL 2021: Sentence-Level Semantics, ACL 2020: Semantics (Long)

Reviewer

- **Conferences:** ACL (ARR): 2015-2022,2025; NAACL : 2015-2022; EMNLP : 2015-2022; NeurIPS : 2018-2022; ICML : 2015, 2019, 2020, 2023; EACL : 2017; AAAI : 2017-2020; CoNLL : 2017-2018, 2020
- **Journals:** Transactions of ACL : 2020 - 2022; Computational Linguistics : 2019 - 2022; Journal of AI Research : 2019

Co-organizer

- EMNLP 2025: Workshop on Uncertainty in NLP 2 — 2025
- ICML 2024: 2nd Workshop on High-dimensional Learning Dynamics (HiLD): The Emergence of Structure and Reasoning — 2024
- EACL 2024: Workshop on Uncertainty in NLP — 2024
- NAACL 2022: Workshop on Deep Learning in Low Resource NLP (DeepLo) — 2022
- EMNLP 2019: Workshop on Deep Learning in Low Resource NLP (DeepLo) — 2019
- West Coast NLP Workshop — 2018

Teaching

Classes

Fall 2025	**CSCI 444: NLP**, Students: 14	***USC***
Fall 2024	**CSCI 544: Applied NLP**, Students: 260	***USC***
Spring 2024	**CSCI 499: Language Models in NLP**, Students: 28	***USC***
Fall 2023	**CSCI 499: Language Models in NLP**, Students: 30	***USC***
Fall 2022	**CSCI 699: Data-Centric NLP**, Students: 29	***USC***

Invited Panels

Oct 18, 2024	**Alfred P. Sloan Foundation Science Seminar**,	*USC School of Cinematic Arts*
Jul 13, 2023	**Limitations of Large Language Models**, Rep4NLP Workshop	*ACL Toronto*
Jun 29, 2023	**Generative AI and Education**, Roundtable	*ASEE 2023*
Apr 5, 2023	**AImaginings**, Polymathic Pizza	*USC Sidney Harmon Academy for Polymathic Studies*
Jul 14, 2022	**Adversarial Data Augmentations**, DADC Workshop	*NAACL Seattle*
Jun 17, 2022	**Role of LLMs**, Responsible AI Symposium	*AILA*

Invited Talks (Since 2020)

Oct 16, 2025	**Cornell Tech Learning Machine Seminar Series (LMSS),** Simple Learning Recipes for Safe and Accountable Language Models	*Cornell Tech*
Sep 16, 2025	**USC School of Social Work,** The heaviest data: Can AI alleviate the weight of suicide mortality research?	*USC*
Apr 10, 2025	**Bloomberg CTO Data Science Speaker Series,** Rethinking Language Model Evaluation	*Bloomberg*
Apr 4, 2025	**Simons Institute: The Future of LMs and Transformers,** The Future of LMs: A Perspective on Evaluation	*UC Berkeley*
Feb 14, 2025	**Simons Research Fellow Introductions,** Rethinking the Evaluation of Large Language Models	*UC Berkeley*
Feb 12, 2025	**UC Berkeley AI and Society Mixer,** Building LLM Assistants for Social Workers	*UC Berkeley*
Nov 07, 2024	**NYU NLP/Text-as-Data Talk Series,** Ensuring Safety and Accountability in LLMs, Pre- and Post-Training	*New York University*
Sep 23, 2024	**Amazon AWS Bedrock AI Seminar,** Ensuring Safety and Accountability in LLMs, Pre- and Post-Training	*Amazon AWS, Virtual*
May 3, 2024	**ISI AI Seminar,** Understanding LLMs through their Generative Behavior, Successes and Shortcomings	*USC ISI*
Mar 26, 2024	**NSF-Open Source Generative AI Workshop,** Towards (Closed-Source) LLM Accountability via Logit Signatures	*Cornell Tech*
Mar 13, 2024	**Utah Data Science Seminar,** Understanding LLMs through their Generative Behavior, Successes and Shortcomings	*University of Utah*
Feb 13, 2024	**Intel Rising Stars Talk,** LLM Inference for Scaling Data Creation	*Intel Labs*
Feb 8, 2024	**Cambridge LTL Seminar,** Understanding LLMs via their Generative Successes and Shortcomings	*University of Cambridge*
Dec 15, 2023	**ATTRIB Workshop Keynote,** Understanding LLMs via their Generative Successes and Shortcomings	*NeurIPS 2023*
Nov 16, 2023	**UCLA Communications Symposium,** Understanding Online Discourse through Social Context and Structured Pragmatics	*UCLA*
Oct 3, 2023	**DataComp Workshop,** Understanding Data with $\mathcal{V}$-Information	*ICCV 2023*
Sep 22, 2023	**LA County of Education CS Speaker Series,** The Role of Language Models in NLP	*Los Angeles CoE*
Aug 24, 2023	**CMU LTI Student Research Symposium,** Understanding Datasets and Explanations through $\mathcal{V}$-Information	*Alumni Keynote: CMU, Pittsburgh*
Jul 13, 2023	**Rep4NLP Workshop,** Contextualizing Representations in Varied Annotator Perspectives	*ACL Toronto*
Jun 14, 2023	**Google Responsible Machine Learning,** Contextualizing Data in Varied Annotator Perspectives	*Google Brain*
Mar 14, 2023	**Cohere for AI Reading Group,** Understanding Dataset Difficulty with V-Usable Information	*Cohere for AI*
Feb 28, 2023	**Spotify Research Seminar,** Designing Controls and Filters for Dataset Generation	*Spotify Research Labs*
Feb 03, 2023	**Amazon Data-Centric AI Seminar,** What's in a Dataset? Interpreting datasets to enable better data creation	*Amazon*
Feb 01, 2023	**USC CAIS++ Seminar,** Contextualizing Bias in Hate Speech Detection	*USC*
Nov 18, 2022	**SoCal NLP Symposium,** Generating Datasets for Robust Generalization	*UC Santa Barbara*
Nov 09, 2022	**USC Center for AI in Society (CAIS) Seminar,** Contextualizing Bias in Hate Speech Detection through Annotator Perspectives	*USC*
May 23, 2022	**ACL Spotlight Talks for Young Rising Stars,** The Devil's in the Data: Mapping and Generating Datasets for Robust Generalization	*Dublin, Ireland*
May 13, 2022	**UC Irvine CS Seminar,** Mapping and Generating Datasets for Robust Generalization	*UC Irvine*
Mar 16, 2022	**UC Santa Cruz IFDS Ethics Seminar,** Rethinking Dataset Construction: : The Role of Generative Modeling and Annotator Perspectives	*UC Santa Cruz*
Oct 20, 2021	**Oracle ML Talks,** What's in your Data? Mapping Datasets and Exploring Data Usability	*Oracle*
Feb 24, 2020	**Georgetown NERT Seminar ,** Addressing Biases for Robust, Generalizable AI	*Georgetown University*
Feb 12, 2020	**Georgia Tech NLP Seminar,** Addressing Biases for Robust, Generalizable AI	*Georgia Tech*
Nov 02, 2020	**Microsoft E+D Product Leads,** Responsible AI: Addressing Biases in Datasets and Models	*Microsoft*