Time and again, we have seen that AI models are just as good as the data they are trained on. The crux of my research is the **acquisition of the next generation data for training and evaluating models of natural language**, with a goal of enabling data-efficient learning, creating trustworthy and interpretable models, and ultimately empowering responsible AI. My research is aimed at *drawing attention to the role of data and its quality*: valuable training data to teach the model to effectively reason about the world and communicate with human users, as well as effective evaluation data to meticulously estimate model capabilities. My research thus far has followed three broad thrusts, which I outline below.

**Training AI on Data with Information-Theoretic Guarantees** My research has shown the critical value of training data in producing language models effective for specialized tasks. This is particularly relevant today, as we have entered an era of general purpose, large language models (LLMs) which while displaying significant language understanding skills, still benefit greatly from task-specific supervision [12]. To this end, I have designed algorithms that quantitatively estimate the difficulty of task data [8], using information-theoretic principles [3]; this won an outstanding paper award at ICML 2022. Intuitively, the harder the training data for a model, the more new information it learns, which translates into better generalizability to unseen domains and novel data distributions; our work on domain adaptation of language models [4] won an honorable mention award at ACL 2020. I have also developed tools to visualize and analytically estimate the degree of training data difficulty, based on point-estimates [16]. These measures have led to **data-efficient learning** in language models, by training only on subsets of data most valuable for the model, and therefore resulting in models robust to spurious correlations and biases in the data.

My future research is aimed at efficient generative algorithms to produce data with such information-theoretic guarantees. Building on my prior work on algorithms for constrained generation [6, 7, 9], I will design algorithms customizable for task-specific outcomes. Specifically, I am currently devising an algorithm which yields provably better truncation policies, given the low rank of hidden dimensions in Transformer architectures. This has implications for generating plausible as well as diverse generations than allowed by the standard truncation sampling inference, such as the one available in OpenAI APIs for the GPT family.

**Training and Evaluating AI on Data with Neuro-Symbolic Knowledge** My investigations into training data have led me to design novel algorithms for creating new and large-scale resources of high quality data, designed to be exempt from aforementioned biases. Based on my prior work [5], it was clear to me that manual creation/curation of such resources is neither exempt of biases, nor scalable. This led me to develop algorithms that employ an LLM assistant for overgenerating examples, while taking into consideration desired aspects of generation (e.g. explainability). Concretely, I have developed generative algorithms for distilling the knowledge of LLMs along desired dimensions, such as for comparative knowledge [7], social common sense knowledge [20] and knowledge about generics [1]. Often the desired aspects of knowledge can be narrowly focused: I have designed algorithms for creating counterfactuals of existing data instances [6], eliciting an ambiguous phrasing of a sentence [10, 11] and generating explanations for a stated fact / model decision [18]. These resources were instrumental in training robust and generalizable models, while offering higher interpretability to model decisions, albeit from a data-centric perspective.

My future research aims to innovate on evaluation of emerging applications of generated language which is still primarily human-dependent or reliant on crude metrics. I propose to reliably and fully automate generative evaluation and disrupt current standards via employing AI critics to evaluate the quality of AI generators. In prior work that won an outstanding paper award at NeurIPS 2021, we introduced evaluation of generated language via measuring distributional differences between machine-generated and human-written text corpora [13, 14]. I am currently working on building on this work to devise measures for close-ended generative tasks such as summarization and translation. I have also explored generative evaluation of specific properties such as information content [2] and acceptability [18] and plan to expand these into newer metrics on additional knowledge criteria. Finally, building on my work on training data quality estimation [3, 16], I plan to design algorithms to select the most valuable evaluation instances for nuanced comparison between systems. Recognizing the key role generative AI is bound to play, I am dedicated to build rigorous evaluation for the safe and responsible deployment of language technologies.

**Training AI for Contextual Cognition and Social Commonsense Reasoning** Human-like performance in AI is only possible when it can handle phenomena such as subjectivity, ambiguity and common sense (physical and social) along with a recognition of contextual cues in natural language [11]. I have studied the crucial role of contextual recognition in models towards improved performance under afore-mentioned

phenomena; this context could either be grounded purely in language or in social situation surrounding the language. I have studied linguistic context by modeling ambiguity in model training, which has resulted in enhanced natural language understanding even in LLMs [10, 11]. I have studied the role of social contexts by modeling the identities and beliefs of annotators [15] as well as conversational situations [20]; avoiding such contextual cues can result in harmful social biases in model decisions which are hard to remove in a post-hoc fashion [19]. To this end, my research has resulted in resources and algorithms that have demonstratively led to better decisions in the wild [20]. Furthermore, social contextual reasoning in models is pivotal for responsible AI, especially given the expected proliferation of AI in the years to come.

Language technologies do not exist in isolation because language is inherently social. But how far can these models reason about societal phenomena? Are they subject to similar biases as humans or are they worse? I am currently exploring **how well LLMs understand broad human attitudes towards widespread societal phenomena** such as homelessness in the US, in collaboration with social scientists at USC. Our plan is to expand this study to several macro- and micro-social phenomena such as attitudes towards social media, AI, the LGBTQ population, accounting for geographical and temporal dimensions.

**Provoking Disruption** The current AI paradigm of improving models simply by scaling training data quantity is neither sustainable nor accessible to the broader research community [17]. My future research agenda aims to **disrupt this paradigm instead by bringing into focus the role of data quality** for training models and **enabling data-efficient learning and generation with LLMs**. Furthermore, in departure to coarse aggregate measures for evaluation, my research will **empower comprehensive and rigorous evaluation of LLMs' generative capabilities via emphasizing on evaluation data quality** to shed light on nuanced yet substantial differences between LLMs. My proposed design of practical data-efficient training and data-first evaluation will be impactful both in industrial and academic settings, paving the path to responsible AI and its safe deployment. My focus on training and evaluating with neuro-symbolic knowledge has implications for commons-sense and social reasoning and cognitive capabilities in models, towards achieving human-like performance, while being consequential for ethical AI. These approaches will fundamentally alter the way the community thinks about data in AI. My proposed methods will further disrupt how and when LLMs are deployed in collaboration with humans. The data and tools my lab builds will be open-sourced for encouraging future scientific development.

**Teaching Innovations and Broadening Participation in STEM** The only way to keep pace with the meteoric rise of AI is a dedicated pursuit towards **AI education**. This has been and will continue to be the focus of my teaching agenda and outreach activities, both being intimately tied to my research goals. Concretely, my pursuits are directed towards the **development of the next generation of ethical engineers intimately familiar with not only AI-enabled technology, but also the ethical considerations underlying the creation of these technologies and their fair / dual use**. In my class on Data-Centric NLP taught in Fall 2022, I focused on the ethical risks behind key language and vision technologies, as well as the value of contextualizing the predictions of AI models in the data they were trained on and the biases of the designers. My future teaching plans will continue anchoring technologies in a strong ethical foundation. My current outreach efforts involve teaching these principles to high-school students through partnerships at USC Viterbi's K-12 Outreach and the LA County Office of Education.

Much of the research proposed above will be done at my lab, which comprises 2 PhD and 2 Masters students (**75% female**) and 5 undergraduates (1 female). I have consistently mentored women and students from underrepresented groups in STEM at both graduate and undergraduate levels. My students and I will continue our concerted effort towards building a diverse and inclusive research group, with women and members from historically underrepresented groups in computer science. I am currently applying for an NSF Research Experiences for Undergraduates (REU) program, designed to encourage undergrads from underrepresented groups towards research pursuits.

**Academic History and Awards** I joined USC as an Assistant Professor in the Thomas Lord Department of Computer Science in Fall 2022. In the nine months since, I have received a Young Investigator Research Award from the Allen Institute for AI as well as the WiSE Gabilan Assistant Professorship at USC. I was selected as one of 5 ACL 2022 Young Rising Stars for spotlight keynote presentations. I have authored over 30 peer-reviewed publications in top-tier selective AI venues (Google Scholar h-index = 23, > 4,800 citations till date), 6 of which have been published since joining USC.

## References


[1] Chandra Bhagavatula, Jena D. Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Lianhui Qin, Keisuke Sakaguchi, **Swabha Swayamdipta**, Peter West, and Yejin Choi. 2023. I2D2: Inductive Knowledge Distillation with NeuroLogic and Self-Imitation. In *Proceedings of ACL (To Appear)*.

[2] Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and **Swabha Swayamdipta**. 2023. REV: Information-theoretic evaluation of free-text rationales. In *Proceedings of ACL (To Appear)*.

[3] Kawin Ethayarajh, Yejin Choi, and **Swabha Swayamdipta**. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proc. of ICML*.

[4] Suchin Gururangan, Ana Marasović, **Swabha Swayamdipta**, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proc. of ACL*.

[5] Suchin Gururangan, **Swayamdipta, Swabha**, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proc. of NAACL-HLT*, pages 107–112.

[6] Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and **Swabha Swayamdipta**. 2022. Neurocounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. In *Findings of EMNLP*.

[7] Phillip Howard, Junlin Wang, Vasudev Lal, Gadi Singer, Yejin Choi, and **Swabha Swayamdipta**. 2023. Neurocomparatives: Neuro-symbolic distillation of comparative knowledge. Under Submission.

[8] Ronan Le Bras, **Swabha Swayamdipta**, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proc. of ICML*.

[9] Alisa Liu, Maarten Sap, Ximing Lu, **Swabha Swayamdipta**, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proc. of ACL*, Online.

[10] Alisa Liu, **Swayamdipta, Swabha**, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation. In *Findings of EMNLP*.

[11] Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, **Swabha Swayamdipta**, Noah A. Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. Under Submission.

[12] Haokun Liu, Derek Tam, Muqeeth Mohammed, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than ICL. In *Proc. of NeurIPS*.

[13] Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, **Swabha Swayamdipta**, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. 2022. MAUVE Scores for Generative Models: Theory and Practice.

[14] Krishna Pillutla, **Swabha Swayamdipta**, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*.

[15] Maarten Sap, **Swabha Swayamdipta**, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proc. of NAACL-HLT*.

[16] **Swayamdipta, Swabha**, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proc. of EMNLP*.

[17] Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning.

[18] Sarah Wiegreffe, Jack Hessel, **Swayamdipta, Swabha**, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proc. of NAACL-HLT*.

[19] Xuhui Zhou, Maarten Sap, **Swabha Swayamdipta**, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proc. of EACL*.

[20] Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, **Swabha Swayamdipta**, and Maarten Sap. 2023. COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements. In *Findings of ACL (To Appear)*.