Robustifying NLP with Humans in the Loop

Divyansh Kaushik

November 2022

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:
Zachary C. Lipton (Co-Chair)
Eduard Hovy (Co-Chair)
Graham Neubig (Carnegie Mellon University)
Chenhao Tan (University of Chicago)
Samuel Bowman (New York University)

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in Language and Information Technologies.*

Keywords: Natural Language Processing, Robustness, Humans in the Loop, Counterfactually Augmented Data, Spurious Correlations

To my parents, who made this day possible.

Abstract

Despite machine learning (ML)'s many practical breakthroughs, formidable obstacles obstruct its deployment in consequential applications. Modern ML models have repeatedly been shown to rely on spurious signals, such as surface level textures in images, and to be sensitive to background scenery, even when the task addresses the recognition of foreground objects. In NLP, these issues have emerged as central concerns in the literature on *annotation artifacts* and *bias*. Moreover, while modern ML performs remarkably well on independent and identically distributed (iid) hold-out data, performance often decays catastrophically under both naturally occurring and adversarial distribution shift. We desire decisions to be based on qualifications, not on distant proxies that are spuriously associated with the outcome of interest. Arguably one key distinction of an actual qualification might be that it actually exerts causal influence on the outcome of interest. In this thesis, we make progress towards these goals: in the first part, we scrutinize benchmarks and problem formulation for popular NLP tasks, such as question answering and how models may ignore crucial parts of the input altogether and yet perform well on a held out test set; in the second part, we focus on introducing methods and datasets to train models to be less reliant on spurious correlations by learning from several forms of human feedback (sought via crowdsourcing); in part three we focus on the human workforce as we discuss the ethical tensions posed by the diverse roles played by crowdworkers in NLP research, and discuss the implications of selecting a diverse cohort of crowdworkers on resulting human-in-the-loop feedback.

Acknowledgements

First and foremost, I would like to express my gratitude to my PhD advisors, Ed and Zack for their guidance, support, and encouragement throughout my Master's and my PhD. They both have always been a great inspiration to me and very supportive in everything I do. I'm especially grateful for the trust Ed and Zack put in me from the first time we met, making me a part of Edvisees and ACMI Lab, the freedom they gave me to pursue ambitious ideas, their contagious optimistic attitude, and the many opportunities they presented to me. They encouraged me to think bolder, become stronger at what I do, read more, learn more, and write more. While Zack shot down my terrible ideas, Ed gave me comfort. I couldn't have asked for better advisors. And thanks to the other members on my thesis committee, Drs. Neubig, Bowman, and Tan, who helped me think through so much of this!

I wouldn't be here without all the hard work my parents put in to get me the best education at all levels even if it meant they had to work double overtime, and my grandfather—who by the way would've sent sweets to the whole town if he was alive today—who cultivated my intellectual curiosity. You're now obliged to call me Dr. Kaushik as you worked harder for this moment than I did. There are two other people who had a big hand in getting me to CMU: Dr. Karamjit Kaur who taught me to question everything and planted the seed of going to grad school in me and Dr. Deepak Garg who made sure I had everything I needed to succeed in my undergraduate studies.

At CMU, I've had the privilege of working with some of the smartest minds in the world: Prof. Sina Fazelpour (though I wish that Liverpool never win the Premier League again), Prof. Maria De Arteaga, Prof. Molly Roberts, Prof. Luke Sanford, and Prof. Alex London. Thanks to Scott Yih and Douwe Kiela (along with Stephen Roller and the whole Dynabench team) for the wonderful time I had at Meta, Bilal Zafar, Matt Lease and Krishnaram Kenthapadi for hosting me at Amazon, and Jason Wu and Caiming Xiong for hosting me at Salesforce. I learned a lot from you. I want to thank all the students at Carnegie Mellon who I had the privilege to mentor at some point: Anurag Katakkar, Clay Yoo, Weiqin Wang, and Amrith Setlur. Special thanks to Shantanu and Saurabh for joining me in exploring all the restaurants in town, and to Laura Winter who made sure Zack did things on time. And generally, thanks to everybody in Edvisees and ACMI Lab who made these five and a half years so memorable and gave me so much to learn from them. You all continue to amaze me. There's a part of me that feels sad that I won't be able to get first dibs on the lab meeting tacos but thanks Nil-Jana for getting amazing food every week.

I've learned that completing a PhD is no small feat, and I could not have done it without the constant encouragement from those closest to me. My roommate, Sanket Mehta, with whom I've had thousands of conversations about how so many NLP researchers don't seem to know the difference between standard error and standard deviation, and shared the 20 years of COVID-19 at home cooking some of the best food I've made (thanks for being the guinea pig as I perfected my recipes), and my roommates for some of the last so many years, Raghuram MA, Ankit Shah, and Kundan Krishna. Tejas Khot and Kate Olszewska, thank you so much for everything! Our intellectual (and shitposting) discussions have shaped my thought so much. And finally, Nidhi Vyas, Shubham Garg, Purvanshi Mehta, Rishiraj Shanmuganandan, and Vasu Sharma, thanks for being there.

I had the best time of any student at Carnegie Mellon and a lot of that has to do with the opportunities I had outside of research. Thank you to all the CMU graduate students who I had the honor and privilege of serving as their Graduate Student Assembly President. Serving on GSA Executive Committee gave me so much more than I could have ever asked for. To all my amazing friends who had to bear my rants about all that is wrong and how to address complex societal problems. Special thanks to my partners in crime, Kristen Allen, Surya Aggarwal, Josh Gyory, Michael Sinko, McKenna McCall, and Evan Feder for all that you've taught me and for taking my 10x rants. Surya, thank you for introducing me to advocacy and mentoring me when I was a mere amateur. Benjamin Lane and Katie Jordan, my partners in crime—may the GRAD Caucus live forever! Thank you to all the partners in university administration, especially Provost Jim Garrett and Dean of Students Gina Casalegno, who worked with me to make the lives of students better. This won't be complete without mentioning all the phenomenal people I've had the opportunity of working with on advancing innovation. Ami Shah, Josh Briesblatt, Sara Barber, Cate Johnson, Heather Painter, Alan McQuinn, Hannah Bristol, Meghan Pearce, Tim McNulty, Debbie Tekavec, Toby Smith, Ishan Sharma, Kumar Garg, Josh Schoop, and Dan Correa, I deeply admire all you do daily. Thank you for teaching me so much.

I would like to express my appreciation to all the research participants on Amazon Mechanical Turk who helped make all this research possible, all the reviewers who helped me make my papers better, Pricewaterhouse Coopers for funding me for the first two years of grad school, Amazon AI for granting me the Graduate Research Fellowship that allowed me to continue working on projects I wanted to, and the Siebel Family Foundation for giving me the Siebel Scholarship this year. Finally, thanks to Independent Brewing Company for the best vegan crab rangoons (Pete, bring them back!) and dealing with me as I pulled out my laptop and started to write papers while sitting there, and Pigeon Bagels for making the best bagels in town!

Finally, I want to thank Prof. Jhilik Bhattacharya who told me during my undergraduate studies that I will never be a successful NLP researcher. Thanks for giving me that C in undergrad NLP, it's kept me going all these years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

What makes a document's sentiment positive? What makes a loan applicant creditworthy? What makes a job candidate qualified? When does a photograph truly depict a dolphin? Moreover, what does it mean for a feature to be relevant to such a determination?

Statistical learning offers one framework for approaching these questions. First, we swap out the semantic question for a more readily answerable associative question. For example, instead of asking *what conveys a document's sentiment*, we recast the question as *which documents are likely to be labeled as positive (or negative)?* Then, in this associative framing, we interpret as *relevant*, those features that are most *predictive* of the label. However, despite the rapid adoption and undeniable commercial success of associative learning, this framing seems unsatisfying.

While deep learning models demonstrate remarkable predictive performance on a wide variety of tasks when evaluated on i.i.d. holdout data, researchers have come to recognize that these models often degrade catastrophically when faced with distribution shift. While papers tackling this problem seldom make their assumptions explicit, they typically proceed under the assumption that the labeling function is deterministic (there is one right answer), and that the *covariate shift assumption* [Shimodaira, 2000] holds (the labeling function $f(x)$ is invariant across domains). Consequently, one might hope to find a single predictor $\hat{f}$ that performs well across a wide variety of real-world scenarios. Note that, absent these assumptions, there is, in general, no reason to believe that any fixed predictor can be expected to perform well across out of distribution. For example, faced with label shift [Lipton et al., 2018] or concept drift, the likelihood of each label under different distributions can vary [Quionero-Candela et al., 2009], necessitating adapting the predictor to each environment.

Even under these strict assumptions, models that perform well on i.i.d. on holdout data might degrade out of distribution. Most importantly, the support (subset of input space with non-zero probability mass) for the distribution from which our dataset is sampled may only be a small subset of the domain of interest. For example, in question answering (QA), we might fear that whole categories of questions that we might encounter in the wild would seldom or never be posed by crowdworkers. Thus, one could produce a predictor that often agrees with the label on the test set but often disagrees on data encountered in the wild. No principle prohibits such an unfortunate scenario.

Thus, alongside deep learning's predictive wins, critical questions have piled up concerning *spurious patterns*, *artifacts*, *robustness*, and *discrimination*, that the purely associative perspective appears ill-equipped to answer. In NLP specifically, an emerging line of work, has catalogued such vulnerabilities. For example, Poliak et al. [2018] demonstrated that classifiers trained for natural language inference (NLI) may depend on hypotheses alone (disregarding premises), while achieving good in-domain performance but vulnerable to catastrophic failure out of domain. In this work, we demonstrate that sentiment classifiers trained on movie reviews often rely on genre for predicting sentiment (*horror* bad, *romance* good). We also show that state-of-the-art QA models may rely on questions (or passages) alone, without loss of performance in domain. Jia and Liang [2017] found that state-of-the-art question answering models trained on SQuAD [Rajpurkar et al., 2016] are vulnerable to synthetic transformations to input passages that do not alter the correct answer.

However, papers seldom make clear what, if anything, *spuriousness* means within the standard supervised learning framework. ML systems are trained to exploit the mutual information between features and a label to make accurate predictions. The standard statistical learning toolkit does not offer a conceptual distinction between spurious and non-spurious associations. So how then, can we exploit the statistical learning framework in a way that can enable us to train predictive models that may generalize well under plausible distribution shifts and what additional challenges we might face?

1.1 Thesis Overview

In part one of this thesis, we scrutinize benchmarks and problem formulation for popular NLP tasks, such as question answering [Kaushik and Lipton, 2018]. Presumably, for a task such as passage based question answering, a model must combine information from both questions and passages to predict corresponding answers. However, we find that models may be able to answer questions by looking at passages alone. We further highlight that while modern ML systems perform remarkably well on independent and identically distributed (iid) holdout data, performance often decays catastrophically under both naturally occurring and adversarial distribution shift due to their reliance on spurious correlations that may not transport out of domain.

In the second part, we study and propose several human-in-the-loop approaches to train ML models that are less reliant on spurious correlations. In one approach, we propose counterfactually manipulating documents via humans-in-the-loop [Kaushik et al., 2020]. We employ crowd workers not to label documents, but rather to edit them, manipulating the text to make a targeted (counterfactual) class applicable. For instance, for sentiment analysis, we direct the worker to revise this negative movie review to make it positive, without making any gratuitous changes. We show that by intervening only upon the factor of interest, we disentangle the spurious and non-spurious associations, yielding classifiers that hold up better when spurious associations do not transport out of domain. We further introduce a toy analog based on linear Gaussian models, observing interesting relationships between causal models, measurement noise, out-of-domain generalization, and reliance on spurious signals [Kaushik et al., 2021b]. Our analysis provides some insights that help to explain the efficacy of CAD. We further investigate whether similar gains in out-of-domain performance can also be achieved by training models with other forms of human feedback, such as feature feedback [Katakkar et al., 2021] or training sets constructed via an adversarial data collection process (ADC) [Kaushik et al., 2021a].

Finally, having demonstrated the effectiveness of several human-in-the-loop approaches, in the third part of this thesis, we look at two practical challenges associated with human-in-the-loop ML research. First, we look at the diversity of the workforce that offers this feedback. We investigate whether diversity among workers who provide this feedback might make a difference in what feedback is received. Our analysis reveals insights into how crowdworkers' demographics play a critical role in the feedback they provide and highlight the need to explicitly consider diversity of crowdworkers as a critical ingredient in any human-in-the-loop study. Second, we highlight how with the increase in creative uses of crowdworkers in ML research (in general) and in NLP research (in particular), the line between laborer and human subject has blurred. We investigate the appropriate designation of ML crowdsourcing studies, focusing our inquiry on natural language processing to expose unique challenges for research oversight [Kaushik et al., 2022]. Crucially, under the U.S. Common Rule, these judgments hinge on determinations of *aboutness*, concerning both whom (or what) the collected data is about and whom (or what) the analysis is about. We highlight two challenges posed by ML: the same set of workers can serve multiple roles and provide many sorts of information; and ML research tends to embrace a dynamic workflow, where research questions are seldom stated ex ante and data sharing opens the door for future studies to aim questions at different targets. Our analysis exposes a potential loophole in the Common Rule, where researchers can elude research ethics oversight by splitting data collection and analysis into distinct studies. Finally, we discuss several policy recommendations to address these concerns.

Chapter 2

Does Reading Comprehension Require *Reading* Or *Comprehension*?

2.1 Overview

Reading comprehension (RC) has emerged as a popular task in NLP, with researchers proposing various end-to-end deep learning algorithms to push the needle on a variety of benchmarks. As characterized by Hermann et al. [2015], Onishi et al. [2016], unlike prior work addressing question answering from general structured knowledge, RC requires that a model extract information from a given, unstructured passage. It's not hard to imagine how such systems could be useful. In contrast to generic text summarization, RC systems could answer targeted questions about specific documents, efficiently extracting facts and insights.

While many RC datasets have been proposed over the years [Hirschman et al., 1999, Breck et al., 2001, Peñas et al., 2011, Peñas et al., 2012, Sutcliffe et al., 2013, Richardson et al., 2013, Berant et al., 2014], more recently, larger datasets have been proposed to accommodate the data-intensiveness of deep learning. These vary both in the source and size of their corpora and in how they cast the prediction problem—as a classification task [Hill et al., 2016, Hermann et al., 2015, Onishi et al., 2016, Lai et al., 2017, Weston et al., 2016, Miller et al., 2016], span selection [Rajpurkar et al., 2016, Trischler et al., 2017], sentence retrieval [Wang et al., 2007, Yang et al., 2015], or free-form answer generation Nguyen et al. [2016].[1] Researchers have steadily advanced on these benchmarks, proposing myriad neural network architectures aimed at attending to both questions and passages to produce answers.

In this chapter, we argue that amid this rapid progress on empirical benchmarks, crucial steps are sometimes skipped. In particular, we demonstrate that the level of difficulty for several of these tasks is poorly characterized. For example, for many RC datasets, it's not reported, either in the papers introducing the datasets, or in those proposing models, how well one can perform while ignoring either the question or the passage. In other datasets, although the passage might consist of many lines of text, it's not clear how many are actually required to answer the question, e.g., the answer may always lie in the first or the last sentence.

[1] We note several other QA datasets [Yang et al., 2015, Miller et al., 2016, Nguyen et al., 2016, Paperno et al., 2016, Clark and Etzioni, 2016, Lai et al., 2017, Trischler et al., 2017, Joshi et al., 2017] not addressed in this chapter.

We describe several popular RC datasets and models proposed for these tasks, analyzing their performance when provided with question-only (Q-only) or passage-only (P-only) information. We show that on many tasks, the results obtained are surprisingly strong, outperforming many baselines, and sometimes even surpassing the same models, supplied with both questions *and* passages. We note that similar problems were shown for datasets in *visual question answering* by Goyal et al. [2017] and for *natural language inference* by Gururangan et al. [2018], Poliak et al. [2018], Glockner et al. [2018]. Several other papers have discussed the weaknesses of various RC benchamrks [Chen et al., 2016, Lee et al., 2016]. We discuss these studies in the paragraphs introducing the corresponding datasets below.

2.2 Datasets

In the following section, we provide context on each dataset that we investigate and then describe our process for corrupting the data as required by our question- and passage-only experiments.

CBT Hill et al. [2016] prepared a cloze-style (*fill in the blank*) RC dataset by using passages from children's books. In their dataset, each passage consists of 20 consecutive sentences, and each question is the 21st sentence with one word removed. The missing word then serves as the answer. The dataset is split into four categories of answers: Named Entities (NE), Common Nouns (CN), Verbs (V) and Prepositions (P). The training corpus contains over $37,000$ candidates and each question is associated with 10 candidates, POS-matched to the correct answer. The authors established LSTM/embedding-based Q-only baselines but did not present the results obtained by their best model using Q-only or P-only information.

CNN Hermann et al. [2015] introduced the CNN/Daily Mail datasets containing more than 1 million news articles, each associated with several highlight sentences. Also adopting the cloze-style dataset preparation, they remove an entity (answer) from a highlight (question). They anonymize all entities to ensure that models rely on information contained in the passage, vs memorizing characteristics of given entities across examples, and thus ignoring passages. On average, passages contain 26 entities, with over 500 total possible answer candidates. Chen et al. [2016] analyzed the difficulty of the CNN and Daily Mail tasks. They hand-engineered a set of eight features for each entity e (does e occur in the question, in the passage, etc.), showing that this simple classifier outperformed many earlier deep learning results.

Who-did-What Onishi et al. [2016] extracted pairs of news articles, each pair referring to the same events. Adopting the cloze-style, they remove a person's name (the answer) from the first sentence of one article (the question). A model must predict the answer based on the question, together with the other article in the pair (passage). Unlike CNN, Who-did-What does not anonymize entities. On average, each question is associated with 3.5 candidate answers. The authors removed several questions from their dataset to thwart simple strategies such as always predicting the name that occurs most (or first) in the passage.

bAbI Weston et al. [2016] presented a set of 20 tasks to help researchers identify and rectify the failings of their reading comprehension systems. Unlike the datasets discussed so far, the questions in this task are not cloze-style and are synthetically generated using templates. This restricts the diversity in clauses appearing in the passages. Further, this also restricts the dataset vocabulary to just 150 words, in contrast, CNN dataset has a vocabulary made of close to $120,000$ words. Memory Networks with adaptive memory, n-grams and non-linear matching were shown to obtain 100% accuracy on 12 out of 20 bAbI tasks. We note that Lee et al. [2016] previously identified that bAbI tasks might fall short as a measure of "AI-complete question answering", proposing two models based on *tensor product representations* that achieve 100% accuracy on many bAbI tasks.

SQuAD Rajpurkar et al. [2016] released the Stanford Question Answering Dataset (SQuAD) containing over $100,000$ crowd-sourced questions addressing 536 passages. Each question is associated with a paragraph (passage) extracted from an article. These passages are shorter than those in CNN and Who-did-What datasets. Models choose answers by selecting (varying-length) spans from these passages.

Generating Corrupt Data To void any information in either the questions or the passages, while otherwise leaving each architecture intact, we create corrupted versions of each dataset by assigning either questions randomly, while preserving the correspondence between passage and answer, or by randomizing the passage. For tasks where question-answering requires selecting spans or candidates from the passage, we create passages that contain the candidates in random locations but otherwise consist of random gibberish.

2.3 Models

In our investigations of the various RC benchmarks, we rely upon the following three QA models: key-value memory networks, gated attention readers, and QA nets. Although space constraints preclude a full discussion of each architecture, we provide references to the source papers and briefly discuss any implementation decisions necessary to reproduce our results.

Key-Value Memory Networks We implement a Key-Value Memory Network (KV-MemNet) [Miller et al., 2016], applying it to bAbI and CBT. KV-MemNets are based on Memory Networks [Sukhbaatar et al., 2015], shown to perform well on both datasets. For bAbI tasks, the keys and values both encode the passage as a bag-of-words (BoW). For CBT, the key is a BoW-encoded 5-word window surrounding a candidate answer and the value is the candidate itself. We fixed the number of hops to 3 and the embedding size to 128.

Gated Attention Reader Introduced by Dhingra et al. [2017], the Gated Attention Reader (GAR)[2] performs multiple hops over a passage, like MemNets. The word representations are refined over each hop and are mapped by an attention-sum module Kadlec et al. [2016] to a

[2] https://github.com/bdhingra/ga-reader

	bAbI Tasks 1-10									
Dataset	1	2	3	4	5	6	7	8	9	10
True dataset	**100%**	**100%**	39%	**100%**	**99%**	**100%**	**94%**	**97%**	**99%**	**98%**
Question only	18%	17%	22%	22%	34%	50%	48%	34%	64%	44%
Passage only	53%	86%	**60%**	59%	31%	48%	85%	79%	63%	47%
$\Delta(min)$	-47	-14	$+21$	-41	-65	-52	-9	-18	-35	-51
	bAbI Tasks 11-20									
	11	12	13	14	15	16	17	18	19	20
True dataset	**94%**	**100%**	**94%**	**96%**	**100%**	**48%**	**57%**	**93%**	**30%**	**100%**
Question only	17%	15%	18%	18%	34%	26%	48%	91%	10%	70%
Passage only	71%	74%	**94%**	50%	64%	**47%**	48%	53%	21%	**100%**
$\Delta(min)$	-23	-26	0	-46	-36	-1	-9	-2	-9	0

Table 2.1: Accuracy on bAbI tasks using our implementation of the Key-Value Memory Networks

probability distribution over the candidate answer set in the last hop. The model nearly matches best-reported results on many cloze-style RC datasets, and thus we apply it to *Who-did-What, CNN, CBT-NE* and *CBT-CN*.

QA Net Introduced by Yu et al. [2018], the QA-Net[3] was recently demonstrated to outperform all previous models on the SQuAD dataset[4]. Passages and questions are passed as input to separate encoders consisting of depth-wise separable convolutions and global self-attention. This is followed by a passage-question attention layer, followed by stacked encoders. The outputs from these encoders are used to predict an answer span inside the passage.

2.4 Experimental Results

bAbI tasks Table 2.1 shows the results obtained by a Key-Value Memory Network on bAbI tasks by nullifying the information present in either questions or passages. On tasks $2, 7, 13$ and 20, P-only models obtain over 80% accuracy with questions randomly assigned. Moreover, on tasks 3, 13, 16, and 20, P-only models match performance of those trained on the full dataset. On task 18, Q-only models achieve an accuracy of 91%, nearly matching the best performance of 93% achieved by the full model. These results show that some of bAbI tasks are easier than one might think.

[3] We use the implementation available at https://github.com/NLPLearn/QANet

[4] When these experiments were conducted in 2018, an ensemble of QA-Net models was at the top of the leader board. A single QA-Net was ranked 4th.

Task	Full	Q-only	P-only	$\Delta(min)$
Key-Value Memory Networks				
CBT-NE	**35.0%**	29.1%	24.1%	-5.9
CBT-CN	**37.6%**	32.4%	24.4%	-5.2
CBT-V	52.5%	**55.7%**	36.0%	$+3.2$
CBT-P	55.2%	**56.9%**	30.1%	$+1.7$
Gated Attention Reader				
CBT-NE	**74.9%**	50.6%	40.8%	-17.5
CBT-CN	**70.7%**	54.0%	36.7%	-16.7
CNN	**77.8%**	25.6%	38.3%	-39.5
WdW	**67.0%**	41.8%	52.2%	-14.8
WdW-R	**69.1%**	50.0%	50.6%	-15.6

Table 2.2: Accuracy on various datasets using KV-MemNets (window memory) and GARs

Task	Complete passage	Last sentence
CBT-NE	22.6%	**22.8%**
CBT-CN	**31.6%**	24.8%
CBT-V	**48.8%**	45.0%
CBT-P	34.1%	**37.9%**

Table 2.3: Accuracy on CBT tasks using KV-MemNets (sentence memory) varying passage size.

Metric	Full	Q-only	P-only	$\Delta(min)$
EM	**70.7%**	0.6%	10.9%	-59.8
F1	**79.1%**	4.0%	14.8%	-64.3

Table 2.4: Performance of QANet on SQuAD

Children's Books Test On the NE and CN CBT tasks, Q-only KV-MemNets obtain an accuracy close to the *full* accuracy and on the Verbs (V) and Prepositions (P) tasks, Q-only models outperform the full model (Table 2.2). Q-only Gated attention readers reach accuracy of 50.6% and 54% on Named Entities (NE) and Common Nouns (CN) tasks, respectively, while P-only models reach accuracies of 40.8% and 36.7%, respectively. We note that our models can outperform 16 of the 19 reported results on the NE task in Hill et al. [2016] using Q-only information. Table 2.3 shows that if we make use of just last sentence instead of all 20 sentences in the passage, our sentence memory based KV-MemNet achieve comparable or better performance *w.r.t* the *full* model on most subtasks.

CNN Table 2.2, shows the performance of Gated Attention Reader on the CNN dataset. Q-only and P-only models obtained 25.6% and 38.3% accuracies respectively, compared to 77.8% on the true dataset. This drop in accuracy could be due to the anonymization of entities which prevents models from building entity-specific information. Notwithstanding the deficiencies noted by Chen et al. [2016], we found that out CNN, out all the cloze-style RC datasets that we evaluated, appears to be the most carefully designed.

Who-did-What P-only models achieve greater than 50% accuracy in both the strict and relaxed setting, reaching within 15% of the accuracy of the *full* model in the strict setting. Q-only models also achieve 50% accuracy on the relaxed setting while achieving an accuracy of 41.8% on the strict setting. Our P-only model also outperforms all the suppressed baselines and 5 additional baselines reported by Onishi et al. [2016]. We suspect that the models memorize attributes of specific entities, justifying the entity-anonymization used by Hermann et al. [2015] to construct the CNN dataset.

SQuAD Our results suggest that SQuAD is an unusually carefully-designed and challenging RC task. The span selection mode of answering requires that models consider the passage thus the abysmal performance of the Q-only QANet (Table 2.4). Since SQuAD requires answering by span selection, we construct Q-only variants here by placing answers from all relevant questions in random order, filling the gaps with random words. Moreover, Q-only and P-only models achieve F1 scores of only 4% and 14.8% resp. (Table 2.4), significantly lower than 79.1 on the proper task.

2.5 Discussion

We briefly discuss our findings, offer some guiding principles for evaluating new benchmarks and algorithms, and speculate on why some of these problems may have gone under the radar. Our goal is not to blame the creators of past datasets but instead to support the community by offering practical guidance for future researchers.

Provide rigorous RC baselines Published RC datasets should contain reasonable baselines that characterize the difficulty of the task, and specifically, the extent to which questions and passages are essential. Moreover, follow-up papers reporting improvements ought to report performance both on the full task and variations omitting questions and passages. While many proposed technical innovations purportedly work by better matching up information in questions and passages, absent these baselines one cannot tell whether gains come for the claimed reason or if the models just do a better job of passage classification (disregarding questions).

Test that *full context* is essential Even on tasks where both questions and passages are required, problems might appear harder than they really are. On first glance the the length-20 passages in CBT, might suggest that success requires reasoning over all 20 sentences to identify the correct answer to each question. However, it turns out that for some models, comparable performance can be achieved by considering only the last sentence. We recommend that researchers provide reasonable ablations to characterize the amount of context that each model truly requires.

Caution with cloze-style RC datasets We note that cloze-style datasets are often created programatically. Thus it's possible for a dataset to be produced, published, and incorporated into many downstream studies, all without many person-hours spent manually inspecting the data. We speculate that, as a result, these datasets tend be subject to less contemplation of what's involved in answering these questions and are therefore especially susceptible to the sorts of overlooked weaknesses described in our study.

A note on publishing incentives We express some concern that the recommended experimental rigor might cut against current publishing incentives. We speculate that papers introducing datasets may be more likely to be accepted at conferences by omitting unfavorable ablations than by including them. Moreover, with reviewers often demanding *architectural novelty*, methods papers may find an easier path to acceptance by providing unsubstantiated stories about the *reasons* why a given architecture works than by providing rigorous ablation studies stripping out spurious explanations and unnecessary model components. For more general discussions of misaligned incentives and empirical rigor in machine learning research, we point the interested reader to Lipton and Steinhardt [2018] and Sculley et al. [2018].

On the name "reading comprehension" Perhaps a first step towards solidifying the intellectual foundations in this area might be to re-evaluate the name "reading comprehension", which strikes us as unjustifiably anthropomorphic. A more neutral name like "passage-based question answering", while perhaps less marketable, more accurately characterizes the tasks. After all, the the systems are concerned with supervised learning, minimizing the error by exploiting correlations in the training data. As demonstrated by Jia and Liang [2017], the resulting systems, which are susceptible to simple adversarial attacks, rely heavily on the superficial statistics of the training data.

Chapter 3

Learning The Difference That Makes A Difference With Counterfactually Augmented Data

3.1 Overview

Alongside deep learning's predictive wins, critical questions have piled up concerning *spurious patterns*, *artifacts*, *robustness*, and *discrimination*, that the purely associative perspective appears ill-equipped to answer. And yet, researchers struggle to articulate precisely why models *should not* rely on such patterns. In natural language processing (NLP), these issues have emerged as central concerns in the literature on *annotation artifacts* and *societal biases*. Across myriad tasks, researchers have demonstrated that models tend to rely on *spurious* associations [Poliak et al., 2018, Gururangan et al., 2018, Kaushik and Lipton, 2018, Kiritchenko and Mohammad, 2018]. However, papers seldom make clear what, if anything, *spuriousness* means within the standard supervised learning framework. ML systems are trained to exploit the mutual information between features and a label to make accurate predictions. The standard statistical learning toolkit does not offer a conceptual distinction between spurious and non-spurious associations.

Causality, however, offers a coherent notion of spuriousness. Spurious associations owe to confounding rather than to a (direct or indirect) causal path. We might consider a factor of variation to be spuriously correlated with a label of interest if intervening upon it would not impact the applicability of the label or vice versa. In this chapter, we introduce a human-in-the-loop system for counterfactually manipulating documents. Our hope is that by intervening only upon the factor of interest, we might disentangle the spurious and non-spurious associations, yielding classifiers that hold up better when spurious associations do not transport out of domain. We employ crowd workers not *to label* documents, but rather *to edit* them, manipulating the text to make a targeted (counterfactual) class applicable. For sentiment analysis, we direct the worker to *revise this negative movie review to make it positive, without making any gratuitous changes*. We might regard the second part of this directive as a least action principle, ensuring that we perturb only those spans necessary to alter the applicability of the label. For NLI, a 3-class classification task (*entailment, contradiction, neutral*), we ask the workers to modify the premise

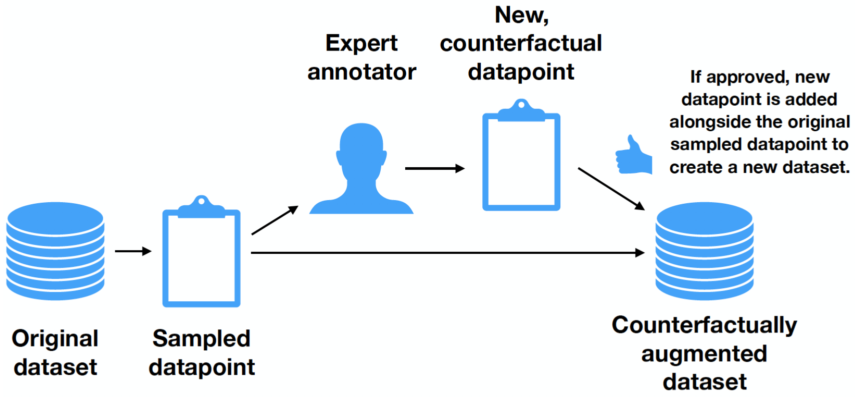

New, counterfactual datapoint
Expert annotator
If approved, new datapoint is added alongside the original sampled datapoint to create a new dataset.
Original dataset
Sampled datapoint
Counterfactually augmented dataset

Figure 3.1: Pipeline for collecting and leveraging counterfactually-altered data

while keeping the hypothesis intact, and vice versa, collecting edits corresponding to each of the (two) counterfactual classes. Using this platform, we collect thousands of counterfactually-manipulated examples for both sentiment analysis and NLI, extending the IMDb [Maas et al., 2011] and SNLI [Bowman et al., 2015] datasets, respectively. The result is two new datasets (each an extension of a standard resource) that enable us to both probe fundamental properties of language and train classifiers less reliant on spurious signal. We show that classifiers trained on original IMDb reviews fail on counterfactually-revised data and vice versa. We further show that spurious correlations in these datasets are even picked up by linear models. However, augmenting the revised examples (creating counterfactually augmented data (CAD)) breaks up these correlations (e.g., genre ceases to be predictive of sentiment).

Following this, we make some initial attempts towards explaining CAD's efficacy and answer certain questions: What is the assumed causal structure underlying settings where CAD might be effective? What are the principles underlying its out-of-domain benefits? Must humans *really* intervene, or could automatic feature attribution methods, e.g., attention [DeYoung et al., 2020], or cheaper feedback mechanisms, e.g., feature feedback [Zaidan et al., 2007], produce similar results?

3.2 Related Work

Several papers demonstrate cases where NLP systems appear not to learn what humans consider to be *the difference that makes the difference*. For example, otherwise state-of-the-art models have been shown to be vulnerable to synthetic transformations such as distractor phrases [Jia and Liang, 2017, Wallace et al., 2019a], to misclassify paraphrased task [Iyyer et al., 2018, Pfeiffer et al., 2019] and to fail on template-based modifications [Ribeiro et al., 2018]. Glockner et al. [2018] demonstrate that simply replacing words by synonyms or hypernyms, which should not alter the applicable label, nevertheless breaks ML-based NLI systems. Gururangan et al. [2018] and Poliak et al. [2018] show that classifiers correctly classified the hypotheses alone in about 69% of SNLI corpus. They further discover that crowd workers adopted specific annotation strategies and heuristics for data generation. Chen et al. [2016] identify similar issues exist with automatically-constructed benchmarks for question-answering [Hermann et al., 2015]. Kaushik and Lipton [2018] discover that reported numbers in question-answering benchmarks could often be achieved by the same models when restricted to be blind either to the question or to the passages. Dixon et al. [2018], Zhao et al. [2018] and Kiritchenko and Mohammad [2018] showed how imbalances in training data lead to unintended bias in the resulting models, and, consequently, potentially unfair applications. Shen et al. [2018] substitute words to test the behavior of sentiment analysis algorithms in the presence of stylistic variation, finding that similar word pairs produce significant differences in sentiment score.

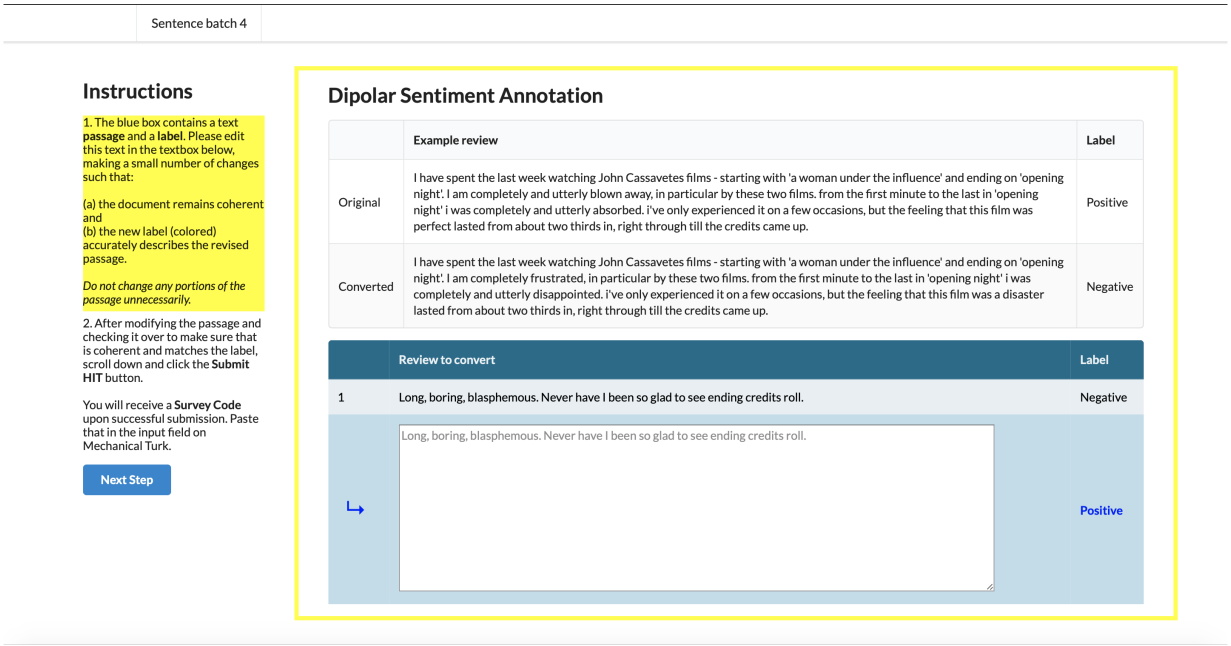
Several papers explore richer feedback mechanisms for classification. Some ask annotators to highlight *rationales*, spans of text indicative of the label [Zaidan et al., 2007, Zaidan and Eisner, 2008, Poulis and Dasgupta, 2017]. For each document, Zaidan et al. remove the *rationales* to generate *contrast* documents, learning classifiers to distinguish original documents from their *contrasting* counterparts. While this feedback is easier to collect than ours, how to leverage it for training deep NLP models, where features are not neatly separated, remains less clear.

Lu et al. [2018] programmatically alter text to invert gender bias and combined the original and manipulated data yielding gender-balanced dataset for learning word embeddings. In the simplest experiments, they swap each gendered word for its other-gendered counterpart. For example, *the doctor ran because he is late* becomes *the doctor ran because she is late*. However, they do not substitute names even if they co-refer to a gendered pronoun. Building on their work, Zmigrod et al. [2019] describe a data augmentation approach for mitigating gender stereotypes associated with animate nouns for morphologically-rich languages like Spanish and Hebrew. They use a Markov random field to infer how the sentence must be modified while altering the grammatical gender of particular nouns to preserve morpho-syntactic agreement. In contrast, Maudslay et al. [2019] describe a method for probabilistic automatic in-place substitution of gendered words in a corpus. Unlike Lu et al., they propose an explicit treatment of first names by pre-defining name-pairs for swapping, thus expanding Lu et al.'s list of gendered word pairs significantly.

A growing body of work has also looked at reducing reliance on spurious correlations by exploiting the stability of relationships between the target variable and its (graph) neighbors. Peters et al. [2016] propose *invariant causal prediction* to obtain a causal predictor from multiple datasets. Ghassami et al. [2017] discuss a similar approach but do not assume that the exogenous noise of the target variable stays fixed among environments. They also demonstrate the benefits of their approach (compared to Peters et al. [2016]) in identifying all direct ancestors of the target variable. Arjovsky et al. [2019] propose *invariant risk minimization*, with the goal of learning a data representation such that the optimal predictor is shared across environments.

3.3 Data Collection

We use Amazon's Mechanical Turk crowdsourcing platform to recruit editors to revise each document. To ensure high quality of the collected data, we restricted the pool to U.S. residents that had already completed at least 500 HITs and had an over 97% HIT approval rate. For each HIT, we conducted pilot tests to identify appropriate compensation per assignment, receive feedback from workers and revise our instructions accordingly. A total of 713 workers contributed throughout the whole process, of which 518 contributed edits reflected in the final datasets.



Instructions

1. The blue box contains a text **passage** and a **label**. Please edit this text in the textbox below, making a small number of changes such that:

(a) the document remains coherent and
(b) the new label (colored) accurately describes the revised passage.

Do not change any portions of the passage unnecessarily.

2. After modifying the passage and checking it over to make sure that is coherent and matches the label, scroll down and click the **Submit HIT** button.

You will receive a **Survey Code** upon successful submission. Paste that in the input field on Mechanical Turk.

Dipolar Sentiment Annotation

	Example review	Label
Original	I have spent the last week watching John Cassavetes films - starting with 'a woman under the influence' and ending on 'opening night'. I am completely and utterly blown away, in particular by these two films. from the first minute to the last in 'opening night' i was completely and utterly absorbed. i've only experienced it on a few occasions, but the feeling that this film was perfect lasted from about two thirds in, right through till the credits came up.	Positive
Converted	I have spent the last week watching John Cassavetes films - starting with 'a woman under the influence' and ending on 'opening night'. I am completely frustrated, in particular by these two films. from the first minute to the last in 'opening night' i was completely and utterly disappointed. i've only experienced it on a few occasions, but the feeling that this film was a disaster lasted from about two thirds in, right through till the credits came up.	Negative

	Review to convert	Label
1	Long, boring, blasphemous. Never have I been so glad to see ending credits roll.	Negative
↳	Long, boring, blasphemous. Never have I been so glad to see ending credits roll.	Positive

Figure 3.2: Annotation platform for collecting counterfactually annotated data for sentiment analysis

Sentiment Analysis The original IMDb dataset consists of $50k$ reviews divided equally across train and test splits. To keep the task of editing from growing unwieldy, we filter out the longest 20% of reviews, leaving $20k$ reviews in the train split from which we randomly sample $2.5k$ reviews, enforcing a $50{:}50$ class balance. Following revision by the crowd workers, we partition this dataset into train/validation/test splits containing 1707, 245 and 488 examples, respectively. We present each review to two workers, instructing them to revise the review such that (a) the counterfactual label applies; (b) the document remains coherent; and (c) no unecessary modifications are made.

Over a four week period, we manually inspected each generated review and rejected the ones that were outright wrong (sentiment was still the same or the review was a spam). After review, we rejected roughly 2% of revised reviews. For 60 original reviews, we did not approve any among the counterfactually-revised counterparts supplied by the workers. To construct the new dataset, we chose one revised review (at random) corresponding to each original review. In qualitative analysis, we identified eight common patterns among the edits (Table 3.2).

By comparing original reviews to their counterfactually-revised counterparts we gain insight into which aspects are causally relevant. To analyze inter-editor agreement, we mark indices corresponding to replacements and insertions, representing the edits in each original review by a binary vector. Using these representations, we compute the Jaccard similarity between the two reviews (Table 3.1), finding it to be negatively correlated with the length of the review.

Natural Language Inference Unlike sentiment analysis, SNLI is 3-way classification task, with inputs consisting of two sentences, a *premise* and a *hypothesis* and the three possible labels being *entailment*, *contradiction*, and *neutral*. The label is meant to describe the relationship between the facts stated in each sentence. We randomly sampled 1750, 250, and 500

Table 3.1: Percentage of inter-editor agreement for counterfactually-revised movie reviews

	Number of tokens							
Type	0-50	51-100	101-150	151-200	201-250	251-300	301-329	Full
Replacement	35.6	25.7	20.0	17.2	15.0	14.8	11.6	19.3
Insertion	27.7	20.8	14.4	12.2	11.0	11.5	07.6	14.3
Combined	41.6	32.7	26.3	23.4	21.6	20.3	16.2	25.5

pairs from the train, validation, and test sets of SNLI respectively, constraining the new data to have balanced classes. In one HIT, we asked workers to revise the hypothesis while keeping the premise intact, seeking edits corresponding to each of the two counterfactual classes. We refer to this data as Revised Hypothesis (RH). In another HIT, we asked workers to revise the original premise, while leaving the original hypothesis intact, seeking similar edits, calling it Revised Premise (RP).

Following data collection, we employed a different set of workers to verify whether the given label accurately described the relationship between each premise-hypothesis pair. We presented each pair to three workers and performed a majority vote. When all three reviewers were in agreement, we approved or rejected the pair based on their decision, else, we verified the data ourselves. Finally, we only kept premise-hypothesis pairs for which we had valid revised data in both RP and RH, corresponding to both counterfactual labels. As a result, we discarded $\approx 9\%$ data. RP and RH, each comprised of 3332 pairs in train, 400 in validation, and 800 in test, leading to a total of 6664 pairs in train, 800 in validation, and 1600 in test in the revised dataset. In qualitative analysis, we identified some common patterns among hypothesis and premise edits (Table 3.3, 3.4).

We collected all data after IRB approval and measured the time taken to complete each HIT to ensure that all workers were paid more than the federal minimum wage. During our pilot studies, workers spent roughly 5 minutes per revised review, and 4 minutes per revised sentence (for NLI). We paid workers $\$0.65$ per revision, and $\$0.15$ per verification, totalling $\$10778.14$ for the study.

3.4 Models

Our experiments rely on the following five models: Support Vector Machines (SVMs), Naïve Bayes (NB) classifiers, Bidirectional Long Short-Term Memory Networks [Bi-LSTMs; Graves and Schmidhuber, 2005], ELMo models with LSTM, and fine-tuned BERT models [Devlin et al., 2019]. For brevity, we discuss only implementation details necessary for reproducibility.

Standard Methods We use `scikit-learn` [Pedregosa et al., 2011] implementations of SVMs and Naïve Bayes for sentiment analysis. We train these models on TF-IDF bag of words feature representations of the reviews. We identify parameters for both classifiers using grid search conducted over the validation set.

Bi-LSTM When training Bi-LSTMs for sentiment analysis, we restrict the vocabulary to

Table 3.2: Most prominent categories of edits performed by humans for sentiment analysis (Original/Revised, in order). Red spans were replaced by Blue spans.

Types of Revisions	Examples
Recasting *fact* as *hoped for*	The world of Atlantis, hidden beneath the earth's core, is fantastic The world of Atlantis, hidden beneath the earth's core is **supposed** to be fantastic
Suggesting sarcasm	thoroughly captivating **thriller-drama, taking a deep and realistic** view thoroughly mind numbing **"thriller-drama", taking a "deep" and "realistic" (who are they kidding?)** view
Inserting modifiers	The presentation of simply Atlantis' landscape and setting The presentation of Atlantis' **predictable** landscape and setting
Replacing modifiers	"Election" is a highly fascinating and thoroughly **captivating** thriller-drama "Election" is a highly expected and thoroughly **mind numbing** "thriller-drama"
Inserting phrases	Although there's hardly any action, the ending is still shocking. Although there's hardly any action **(or reason to continue watching past 10 minutes)**, the ending is still shocking.
Diminishing via qualifiers	which, while usually containing some reminder of harshness, become **more and more intriguing**. which, usually containing some reminder of harshness, became **only slightly more intriguing**.
Differing perspectives	Granted, **not all of the story makes full sense**, but the film doesn't feature any amazing new computer-generated visual effects. Granted, **some of the story makes sense**, but the film doesn't feature any amazing new computer-generated visual effects.
Changing ratings	one of the worst ever scenes in a sports movie. **3 stars out of 10**. one of the wildest ever scenes in a sports movie. **8 stars out of 10**.

the most frequent $20k$ tokens, replacing out-of-vocabulary tokens by UNK. We fix the maximum input length at 300 tokens and pad smaller reviews. Each token is represented by a randomly-initialized 50-dimensional embedding. Our model consists of a bidirectional LSTM (hidden dimension 50) with recurrent dropout (probability 0.5) and global max-pooling following the embedding layer. To generate output, we feed this (fixed-length) representation through a fully-connected hidden layer with ReLU [Nair and Hinton, 2010] activation (hidden dimension 50),

Table 3.3: Analysis of edits performed by humans for NLI hypotheses. P denotes *Premise*, OH denotes *Original Hypothesis*, and NH denotes *New Hypothesis*.

Types of Revisions	Examples
Modifying/removing actions	**P:** A young dark-haired woman crouches on the banks of a river while washing dishes. **OH:** A woman washes dishes in the river **while camping**. (Neutral) **NH:** A woman washes dishes in the river. (Entailment)
Substituting entities	**P:** Students are inside of a lecture hall. **OH:** Students are **indoors**. (Entailment) **NH:** Students are **on the soccer field**. (Contradiction)
Adding details to entities	**P:** An older man with glasses raises his eyebrows in surprise. **OH:** The man **has no glasses**. (Contradiction) **NH:** The man **wears bifocals**. (Neutral)
Inserting relationships	**P:** A blond woman speaking to a brunette woman with her arms crossed. **OH:** A woman is talking to **another woman**. (Entailment) **NH:** A woman is talking to **a family member**. (Neutral)
Numerical modifications	**P:** Several farmers bent over working on the fields while lady with a baby and four other children accompany them. **OH:** The lady has **three** children. (Contradiction) **NH:** The lady has **many** children. (Entailment)
Using/Removing negation	**P:** An older man with glasses raises his eyebrows in surprise. **OH:** The man **has no** glasses. (Contradiction) **NH:** The man **wears** glasses. (Entailment)
Unrelated hypothesis	**P:** A female athlete in crimson top and dark blue shorts is running on the street. **OH:** A woman is **sitting on** a white couch. (Contradiction) **NH:** A woman **owns** a white couch. (Neutral)

and then a fully-connected output layer with softmax activation. We train all models for a maximum of 20 epochs using Adam [Kingma and Ba, 2015], with a learning rate of $1e{-}3$ and a batch size of 32. We apply early stopping when validation loss does not decrease for 5 epochs. We also experimented with a larger Bi-LSTM which led to overfitting. We use the architecture due to Poliak et al. [2018] to evaluate hypothesis-only baselines.[1]

ELMo-LSTM We compute contextualized word representations (ELMo) using character-based word representations and bidirectional LSTMs [Peters et al., 2018]. The module outputs a 1024-dimensional weighted sum of representations from the 3 Bi-LSTM layers used in ELMo. We represent each word by a 128-dimensional embedding concatenated to the resulting 1024-

[1] https://github.com/azpoliak/hypothesis-only-NLI

Table 3.4: Analysis of edits performed by humans for NLI premises. OP denotes *Original Premise*, NP denotes *New Premise*, and H denotes *Hypothesis*.

Types of Revisions	Examples
Introducing direct evidence	**OP:** Man walking with tall buildings with reflections behind him. (Neutral) **NP:** Man walking **away from his friend**, with tall buildings with reflections behind him. (Contradiction) **H:** The man was walking to meet a friend.
Introducing indirect evidence	**OP:** An Indian man standing on the bank of a river. (Neutral) **NP:** An Indian man standing **with only a camera** on the bank of a river. (Contradiction) **H:** He is fishing.
Substituting entities	**OP:** A young man in front of a **grill** laughs while pointing at something to his left. (Entailment) **NP:** A young man in front of a **chair** laughs while pointing at something to his left. (Neutral) **H:** A man is outside
Numerical modifications	**OP:** The exhaustion in the woman's face while she continues to ride her bicycle in the competition. (Neutral) **NP:** The exhaustion in the woman's face while she continues to ride her bicycle in the competition **for people above 7 ft**. (Entailment) **H:** A tall person on a bike
Reducing evidence	**OP:** The girl in yellow shorts and white jacket has a tennis ball **in her left pocket**. (Entailment) **NP:** The girl in yellow shorts and white jacket has a tennis ball. (Neutral) **H:** A girl with a tennis ball in her pocket.
Using abstractions	**OP:** An elderly **woman** in a crowd pushing a wheelchair. (Entailment) **NP:** An elderly **person** in a crowd pushing a wheelchair. (Neutral) **H:** There is an elderly woman in a crowd.
Substituting evidence	**OP:** A woman is **cutting something with scissors**. (Entailment) **NP:** A woman is **reading something about scissors**. (Contradiction) **H:** A woman uses a tool

dimensional ELMo representation, leading to a 1152-dimensional hidden representation. Following Batch Normalization, this is passed through an LSTM (hidden size 128) with recurrent dropout (probability 0.2). The output from this LSTM is then passed to a fully-connected output layer with softmax activation. We train this model for up to 20 epochs with same early stopping criteria as for Bi-LSTM, using the Adam optimizer with a learning rate of $1\mathrm{e}{-3}$ and a batch size of 32.

BERT We use an off-the-shelf uncased BERT Base model, fine-tuning for each task.[2] To account for BERT's sub-word tokenization, we set the maximum token length is set at 350 for sentiment analysis and 50 for NLI. We fine-tune BERT up to 20 epochs with same early stopping criteria as for Bi-LSTM, using the BERT Adam optimizer with a batch size of 16 (to fit on a Tesla V-100 GPU). We found learning rates of $5\mathrm{e}{-5}$ and $1\mathrm{e}{-5}$ to work best for sentiment analysis and NLI respectively.

3.5 Experiments with CAD

Sentiment Analysis We find that for sentiment analysis, linear models trained on the original $1.7k$ reviews achieve 80% accuracy when evaluated on original reviews but only 51% (level of random guessing) on revised reviews (Table 3.5). Linear models trained on revised reviews achieve 91% accuracy on revised reviews but only 58.3% on the original test set. We see similar pattern for Bi-LSTMs where accuracy drops substantially in both directions. Interestingly, while BERT models suffer drops too, they are less pronounced, perhaps a benefit of the exposure to a larger dataset where the spurious patterns may not have held. Classifiers trained on combined datasets perform well on both, often within ≈ 3 pts of models trained on the same amount of data taken only from the original distribution. Thus, there may be a price to pay for breaking the reliance on spurious associations, but it may not be substantial.

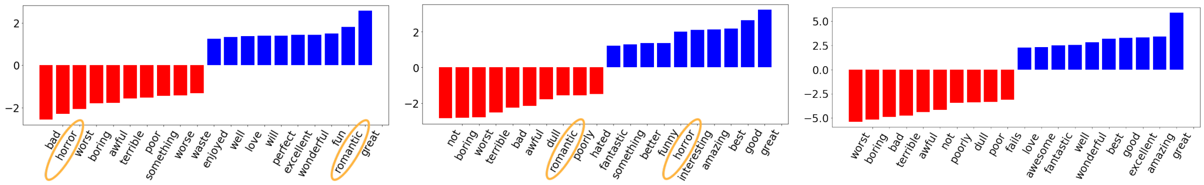
We also conduct experiments to evaluate our sentiment models vis-a-vis their generalization out-of-domain to new domains. We evaluate models on Amazon reviews [Ni et al., 2019a] on data aggregated over six genres: *beauty, fashion, appliances, giftcards, magazines,* and *software*, the Twitter sentiment dataset [Nakov et al., 2013],[3] and Yelp reviews released as part of the Yelp dataset challenge. We show that in almost all cases, models trained on the counterfactually-augmented IMDb dataset perform better than models trained on comparable quantities of original data.

To gain intuition about what is learnable absent the edited spans, we tried training several models on passages where the edited spans have been removed from training set sentences (but not test set). SVM, Naïve Bayes, and Bi-LSTM achieve $57.8\%, 59.1\%, 60.2\%$ accuracy, respectively, on this task. Notably, these passages are predictive of the (true) label despite being semantially compatible with the counterfactual label. However, BERT performs worse than random guessing.

In one simple demonstration of the benefits of our approach, we note that seemingly irrelevant words such as: *romantic, will, my, has, especially, life, works, both, it, its, lives* and *gives* (correlated with positive sentiment), and *horror, own, jesus, cannot, even, instead, minutes, your,*

[2] https://github.com/huggingface/pytorch-transformers

[3] We use the development set as test data is not public.



(a) Trained on the original dataset (b) Trained on the revised dataset (c) Trained on combined dataset

Figure 3.3: Most important features learned by an SVM classifier trained on TF-IDF bag of words.

Table 3.5: Accuracy of various models for sentiment analysis trained with various datasets. Orig. denotes *original*, *Rev.* denotes revised, and *Orig. - Edited* denotes the original dataset where the edited spans have been removed.

Training data	SVM		NB		ELMo		Bi-LSTM		BERT	
	O	R	O	R	O	R	O	R	O	R
Orig. (1.7k)	**80.0**	51.0	**74.9**	47.3	**81.9**	66.7	**79.3**	55.7	**87.4**	82.2
Rev. (1.7k)	58.3	**91.2**	50.9	**88.7**	63.8	**82.0**	62.5	**89.1**	80.4	**90.8**
Orig. − Edited	57.8	–	59.1	–	50.3	–	60.2	–	49.2	–
Orig. & Rev. (3.4k)	83.7	**87.3**	**86.1**	**91.2**	**85.0**	**92.0**	**81.5**	**92.0**	88.5	**95.1**
Orig. (3.4k)	**85.1**	54.3	82.4	48.2	82.4	61.1	80.4	59.6	**90.2**	86.1
Orig. (19k)	**87.8**	60.9	84.3	42.8	86.5	64.3	86.3	68.0	93.2	88.3
Orig. (19k) & Rev.	**87.8**	**76.2**	**85.2**	**48.4**	**88.3**	**84.6**	**88.7**	**79.5**	**93.2**	**93.9**

effort, script, seems and *something* (correlated with negative sentiment) are picked up as high-weight features by linear models trained on either original or revised reviews as top predictors. However, because humans never edit these during revision owing to their lack of semantic relevance, combining the original and revised datasets breaks these associations and these terms cease to be predictive of sentiment (Fig 3.3). Models trained on original data but at the same scale as combined data are able to perform slightly better on the original test set but still fail on the revised reviews. All models trained on $19k$ original reviews receive a slight boost in accuracy on revised data (except Naïve Bayes), yet their performance significantly worse compared to specialized models. Retraining models on a combination of the original $19k$ reviews with revised $1.7k$ reviews leads to significant increases in accuracy for all models on classifying revised reviews, while slightly improving the accuracy on classifying the original reviews. This underscores the importance of including counterfactually-revised examples in training data.

Natural Language Inference Fine-tuned on $1.67k$ original sentence pairs, BERT achieves 72.2% accuracy on SNLI dataset but it is only able to accurately classify 39.7% sentence pairs from the RP set (Table 3.7). Fine-tuning BERT on the full SNLI training set ($500k$ sentence pairs) results in similar behavior. Fine-tuning it on RP sentence pairs improves its accuracy to

Table 3.6: Accuracy of various sentiment analysis models on out-of-domain data

Training data	SVM	NB	ELMo	Bi-LSTM	BERT
Accuracy on Amazon Reviews					
Orig. & Rev. ($3.4k$)	**77.1**	**82.6**	**78.4**	**82.7**	**85.1**
Orig. ($3.4k$)	74.7	66.9	**79.1**	65.9	80.0
Accuracy on Semeval 2017 (Twitter)					
Orig. & Rev. ($3.4k$)	**66.5**	**73.9**	**70.0**	**68.7**	**82.9**
Orig. ($3.4k$)	61.2	64.6	**69.5**	55.3	79.3
Accuracy on Yelp Reviews					
Orig. & Rev. ($3.4k$)	**87.6**	**89.6**	**87.2**	**86.2**	**89.4**
Orig. ($3.4k$)	81.8	77.5	82.0	78.0	85.3

Table 3.7: Accuracy of BERT on NLI with various train and eval sets.

Train/Eval	Original	RP	RH	RP & RH
Original ($1.67k$)	72.2	39.7	59.5	49.6
Revised Premise (RP; $3.3k$)	50.6	66.3	50.1	58.2
Revised Hypothesis (RH; $3.3k$)	71.9	47.4	67.0	57.2
RP & RH ($6.6k$)	64.7	64.6	67.8	66.2
Original w/ RP & RH ($8.3k$)	73.5	**64.6**	**69.6**	**67.1**
Original ($8.3k$)	**77.8**	44.6	66.1	55.4
Original ($500k$)	90.4	54.3	74.3	64.3

66.3% on RP but causes a drop of roughly 20 pts on SNLI. On RH sentence pairs, this results in an accuracy of 67% on RH and 71.9% on SNLI test set but 47.4% on the RP set. To put these numbers in context, each individual hypothesis sentence in RP is associated with two labels, each in the presence of a different premise. A model that relies on hypotheses only would at best perform slightly better than choosing the majority class when evaluated on this dataset. However, fine-tuning BERT on a combination of RP and RH leads to consistent performance on all datasets as the dataset design forces models to look at both premise and hypothesis. Combining original sentences with RP and RH improves these numbers even further. We compare this with the performance obtained by fine-tuning it on $8.3k$ sentence pairs sampled from SNLI training set, and show that while the two perform roughly within 4 pts of each other when evaluated on SNLI, the former outperforms latter on both RP and RH.

Table 3.8: Accuracy of Bi-LSTM classifier trained on hypotheses only

Train/Test	Original	RP	RH	RP & RH
Majority class	34.7	34.6	34.6	34.6
RP & RH $(6.6k)$	**32.4**	**35.1**	**33.4**	**34.2**
Original w/ RP & RH $(8.3k)$	44.0	25.8	43.2	**34.5**
Original $(8.3k)$	60.2	20.5	46.6	**33.6**
Original $(500k)$	69.0	15.4	53.2	**34.3**

Table 3.9: Accuracy of models trained to differentiate between original and revised data

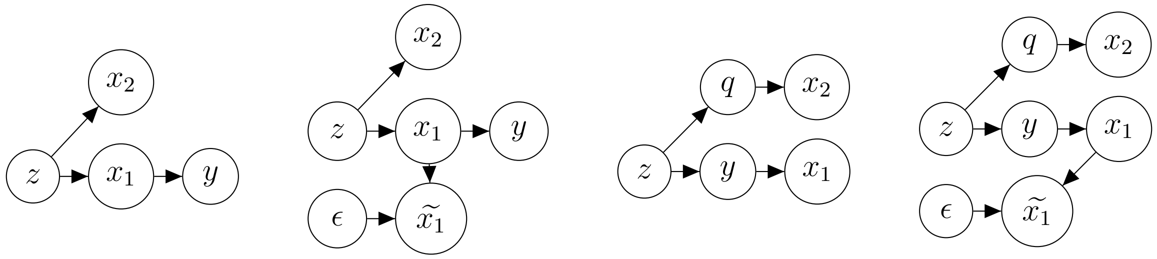
Model	IMDb	SNLI/RP	SNLI/RH
Majority class	50.0	66.7	66.7
SVM	67.4	46.6	51.0
NB	69.2	**66.7**	66.6
BERT	**77.3**	64.8	**69.7**

To further isolate this effect, Bi-LSTM trained on SNLI hypotheses only achieves 69% accuracy on SNLI test set, which drops to 44% if it is retrained on combination of original, RP and RH data (Table 3.8). Note that this combined dataset consists of five variants of each original premise-hypothesis pair. Of these five pairs, three consist of the same hypothesis sentence, each associated with different truth value given the respective premise. Using these hypotheses only would provide conflicting feedback to a classifier during training, thus causing the drop in performance. Further, we notice that the gain of the latter over majority class baseline comes primarily from the original data, as the same model retrained only on RP and RH data experiences a further drop of 11.6% in accuracy, performing worse than just choosing the majority class at all times.

One reasonable concern might be that our models would simply distinguish whether an example were from the original or revised dataset and thereafter treat them differently. The fear might be that our models would exhibit a hypersensitivity (rather than insensitivity) to domain. To test the potential for this behavior, we train several models to distinguish between original and revised data (Table 3.9). BERT identifies original reviews from revised reviews with 77.3% accuracy. In case of NLI, BERT and Naïve Bayes perform roughly within 3 pts of the majority class baseline (66.7%) whereas SVM performs substantially worse.

3.6 Explaining the Efficacy of CAD: A Toy Formulation

We briefly review the OLS estimator for the model $Y = X\beta + \epsilon$, where $Y \in \mathrm{R}^n$ is the target, $X \in \mathrm{R}^{n \times p}$ the design matrix, $\beta \in \mathrm{R}^p$ the coefficient vector we want to estimate, and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 \mathbf{I}_n)$ an iid noise term. The OLS estimate β^{ols} is given by $\mathrm{Cov}(X, X)\beta^{ols} = \mathrm{Cov}(X, Y)$. Representing



(a) Causal setting

(b) Noisy measurement in causal setting

(c) Anticausal setting

(d) Noisy measurements in anticausal setting

Figure 3.4: Toy causal models with one hidden confounder. In 3.4a and 3.4c, the observed covariates are x_1, x_2. In 3.4b and 3.4d, the observed covariates are $\widetilde{x_1}, x_2$. In all cases, y denotes the label.

Var$[X_i]$ as $\sigma^2_{x_i}$ and $\mathrm{Cov}(X_i, X_j)$ as σ_{x_i,x_j}, if we observe only two covariates $(p = 2)$, then:

$$\beta_1^{ols} = \frac{\sigma^2_{x_2}\sigma_{x_1,y} - \sigma_{x_1,x_2}\sigma_{x_2,y}}{\sigma^2_{x_1}\sigma^2_{x_2} - \sigma^2_{x_1,x_2}} \qquad \beta_2^{ols} = \frac{\sigma^2_{x_1}\sigma_{x_2,y} - \sigma_{x_1,x_2}\sigma_{x_1,y}}{\sigma^2_{x_1}\sigma^2_{x_2} - \sigma^2_{x_1,x_2}} \tag{3.1}$$

Our analysis adopts the structural causal model (SCM) framework [Pearl, 2009], formalizing causal relationships via Directed Acyclic Graphs (DAGs). Each edge of the form $A \to B \in \mathcal{E}$ in a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ indicates that the variable A is (potentially) a direct cause of variable B. All measured variables $X \in \mathcal{V}$ in the model are deterministic functions of their corresponding parents Pa$(X) \subseteq \mathcal{V}$ and a set of jointly independent noise terms. For simplicity, we work with linear Gaussian SCMs in the presence of a single confounder where each variable is a linear function of its parents and the noise terms are assumed to be additive and Gaussian. We look at both causal and anticausal learning settings. In the former, we assume that a document causes the applicability of the label (as in annotation, where the document truly causes the label). In the latter interpretation, we assume that the label is one latent variable (among many) that causes features of the document (as when a reviewer's "actual sentiment" influences what they write). For simplicity, we assume that the latent variables are correlated due to confounding but that each latent causes a distinct set of observed features. Without loss of generality, we assume that all variables have zero mean. Both DAGs contain the four random variables z, x_1, x_2, y and the anticausal DAG also contains some additional latent variables q (Figure 3.4).

3.6.1 The Causal Setting

We now focus on the causal setting (Figure 3.4a, 3.4b) Let the Gaussian SCM be defined as follows where the noise term for variable x is defined as u_x:

$$\begin{aligned}
z &= u_z, & u_z &\sim \mathcal{N}(0, \sigma^2_{u_z}) \\
x_1 &= bz + u_{x_1}, & u_{x_1} &\sim \mathcal{N}(0, \sigma^2_{u_{x1}}) \\
x_2 &= cz + u_{x_2}, & u_{x_2} &\sim \mathcal{N}(0, \sigma^2_{u_{x2}}) \\
y &= ax_1 + u_y, & u_y &\sim \mathcal{N}(0, \sigma^2_{u_y}).
\end{aligned} \tag{3.2}$$

Applying OLS, we obtain $\beta_1^{ols} = a$ and $\beta_2^{ols} = 0$. However, consider what happens if we only observe x_1 via a noisy proxy $\widetilde{x_1} \sim \mathcal{N}(x_1, \sigma_{u_{x_1}}^2 + \sigma_{\epsilon_{x_1}}^2)$ (Figure 3.4b). Assuming, $\epsilon_{x_1} \perp\!\!\!\perp (x_1, x_2, y)$, from Eq. 3.1 we get the estimates $\widehat{\beta_1^{ols}}$ and $\widehat{\beta_2^{ols}}$ (Eq. 3.3) in the presence of observation noise on x_1.

$$\begin{aligned}
\widehat{\beta_1^{ols}} &= \frac{a(\sigma_{u_z}^2(b^2\sigma_{u_{x2}}^2 + c^2\sigma_{u_{x1}}^2) + \sigma_{u_{x1}}^2\sigma_{u_{x2}}^2)}{\sigma_{u_z}^2(b^2\sigma_{u_{x2}}^2 + c^2\sigma_{u_{x1}}^2) + \sigma_{u_{x1}}^2\sigma_{u_{x2}}^2 + \sigma_{\epsilon_{x1}}^2(c^2\sigma_{u_z}^2 + \sigma_{u_{x2}}^2)} \\
\widehat{\beta_2^{ols}} &= \frac{acb\sigma_{\epsilon_{x1}}^2\sigma_{u_z}^2}{\sigma_{u_z}^2(b^2\sigma_{u_{x2}}^2 + c^2\sigma_{u_{x1}}^2) + \sigma_{u_{x1}}^2\sigma_{u_{x2}}^2 + \sigma_{\epsilon_{x1}}^2(c^2\sigma_{u_z}^2 + \sigma_{u_{x2}}^2)}
\end{aligned} \tag{3.3}$$

As we can see, $\widehat{\beta_1^{ols}} \propto \frac{1}{\sigma_{\epsilon_{x1}}^2}$. This shows us that as $\sigma_{\epsilon_{x1}}^2$ increases, $|\widehat{\beta_1^{ols}}|$ (the magnitude of the coefficient for x_1) decreases and $|\widehat{\beta_2^{ols}}|$ (the magnitude of the coefficient for x_2) increases. The asymptotic OLS estimates in the presence of infinite observational noise is $\lim_{\sigma_{\epsilon_{x1}}^2 \to \infty} \widehat{\beta_1^{ols}} = 0$, whereas $\widehat{\beta_2^{ols}}$ converges to a finite non-zero value. On the other hand, observing a noisy version of x_2 will not affect our OLS estimates if there is no measurement error on x_1.

These simple graphs provide qualitative insights into when we should expect a model to rely on spurious patterns. In the causal setting, under perfect measurement, the causal variable d-separates the non-causal variable from the label (Figure 3.4a). However, under observation noise, a predictor will rely on the non-causal variable (Eq. 3.3). Moreover, when the causal feature is noisily observed, additional observation noise on non-causal features yields models that are more reliant on causal features. We argue that while review text is not noisily observed per se, learning with imperfect feature representations acquired by training deep networks on finite samples has an effect that is analogous to learning with observation noise.

Connection to Counterfactually Augmented Data In the causal setting, intervening on the causal feature, d-separates the label y from the non-causal feature x_2, and thus models trained on samples from the interventional distribution will rely solely on the causal feature, even when it is noisily observed. We argue that in a qualitative sense, the process of generating CAD resembles such an intervention, however instead of intervening randomly, we ensure that for each example, we produce two sets of values of x_1, one such that the label is applicable and one such that it is not applicable. One is given in the dataset, and the other is produced via the revision.

3.6.2 An Anticausal Interpretation

Alternatively, rather than thinking of features causing the applicable label, we might think of the "causal feature" as a direct effect of the label (not a cause). In this case, so long as the relationship is truly not deterministic, even absent noisy observation, conditioning on the causal feature does not d-separate the label from the non-causal feature and thus models should be expected to assign weight to both causal and non-causal variables.

As in the causal setting, as we increase observation noise on the causal variable, the weight assigned to the non-causal variable should increase. Conversely, as in the causal setting with observation noise on x_1, as observation noise on the non-causal feature x_2 increases, we expect the learned predictor to rely more on the causal feature. The OLS coefficients for Fig. 3.4c are as follows:

$$\beta_1^{ols} = \frac{d(a^2c^2\sigma^2_{u_z}\sigma^2_{u_y} + (c^2\sigma^2_{u_q} + \sigma^2_{u_{x2}})(b^2\sigma^2_{u_z} + \sigma^2_{u_y}))}{(d^2b^2\sigma^2_{u_z} + \sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})(\sigma^2_{u_{x2}} + c^2\sigma^2_{u_q}) + (\sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})c^2a^2\sigma^2_{u_z}}$$
$$\beta_2^{ols} = \frac{abc\sigma^2_{u_z}\sigma^2_{u_{x1}}}{(d^2b^2\sigma^2_{u_z} + \sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})(\sigma^2_{u_{x2}} + c^2\sigma^2_{u_q}) + (\sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})c^2a^2\sigma^2_{u_z}} \tag{3.4}$$

If we observe a noisy version of x_1, given by $\widetilde{x_1}$:

$$\widetilde{x_1} = x_1 + \epsilon_{x_1}, \qquad \epsilon_{x_1} \sim \mathcal{N}(0, \sigma^2_{\epsilon_{x1}}) \tag{3.5}$$

Since $\epsilon_{x_1} \perp\!\!\!\perp x_2, y$, in order to obtain expressions for the OLS estimates $\widehat{\beta_1^{ols}}, \widehat{\beta_2^{ols}}$ in the presence of observation noise, in Eq. 3.4 we only need to replace $\sigma^2_{u_{x1}}$ with $\sigma^2_{u_{\widetilde{x1}}}$, which is given by:

$$\sigma^2_{u_{\widetilde{x1}}} = \sigma^2_{u_{x1}} + \sigma^2_{\epsilon_{x1}} \tag{3.6}$$

$$\widehat{\beta_1^{ols}} = \frac{d(a^2c^2\sigma^2_{u_z}\sigma^2_{u_y} + (c^2\sigma^2_{u_q} + \sigma^2_{u_{x2}})(b^2\sigma^2_{u_z} + \sigma^2_{u_y}))}{(d^2b^2\sigma^2_{u_z} + (\sigma^2_{u_{x1}} + \sigma^2_{\epsilon_{x1}}) + d^2\sigma^2_{u_y})(\sigma^2_{u_{x2}} + c^2\sigma^2_{u_q}) + ((\sigma^2_{u_{x1}} + \sigma^2_{\epsilon_{x1}}) + d^2\sigma^2_{u_y})c^2a^2\sigma^2_{u_z}} \tag{3.7}$$

$$\widehat{\beta_2^{ols}} = \frac{abc\sigma^2_{u_z}(\sigma^2_{u_{x1}} + \sigma^2_{\epsilon_{x1}})}{(d^2b^2\sigma^2_{u_z} + (\sigma^2_{u_{x1}} + \sigma^2_{\epsilon_{x1}}) + d^2\sigma^2_{u_y})(\sigma^2_{u_{x2}} + c^2\sigma^2_{u_q}) + ((\sigma^2_{u_{x1}} + \sigma^2_{\epsilon_{x1}}) + d^2\sigma^2_{u_y})c^2a^2\sigma^2_{u_z}} \tag{3.8}$$

$$\widehat{\beta_1^{ols}} = \frac{\beta_1^{ols}}{1 + \lambda_{ac}^{x_1}} \qquad \widehat{\beta_2^{ols}} = \frac{\beta_2^{ols}}{1 + \lambda_{ac}^{x_1}}\left[1 + \frac{\sigma^2_{\epsilon_{x_1}}}{\sigma^2_{u_{x_1}}}\right] \tag{3.9}$$

$$\lambda_{ac}^{x_1} = \frac{\sigma^2_{\epsilon_{x_1}}(c^2a^2\sigma^2_{u_z} + c^2\sigma^2_{u_q} + \sigma^2_{u_{x2}})}{(d^2b^2\sigma^2_{u_z} + \sigma^2_{u_{x_1}} + d^2\sigma^2_{u_y})(\sigma^2_{u_{x2}} + c^2\sigma^2_{u_q}) + (\sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})c^2a^2\sigma^2_{u_z}} \tag{3.10}$$

where $\lambda_{ac}^{x_1} > 0$ and $\lambda_{ac}^{x_1} \propto \sigma^2_{\epsilon_{x1}}$. Thus, as $\sigma^2_{\epsilon_{x1}}$ increases, $|\widehat{\beta_1^{ols}}|$ decreases. The asymptotic OLS estimates in the presence of infinite observational noise can be seen to be: $\lim_{\sigma^2_{\epsilon_{x1}} \to \infty} \widehat{\beta_1^{ols}} = 0$, where as $\lim_{\sigma^2_{\epsilon_{x1}} \to \infty} \widehat{\beta_2^{ols}} = \beta_2^{ols}\frac{((d^2b^2\sigma^2_{u_z} + \sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})(\sigma^2_{u_{x2}} + c^2\sigma^2_{u_q}) + (\sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})c^2a^2\sigma^2_{u_z})}{(\sigma^2_{u_{x1}}(c^2a^2\sigma^2_{u_z} + c^2\sigma^2_{u_q} + \sigma^2_{u_{x2}}))}$.

Similarly, if we observe a noisy version of X_2, given by $\widetilde{X_2}$:

$$\widetilde{x_2} = x_2 + \epsilon_{x_2}, \qquad \epsilon_{x_2} \sim \mathcal{N}(0, \sigma^2_{\epsilon_{x2}}) \tag{3.11}$$

Since $\epsilon_{x_2} \perp\!\!\!\perp x_1, y$, in order to obtain expressions for the OLS estimates $\widehat{\beta_1^{ols}}, \widehat{\beta_2^{ols}}$ in the presence of observation noise on non-causal features, in Eq. 3.4 we only need to replace $\sigma^2_{u_{x2}}$ with $\sigma^2_{u_{\widetilde{x2}}}$, which is given by:

$$\sigma^2_{u_{\widetilde{x2}}} = \sigma^2_{u_{x2}} + \sigma^2_{\epsilon_{x2}} \tag{3.12}$$

$$\widehat{\beta_1^{ols}} = \frac{d(a^2c^2\sigma^2_{u_z}\sigma^2_{u_y} + (c^2\sigma^2_{u_q} + (\sigma^2_{u_{x2}} + \sigma^2_{\epsilon_{x2}}))(b^2\sigma^2_{u_z} + \sigma^2_{u_y}))}{(d^2b^2\sigma^2_{u_z} + \sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})((\sigma^2_{u_{x2}} + \sigma^2_{\epsilon_{x2}}) + c^2\sigma^2_{u_q}) + (\sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})c^2a^2\sigma^2_{u_z}} \tag{3.13}$$

$$\widehat{\beta_2^{ols}} = \frac{abc\sigma^2_{u_z}\sigma^2_{u_{x1}}}{(d^2b^2\sigma^2_{u_z} + \sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})((\sigma^2_{u_{x_2}} + \sigma^2_{\epsilon_{x2}}) + c^2\sigma^2_{u_q}) + (\sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})c^2a^2\sigma^2_{u_z}} \tag{3.14}$$

$$\begin{aligned} \widehat{\beta_1^{ols}} &= \frac{\beta_1^{ols}}{1+\lambda_{ac}^{x_2}}\left[1 + \frac{\sigma^2_{\epsilon_{x2}}(b^2\sigma^2_{u_z} + \sigma^2_{u_y})}{a^2c^2\sigma^2_{u_z}\sigma^2_{u_y} + (c^2\sigma^2_{u_q} + \sigma^2_{u_{x2}})(b^2\sigma^2_{u_z} + \sigma^2_{u_y})}\right] \quad \widehat{\beta_2^{ols}} = \frac{\beta_2^{ols}}{1+\lambda_{ac}^{x_2}} \\ \lambda_{ac}^{x_2} &= \frac{\sigma^2_{\epsilon_{x2}}(d^2b^2\sigma^2_{u_z} + \sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})}{(d^2b^2\sigma^2_{u_z} + \sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})(\sigma^2_{u_{x2}} + c^2\sigma^2_{u_q}) + (\sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})c^2a^2\sigma^2_{u_z}} \end{aligned} \tag{3.15}$$

where $\lambda_{ac}^{x_2} > 0$ and $\lambda_{ac}^{x_2} \propto \sigma^2_{\epsilon_{x2}}$. Thus, as $\sigma^2_{\epsilon_{x2}}$ increases, $|\widehat{\beta_1^{ols}}|$ increases. The asymptotic OLS estimates in the presence of infinite observational noise can be seen to be: $\lim_{\sigma^2_{\epsilon_{x2}} \to \infty} \widehat{\beta_2^{ols}} = 0$, where as $\lim_{\sigma^2_{\epsilon_{x2}} \to \infty} \widehat{\beta_1^{ols}} = \beta_1^{ols} \frac{(b^2\sigma^2_{u_z} + \sigma^2_{u_y})((d^2b^2\sigma^2_{u_z} + \sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})(\sigma^2_{u_{x2}} + c^2\sigma^2_{u_q}) + (\sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})c^2a^2\sigma^2_{u_z})}{(a^2c^2\sigma^2_{u_z}\sigma^2_{u_y} + (c^2\sigma^2_{u_q} + \sigma^2_{u_{x2}})(b^2\sigma^2_{u_z} + \sigma^2_{u_y}))(d^2b^2\sigma^2_{u_z} + \sigma^2_{u_{x1}} + d^2\sigma^2_{u_y})}$.

Connection to Counterfactually Augmented Data In this interpretation, we think of CAD as a process by which we (the designers of the experiment) intervene on the label itself and the human editors, play the role of a simulator that we imagine to be capable of generating a counterfactual example, holding all other latent variables constant. In the sentiment case, we could think of the editors as providing us with the review that would have existed had the sentiment been flipped, holding all other aspects of the review constant. Note that by intervening on the label, we d-separate it from the spurious correlate x_2 (Figure 3.4c).

3.6.3 Insights and Testable Hypotheses

In both the causal and anticausal models, the mechanism underlying the causal relationship that binds x_1 to y (regardless of direction) is that binding language to a semantic concept (such as sentiment), which we expect to be more stable across settings than the more capricious relationships among the background variables, e.g., those linking genre and production quality.

In that spirit, if spans edited to generate *counterfactually revised data* (CRD) are analogous to the causal (or anticausal) variables, in the causal (or anticausal) graphs, then we might expect that noising those spans (e.g. by random word replacement) should lead to models that rely more on non-causal features and perform worse on out of domain data. On the other hand, we expect that noising unedited spans should have the opposite behavior, leading to degraded in-domain performance, but comparatively better out-of-domain performance. In the remainder of the paper, we investigate these hypotheses, finding evidence that qualitatively confirms the predictions of our theory.

We freely acknowledge the speculative nature of this analysis and concede that the mapping between the messy unstructured data we wish to model and the neatly disentangled portrait captured by our linear Gaussian models leaves a gap to be closed through further iterations of theoretical refinement and scientific experiment. Ultimately, our argument is not that this simple analysis fully accounts for counterfactually augmented data but instead that it is a useful abstraction for formalizing two (very different) perspectives on how to conceive of CAD, and for suggesting interesting hypotheses amenable to empirical verification.

3.7 Experiments and Discussion

If spans marked as rationales by humans via editing or highlighting are analogous to causal features, then noising those spans should lead to models that rely more on non-causal features and thus perform worse on out-of-domain data, and noising the unmarked spans (analagous to non-causal features) should have the opposite behavior. In this section, we test these hypotheses empirically on real-world datasets. Additionally, we investigate whether the feedback from human workers is yielding anything qualitatively different from what might be seen with spans marked by automated feature attribution methods such as attention and saliency. Along similar, lines we ask whether CAD in the first place offers qualitative advantages over what might be achieved via automatic sentiment-flipping methods through experiments with text style transfer algorithms.

We conduct experiments on sentiment analysis [Zaidan et al., 2007, Kaushik et al., 2020] and NLI [DeYoung et al., 2020]. All datasets are accompanied with human feedback (tokens deemed relevant to the label's applicability) which we refer to as *rationales*. For the first set of experiments, we rely on four models: Support Vector Machines (SVMs), Bidirectional Long Short-Term Memory Networks (BiLSTMs) with Self-Attention [Graves and Schmidhuber, 2005], BERT [Devlin et al., 2019], and Longformer [Beltagy et al., 2020]. For the second set of experiments, we rely on four state-of-the-art style transfer models representative of different methodologies, each representative of a different approach to automatically generate new examples with flipped labels [Hu et al., 2017, Li et al., 2018, Sudhakar et al., 2019, Madaan et al., 2020]. To evaluate classifier performance on the resulting augmented data, we consider SVMs, Naive Bayes (NB), BiLSTMs with Self Attention, and BERT.

For sentiment analysis, we use SVM, BiLSTM with Self Attention, BERT, and Longformer models. In each document, we replace a fraction of *rationale* (or *non-rationale*) tokens with random tokens sampled from the vocabulary, and train our models, repeating the process 5 times. We perform similar experiments for NLI using BERT. As an individual premise-hypothesis pair is often not as long as a movie review, many pairs only have one or two words marked as *rationales*. To observe the effects from gradually injecting noise on *rationales* or *non-rationales*, we select only those premise-hypothesis pairs that have a minimum 10 tokens marked as *rationales*.

Table 3.10: Accuracy of sentiment analysis classifiers trained on $1.7k$ original reviews from Kaushik et al. [2020] as noise is injected on *rationales/non-rationales* identified via humans.

Dataset	Percent noise in rationales										
SVM											
	0	10	20	30	40	50	60	70	80	90	100
In-sample test	87.8	88.2	85.7	86.9	86.9	84.5	83.3	81.6	80	79.2	76.7
CRD	51.8	47.3	45.7	42.9	39.2	33.5	28.2	25.7	24.1	19.6	17.1
Amazon	73.2	72.2	71.3	69.4	67.3	63.7	63.7	58.2	57	50.1	46.5
Semeval	62.5	62.2	61.9	61.1	60.9	58.3	57.1	55.4	54.5	51.3	50.1
Yelp	79.9	79	77.7	76.7	74.1	71.4	69	65.5	62.4	55.8	51.5
BiLSTM with Self Attention											
In-sample test	81.5	78.8	77.6	76.7	75.3	75.2	74.5	72.8	67.3	64.2	63.8
CRD	49.4	49.3	46.3	45.1	39.5	38.1	38.9	38.7	32.6	32.6	29.7
Amazon	65.4	69.1	68.5	66.6	63.2	63.9	58.8	50.6	50.6	47.1	44.2
Semeval	59.3	59.8	57.6	56.4	58.6	56.6	55.3	54.3	54.3	52.3	50
Yelp	71.2	70.8	67.4	65.9	65.3	64.1	63.4	60.1	62.4	49.8	46.4
BERT											
In-sample test	87.4	87.4	86.5	85.7	85.3	84.3	83.6	81	76.6	71	69
CRD	82.2	78.1	78.4	75.4	67.6	67.5	65.5	53.9	42.7	36.2	31.8
Amazon	76.2	75.5	75.1	74.2	73.5	73	72.5	70.7	63.4	57.8	56.1
Semeval	76.4	69.7	66.9	69.8	67.8	67.4	66.8	65.5	62.2	54.9	52.6
Yelp	83.7	82.5	82	81.5	80.9	80.2	79.9	75.6	64.3	54.6	52.3
Dataset	Percent noise in non-rationales										
SVM											
In-sample test	87.8	88.6	89	86.9	85.3	82.4	86.5	83.7	82	81.6	78
CRD	51.8	55.9	53.5	57.1	58.8	63.7	63.3	65.7	70.2	73.9	74.3
Amazon	73.2	74.9	75.3	77.3	75.8	76.6	76.5	77.4	75.5	75.4	76.9
Semeval	62.5	63.3	62.7	64.3	64.3	65.6	66	65.8	65	66.4	66.4
Yelp	79.9	80.9	80.1	82.2	83.6	84.1	83.5	83.4	82.7	82.1	81.4
BiLSTM with Self Attention											
In-sample test	81.5	77.5	77	75.9	75.4	75.2	75.1	73.8	73	72.4	71.7
CRD	49.4	53.1	56.25	56.6	57.5	58.4	58.6	60.3	61.5	65.5	66.1
Amazon	65.4	66.5	66.6	66.6	67.6	67.7	68.3	68.6	68.8	68.5	68.4
Semeval	59.3	58.6	58.9	59.3	58.1	57.5	59.2	59.5	59.8	59.5	58
Yelp	71.2	74.7	72.5	73.3	73.9	73.6	72.2	74.3	73.7	75.6	75.4
BERT											
In-sample test	87.4	88.2	87	86.9	87	85.8	83.6	78.9	72.5	72.1	71.3
CRD	82.2	92.8	92.8	92.3	93.1	92.8	89.8	88.6	84.5	81.3	81
Amazon	76.2	78.6	78.9	79.2	75.1	71.7	67.6	65.3	65.2	63.7	61.8
Semeval	76.4	74.6	76.3	75.8	70.9	62.1	64.8	63.3	60.8	58.7	58.7
Yelp	83.7	85.4	85.3	85.1	82.1	78.3	77.2	76.2	74.3	71.6	70.1

Since no neutral pairs exist with 10 or more rationale tokens, we consider only a binary classification setting (entailment-contradiction), and downsample the majority class to ensure a 50:50

Table 3.11: Accuracy of sentiment analysis classifiers trained on $1.7k$ original reviews from Kaushik et al. [2020] as noise is injected on *rationales/non-rationales* identified via Attention masks.

Dataset	Percent noise in rationales										
SVM											
	0	10	20	30	40	50	60	70	80	90	100
In-sample test	87.8	85	85.9	86.3	86.3	85.2	84.6	86.3	83.6	84.2	83.6
CRD	51.8	50.6	51.8	52	51.8	50	50.6	48.6	48.6	47.5	46.1
Amazon	73.2	74.3	73.4	72.8	72.8	72.9	72	72.3	71.1	72	70.3
Semeval	62.5	62.8	62.9	61.8	62.5	61.9	61.4	60.7	61.1	60.6	60.1
Yelp	79.9	80.1	79.3	78.7	78.9	78.5	77.8	77.5	77.8	76.2	75.9
BiLSTM with Self Attention											
In-sample test	81.5	78.8	78.6	78.3	78.2	76.2	77.3	76.8	71.8	73.2	74.2
CRD	49.4	53.3	50	53.4	52.4	49.7	49.2	47.4	47.7	47	44.1
Amazon	65.4	66.8	71	64.7	60.7	61.7	65.2	64.6	51.6	57.1	66.4
Semeval	59.3	59.5	60.1	57.4	55.9	57.2	52.2	57.6	51.5	51.8	56.1
Yelp	71.2	72.3	74.2	69.6	70.5	67.3	70.7	72.8	62.8	65	66.2
BERT											
In-sample test	87.4	93	90.8	90.3	90.6	91.2	90.3	90.4	90.7	90.6	90.3
CRD	82.2	91.2	92	90.8	90.8	90.9	90.3	90.9	90.2	89.8	90.4
Amazon	76.2	77.3	79.1	78.7	79.8	79.1	79.8	79.5	79.2	78.9	79.3
Semeval	76.4	71.4	73.5	73.2	74.4	76.1	77.6	79.8	78.4	79.2	77.8
Yelp	83.7	83.5	85.4	84.9	86	85.7	85.9	85.6	85.5	85.4	68.9
Dataset	Percent noise in non-rationales										
SVM											
In-sample test	87.8	85	85.7	84.8	85	84	83.6	84.6	80.7	81.1	77.3
CRD	51.8	50.4	52.2	53.9	50.2	50.8	52.9	54.1	54.1	56.8	56.4
Amazon	73.2	73.5	75.3	74.3	76.2	73.9	73.4	73.6	71	70	67.8
Semeval	62.5	62.6	63.7	63.7	63.1	62.6	63.5	61.5	62.1	62	59.9
Yelp	79.9	79.8	80.9	81.7	80.9	80.5	80	80.1	78.5	77.5	74.4
BiLSTM with Self Attention											
In-sample test	81.5	77.6	76	77.1	77.3	75.4	73.7	67.9	68.6	54.2	52.3
CRD	49.4	53.1	52.1	52.1	65	54.1	51.9	53.4	55	52.3	51.6
Amazon	65.4	63.7	65.7	64	58.8	65.5	60.3	58.7	61	58.1	56.2
Semeval	59.3	54.8	58.4	57.3	60.7	56.8	55.2	54	51.2	50	49.9
Yelp	71.2	72	73.6	70.2	61.3	71.5	68.4	64.9	66.3	58.2	55.8
BERT											
In-sample test	87.4	86.9	86.7	85.3	84	81.9	80.6	74	74	73	67.2
CRD	82.2	92.3	92.4	92.1	90	86.8	83	73.2	77.7	72.5	68.5
Amazon	76.2	79.5	78.5	77.9	69.2	67.4	58.1	55.9	53.5	55.8	52.6
Semeval	76.4	76.5	75.7	77.1	65.7	61.8	54.6	58.8	51.8	54	50.8
Yelp	83.7	85.8	85	85.5	79.3	78.7	67.8	66.5	59.5	63.2	57.5

Table 3.12: Accuracy of sentiment analysis classifiers trained on $1.7k$ original reviews from Kaushik et al. [2020] as noise is injected on *rationales/non-rationales* identified via Allen NLP Saliency Interpreter.

Dataset	Percent noise in rationales										
					SVM						
	0	10	20	30	40	50	60	70	80	90	100
In-sample test	87.8	85.1	85.4	85.1	85.1	83.9	82.5	82.8	81.8	80	77.5
CRD	51.8	51.2	52.5	51.1	51	50.1	49.5	46.6	43.7	42.1	40.5
Amazon	73.2	73.4	73.65	73.2	72.7	72.9	71.8	72.1	70.5	69.6	68.9
Semeval	62.5	62.8	62.5	62.4	61.9	61.2	60.7	60.5	59.6	58.4	57.9
Yelp	79.9	79.8	79.7	79.1	78.7	78.2	78.1	76.6	75.1	74.1	72.2
					BiLSTM with Self Attention						
In-sample test	81.5	82.3	83.5	80.4	78.2	81.9	80.6	77.8	79.2	76	77
CRD	49.4	48.2	48.6	51.2	48.6	47.3	47.1	46.9	44.3	42.6	37.9
Amazon	65.4	46.6	72.8	66.9	49.7	55.4	53.7	68.5	54.7	49.8	51.8
Semeval	59.3	42.1	49.5	56.2	54.7	52.7	53.7	50.1	51.2	50.2	50
Yelp	71.2	69	73.3	73.2	67.8	69.2	69.5	68.8	67	54.4	56.9
					BERT						
In-sample test	87.4	91.1	90.6	90	88	89.1	87.4	86.3	83.6	84.5	81.6
CRD	82.2	93.4	92.3	91.9	90.3	90.2	87.7	83.8	78	79.3	70
Amazon	76.2	82.4	81.3	79.8	77.2	77.6	77.8	75.6	69.7	69.4	73.6
Semeval	76.4	82.6	82.8	81.3	79.2	78.1	76.7	74.7	67.4	65.8	67.4
Yelp	83.7	88.3	88.8	88.5	87.8	88.1	87	86.2	84.4	83.3	82.7
Dataset	Percent noise in non-rationales										
					SVM						
In-sample test	87.8	85.9	85.7	86.9	83.6	86.9	85.9	83.2	85	81.8	79.1
CRD	51.8	52.3	53.7	53.9	56.8	55.3	53.5	54.3	58	60	61.5
Amazon	73.2	73.9	74.1	71.8	73.6	72.5	73.8	72.6	70.6	70.6	70.8
Semeval	62.5	62.7	62.8	61.3	62.7	62	61.9	63.2	62.3	62.4	63.6
Yelp	79.9	79.8	79.8	81.4	81	80.7	81	80.5	80.3	79.8	78.6
					BiLSTM with Self Attention						
In-sample test	81.5	81	81.7	80.8	79.8	78	75.6	73	70.4	51	50
CRD	49.4	49	49.8	48	47.9	51.6	46.7	53.3	50.2	51.6	48.4
Amazon	65.4	65.3	64.9	62.7	63.3	65.3	67.1	65.3	64	58.3	41.8
Semeval	59.3	55	61.3	50.1	54.6	58.5	55.2	55.7	49.4	49.6	44
Yelp	71.2	73.8	75.1	71.4	74.1	73.4	74.5	72.5	66.9	55.9	53.6
					BERT						
In-sample test	87.4	90.5	89.1	88.6	80.6	75.1	70.1	63.7	53.8	54.1	53.1
CRD	82.2	92.1	92.2	91.3	79.9	73.3	67.1	59.2	50.1	49.8	49.6
Amazon	76.2	77.5	79.2	77.3	69	65.9	61.1	61.7	52.9	52.7	51.4
Semeval	76.4	82	83.6	83.1	78.1	77.9	71.1	69.7	55.9	56.5	51.8
Yelp	83.7	88	87.4	87.8	76.8	73	66.9	66.3	55.3	54.4	53.1

Table 3.13: Accuracy of sentiment analysis classifiers trained on reviews from Zaidan et al. [2007] as noise is injected on *rationales/non-rationales* identified via humans.

Dataset	Percent noise in rationales										
	SVM										
	0	10	20	30	40	50	60	70	80	90	100
In-sample test	87.5	86.2	85.5	85	84.5	83.3	82.5	81.1	78.9	77.5	76.5
CRD	46.1	45.6	44.4	43.7	44.1	41.2	38.8	36	34.4	33.1	30.9
Amazon	68.6	67.1	65.1	64.2	62.2	60.4	57.9	50.5	54.9	53.5	51.8
Semeval	56.7	56.1	55.4	54.8	54.1	53.5	52.7	52	51.6	50.8	50.4
Yelp	76.2	75	73.5	72	70.2	68.8	66.6	65.1	63.3	61.1	59.3
	BiLSTM with Self Attention										
In-sample test	80.3	82.1	83.2	81.3	78.4	71.1	78.8	77.4	76.9	77.4	75.5
CRD	49.2	50.6	51	48.8	48	49.6	49.4	48.8	48.8	47.5	48.4
Amazon	50	50.5	49.4	49.7	49.8	49.7	49.7	49.7	49.6	49.5	49.4
Semeval	50	50	50	50	50	50	50	50	50	50	50
Yelp	50.5	50	53.1	52.1	50.5	50.2	50.1	50	50	50.2	50.1
	Longformer										
In-sample test	97.5	96.7	94	90.5	88.3	78.9	81.4	72.6	79.4	78.7	83.5
CRD	93.4	93.6	87.5	85.4	84.2	64.1	61.5	54.2	52.7	50.3	48
Amazon	81.8	77.9	65.3	65.7	64.7	63.6	61.9	62.1	61.3	60.6	57.9
Semeval	80.3	74.9	64	66.9	71.6	61.3	58.4	56.7	58.9	62.1	58.6
Yelp	88.6	85.8	77.7	74.6	72.5	68.4	66.5	64.8	64.3	64.9	62.2
Dataset	Percent noise in non-rationales										
	SVM										
In-sample test	87.5	85.5	86	83	82	83	81	80.5	75.5	60	50
CRD	46.1	46.1	49	49.4	57.1	55.5	58.4	58.4	56.5	56.3	54
Amazon	68.6	67.7	68	67.2	69.4	69	69.7	68.9	69.2	64.9	62.3
Semeval	56.7	56.9	57.5	57.4	58.3	57.6	58.8	59.4	59.3	57.4	56.3
Yelp	76.2	76.1	76.9	75.9	77	77.4	75.2	74.1	73.3	68.5	61.6
	BiLSTM with Self Attention										
In-sample test	80.3	80.8	79.8	75.2	75	62.5	62	57.7	56.7	58.7	57.7
CRD	49.2	50	51.1	50.8	52.9	53.9	58.6	58.6	60	60.4	60.8
Amazon	50	50	50.7	50.7	50.9	52.2	52.3	53.2	55	55.1	56.7
Semeval	50	50	50	50	50	51	51.8	52.7	53.5	53.8	53.9
Yelp	50.5	50.4	52.7	52.9	52.9	55.2	58	58.9	64.6	64.6	70
	Longformer										
In-sample test	97.5	97.9	98.1	97.4	94.8	93.4	86.4	82.3	76.3	77.4	80.2
CRD	93.4	94.7	94.1	91.8	91.4	91.8	88	83.4	83.7	83.6	83.4
Amazon	81.8	79	80	81.5	83.2	84.2	84.1	76.3	78.5	79.4	76.9
Semeval	80.3	79.4	77.2	80.6	80.6	84.6	85.3	71.8	79.9	83.7	76.6
Yelp	88.6	85.3	86.4	89	89.5	89.9	89.9	86.2	86.5	86.4	84.7

label split.

Figures 3.5 and 3.6 show the difference in mean accuracy over 5 runs. For all classifiers, as

Table 3.14: Accuracy of sentiment analysis classifiers trained on reviews from Zaidan et al. [2007] as noise is injected on *rationales/non-rationales* identified via Attention masks.

Dataset	Percent noise in rationales										
SVM											
	0	10	20	30	40	50	60	70	80	90	100
In-sample test	87.5	85	84.5	84	82.5	83	81	80	77.5	75.5	75.5
CRD	46.1	51	50.6	52	51.8	52.3	52.3	51.8	50.2	49.8	49.8
Amazon	68.6	68.1	67.1	66.8	66.9	66.5	66.2	65.4	66.1	66.6	65.7
Semeval	56.7	56.6	56.3	56.4	56.2	56.4	56.4	56.2	56.8	56.4	56.4
Yelp	76.2	76.1	76	76.2	76.4	76.5	76.9	76.9	76.7	76.9	76.5
BiLSTM with Self Attention											
In-sample test	80.3	78.8	77.9	77.9	78.8	67.3	65.9	63.9	62	65.4	58.7
CRD	49.2	49.4	50.2	50.2	52.1	51	52.1	52.3	56.3	51.8	54.7
Amazon	50	49.7	49.9	49.9	50.4	50.2	51	51.7	51.1	50.7	50.7
Semeval	50	50	50	50	50	50	50	50.2	50.1	50	50.1
Yelp	50.5	50.1	50.5	50.5	52.1	52.4	56.1	54.9	54.9	52.2	54.9
Longformer											
In-sample test	97.5	97.3	97	96.5	88.3	94	93.8	91.2	91.5	87.2	84
CRD	93.4	93.5	93.1	92.8	91.7	91.8	90.7	88	87.5	83.7	80.8
Amazon	81.8	76.3	69.5	75.4	70.4	64.5	66.3	60.8	64.7	57.3	55.3
Semeval	80.3	73	67.2	75.1	69.6	61.5	67	58.8	67.6	56.4	55.3
Yelp	88.6	85.1	79.3	83.9	79.8	75.4	76.8	69.1	75.4	65.7	61
Dataset	Percent noise in non-rationales										
SVM											
In-sample test	87.5	87	86.5	87.5	81	82.5	73	52	50	50	50
CRD	46.1	50.4	49.6	48.6	50	46.9	50.6	49.6	50.4	50.2	50.2
Amazon	68.6	66.7	66.8	64.1	65.9	63.2	62.2	60	57.8	56.2	56.3
Semeval	56.7	56.3	56.8	55.9	56.7	55	54.2	53.8	51.8	51.1	51
Yelp	76.2	74.8	74.2	71.1	71	64.9	59.7	55.2	52.3	51	50
BiLSTM with Self Attention											
In-sample test	80.3	79.8	81.3	78.4	63.5	67.3	49.5	49	48.1	48.4	48.1
CRD	49.2	51.4	51.4	54.5	49.8	49.4	49.6	49.4	49.4	49.4	49.4
Amazon	50	49.9	50.6	50.4	50.1	49.7	49.6	49.5	49.5	49.5	49.5
Semeval	50	50	50	50.2	50	50	50	50	50	50	50
Yelp	50.5	52.3	52.7	56.9	51	50.4	50	50	50	50	50
Longformer											
In-sample test	97.5	98.2	97.8	95	90.2	83.3	67.3	62.8	69.3	64.2	52.8
CRD	93.4	93.6	93.5	88.8	83.1	76.5	67.8	69.7	77.6	54.5	51.4
Amazon	81.8	81.6	97.8	95	90.2	83.3	67.3	62.8	79.3	64.2	52.8
Semeval	80.3	74.8	70.3	79.1	79	78.9	69.5	67.9	64	63.3	58.6
Yelp	88.6	83.9	83.1	89.5	90.2	89.7	87.6	83	78.8	62.4	59.4

the noise in *rationales* increases, in-sample accuracy stays relatively stable compared to out-of-domain accuracy. An SVM classifier trained on the original $1.7k$ IMDb reviews from Kaushik

Table 3.15: Accuracy of sentiment analysis classifiers trained on reviews from Zaidan et al. [2007] as noise is injected on *rationales/non-rationales* identified via Allen NLP Saliency interpreter.

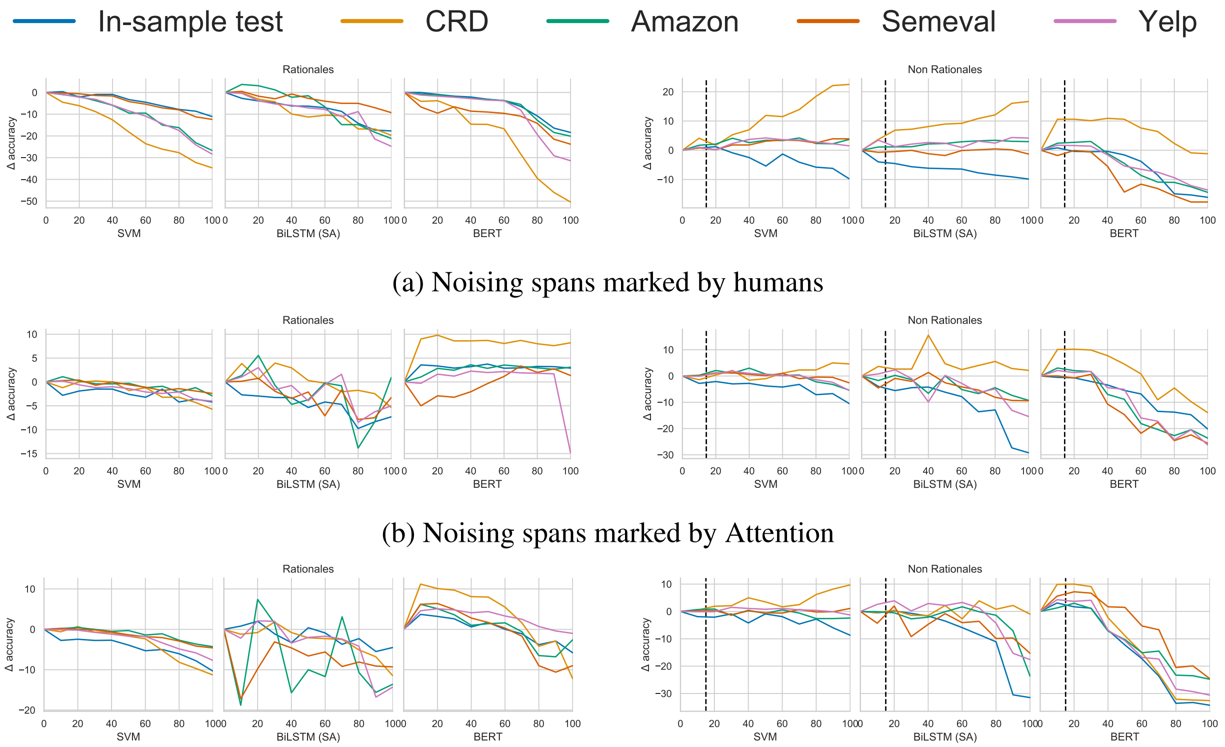
Dataset	Percent rationales tokens replaced by noise										
						SVM					
	0	10	20	30	40	50	60	70	80	90	100
In-sample test	87.5	85	84.5	84	82.5	83	81	80	77.5	75.5	75.5
CRD	46.1	51	50.6	52	51.8	52.3	52.3	51.8	50.2	49.8	49.8
Amazon	68.6	68.1	67.1	66.8	66.9	66.5	66.2	65.4	66.1	66.6	65.7
Semeval	56.7	56.6	56.3	56.4	56.2	56.4	56.4	56.2	56.8	56.4	56.4
Yelp	76.2	76.1	76	76.2	76.4	76.5	76.9	76.9	76.7	76.9	76.5
						BiLSTM with Self Attention					
In-sample test	80.3	83.2	78.1	76.9	73.6	80.3	81.7	76.4	76.4	74	74.5
CRD	49.2	49.8	50.6	50.8	50.8	49.2	49.2	49.2	52	49.4	49.8
Amazon	50	49.8	50.5	49.8	50	49.7	49.7	50.1	50	50.3	49.8
Semeval	50	50	50	50	50	50	50	50	50	50	50
Yelp	50.5	50.4	50	50.5	50.8	50.3	50.1	50.9	50.7	50.8	50.5
						Longformer					
In-sample test	97.5	98	98	97.5	97.5	97	97	97	96.5	94.3	92.8
CRD	93.4	93.4	93.9	94	92.4	91	92.2	91.7	90.2	86.6	74.5
Amazon	81.8	81	74.2	66.3	74.7	78.3	80.6	76.2	63.2	77.3	55.8
Semeval	80.3	79.9	69.4	64	73.4	77	78	74.6	60.6	78.3	56.4
Yelp	88.6	87.3	84.5	76.6	83.1	86.4	87.6	85.6	72.8	84.1	61.7
Dataset	Percent noise in non-rationales										
						SVM					
In-sample test	87.5	79.7	79.9	79.5	81.1	79.9	80.3	78.9	78.7	79.3	73.4
CRD	46.1	52.7	52	53.1	50	54.3	50.6	54.3	52	57.2	57.2
Amazon	68.6	68.1	66.2	67	65.8	68.8	65.3	64.4	65.3	63.1	61.9
Semeval	56.7	57.4	56.2	56.9	55.9	57.3	55.6	58.1	57	57.9	58
Yelp	76.2	76.4	75.4	76	75.6	75.8	74.2	74.3	73.6	73.7	71.5
						BiLSTM with Self Attention					
In-sample test	80.3	76.9	80.8	79.3	78.8	77.9	76	76	63.5	73.6	66.8
CRD	49.2	50	50.2	50.4	50.2	50.8	51.4	47.9	48.8	47.5	48.4
Amazon	50	50	49.7	50.1	50.2	50.8	50.2	50	50.3	49.8	47.9
Semeval	50	50	50	50	50	50.1	50	50	50	50	50
Yelp	50.5	50.5	50.3	51.3	54.5	54.9	52.2	51.4	52.7	50.5	54.9
						Longformer					
In-sample test	97.5	97.8	98	97.8	97.5	98.3	95	92.8	84.5	83.5	74.5
CRD	93.4	94.4	94.1	93.6	93.1	93.3	92.8	91	86.9	70.9	67.6
Amazon	81.8	80.9	75.9	75.8	79.7	68.9	81.4	72.4	71.2	63.5	55.2
Semeval	80.3	78.6	72.7	74.4	79.1	68.9	81.6	73.5	76.2	59.2	55.7
Yelp	88.6	88.1	84.1	84.8	87.5	81.3	89.3	82.2	82.2	70.3	61.5

Table 3.16: Out-of-domain accuracy of models trained on original only, CAD, and original and *sentiment-flipped* reviews

Training data	SVM	NB	BiLSTM (SA)	BERT
Accuracy on Amazon Reviews				
CAD ($3.4k$)	**79.3**	**78.6**	**71.4**	**83.3**
Orig. & Hu et al. [2017]	66.4	71.8	62.6	78.4
Orig. & Li et al. [2018]	62.9	65.4	57.6	61.8
Orig. & Sudhakar et al. [2019]	64.0	69.3	54.7	77.2
Orig. & Madaan et al. [2020]	74.3	73.0	63.8	71.3
Orig. ($3.4k$)	74.5	74.3	68.9	80.0
Accuracy on Semeval 2017 (Twitter)				
CAD ($3.4k$)	**66.8**	**72.4**	**58.2**	**82.8**
Orig. & Hu et al. [2017]	60.9	63.4	56.6	79.2
Orig. & Li et al. [2018]	57.6	60.8	54.7	62.7
Orig. & Sudhakar et al. [2019]	59.4	62.6	54.9	72.5
Orig. & Madaan et al. [2020]	62.8	63.6	54.6	79.3
Orig. ($3.4k$)	63.1	63.7	50.7	72.6
Accuracy on Yelp Reviews				
CAD ($3.4k$)	**85.6**	**86.3**	**73.7**	**86.6**
Orig. & Hu et al. [2017]	77.4	80.4	68.8	84.7
Orig. & Li et al. [2018]	67.8	73.6	63.1	77.1
Orig. & Sudhakar et al. [2019]	69.4	75.1	66.2	84.5
Orig. & Madaan et al. [2020]	81.3	82.1	68.6	78.8
Orig. ($3.4k$)	81.9	82.3	72.0	84.3

Table 3.17: Accuracy of BERT trained on SNLI [DeYoung et al., 2020] as noise is injected on human identified *rationales/non-rationales*. RP and RH are Revised Premise and Revised Hypothesis test sets in Kaushik et al. [2020]. MNLI-M and MNLI-MM are MNLI [Williams et al., 2018a] dev sets.

	Percent noise added to train data rationales										
Dataset	0	10	20	30	40	50	60	70	80	90	100
In-sample test	91.6	90.7	90.0	88.9	87.3	86.2	84.4	80.2	78.0	72.2	71.9
RP	72.7	70.7	69.1	67.1	65.7	62.4	61.8	57.7	55.6	53.8	51.4
RH	84.7	80.8	80.4	79.5	77.2	75.7	73.3	67.7	64.0	57.9	53.2
MNLI-M	75.6	74.7	73.9	72.0	70.6	69.1	64.7	59.1	55.8	54.4	53.3
MNLI-MM	77.9	76.7	75.6	73.9	72.3	70.8	65.6	58.4	55.1	53.6	52.5
	Percent noise added to train data non-rationales										
Dataset	0	10	20	30	40	50	60	70	80	90	100
In-sample test	91.6	91.4	91.3	90.9	90.8	89.9	89.0	88.7	87.8	86.7	85.4
RP	72.7	73.5	73.2	72.1	71.5	70.7	70.6	70.6	70.6	70.6	70.4
RH	84.7	83.6	82.6	81.9	81.3	81.1	80.5	79.8	79.4	79.4	79.2
MNLI-M	75.6	74.9	74.4	72.6	72.4	71.8	71.3	71.3	70.9	70.9	70.8
MNLI-MM	77.9	76.2	75.8	75.0	74.6	74.3	73.9	73.7	73.3	73.0	72.8



(a) Noising spans marked by humans

(b) Noising spans marked by Attention

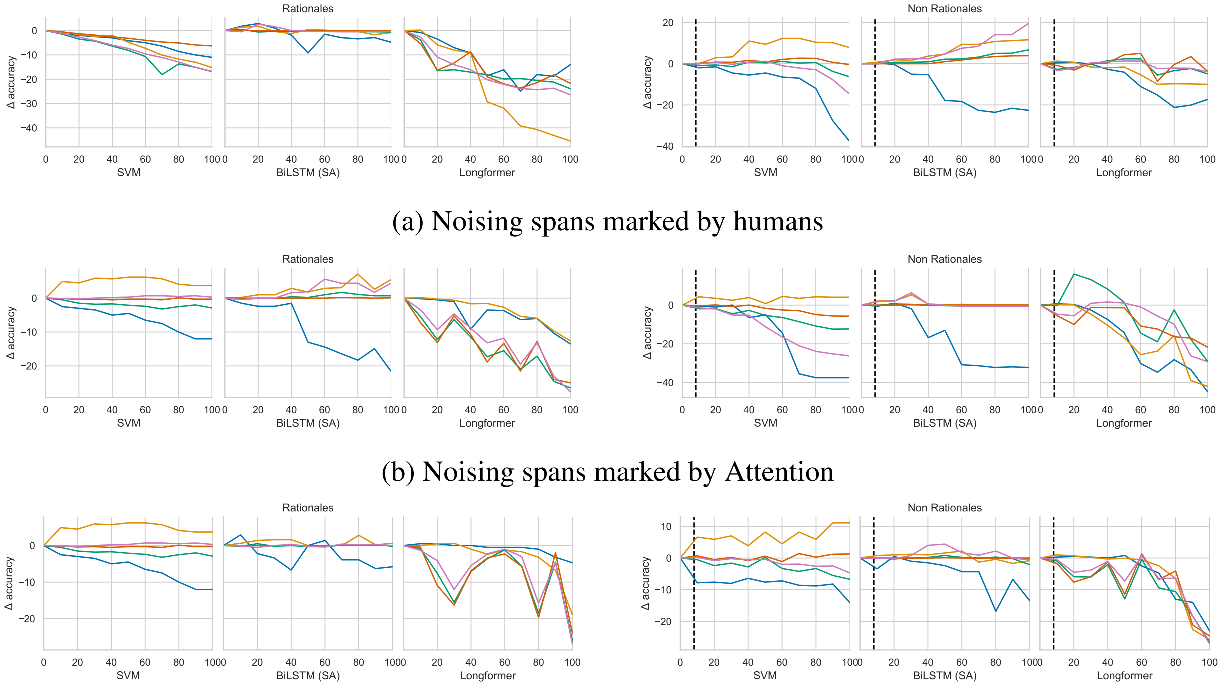
(c) Noising spans marked via gradient based feature attribution

Figure 3.5: Change in classifier accuracy as noise is injected on *rationales/non-rationales* for IMDb reviews from Kaushik et al. [2020].

et al. [2020] obtains 87.8% accuracy on the IMDb test set and 79.9% on Yelp reviews.[4] As a greater fraction of *rationales* are replaced with random words from the vocabulary, the classifier experiences a drop of $\approx 11\%$ by the time all *rationale* tokens are replaced with noise. However, it experiences an 28.7% drop in accuracy on Yelp reviews. Similarly, on the same datasets, a fine-tuned BERT classifier sees its in-sample accuracy drop by 18.4%, and by 31.4% on Yelp as *rationale* tokens replaced by noise go from 0 to 100%. However, as more *non-rationales* are replaced with noise, in-sample accuracy for SVM goes down by $\approx 10\%$ but increases by 1.5% on Yelp. For BERT, in-sample accuracy decreases by only 16.1% and only 13.6% on Yelp (Table 3.10).

We obtain similar results using *rationales* identified via feature feedback. An SVM classifier trained on reviews from Zaidan et al. [2007] sees in-sample accuracy drop by 11%, and accuracy on Yelp drop by 16.9% as noise is inserted on *rationales* but goes down by 17.3% and 14.6%, respectively when noise is inserted in *non-rationales*. For Longformer, in-sample accuracy drops by 14% and accuracy on Yelp goes down by 26.4% compared to a drop of 17.3% and gain of 3.9%, respectively, when noise is inserted in *non-rationales*. Similar patterns are observed across

[4] The out-of-domain evaluation sets in Kaushik et al. [2020] do not have 50:50 label split. We enforce this split to observe when a classifier approaches random baseline performance. All datasets can be found at https://github.com/acmi-lab/counterfactually-augmented-data



(a) Noising spans marked by humans

(b) Noising spans marked by Attention

(c) Noising spans marked via gradient based feature attribution

Figure 3.6: Change in classifier accuracy as noise is injected on *rationales/non-rationales* for IMDb reviews from Zaidan et al. [2007]. In both Figures 3.5 and 3.6, the vertical dashed line indicates the fraction of median length of *non-rationales* equal to the median length of *rationales*.

datasets and models (see Figure 3.6a and Table 3.13).[5]

For NLI, the in-sample accuracy of BERT fine-tuned on an SNLI subsample drops by $\approx 20\%$ when *rationales* are replaced with noise, and out-of-domain accuracy goes down by 21.3–31.5% on various datasets (Table 3.17). Whereas, if *non-rationales* are replaced with noise, in-sample accuracy goes down by 6.2% but out-of-domain accuracy drops by only 2.3–5.5%. These results support our hypothesis that spans marked by humans as causing a label are analogous to causal variables.

Interestingly, in our NLI experiments, for various models the drops in both in-sample and out-of-domain accuracy are greater in magnitude when noise is injected in *rationales* versus when it is injected in *non-rationales*. This is opposite to what we observe in sentiment analysis. We conjecture that these results are due to the fact that in our experiment design for NLI, we only keep those premise-hypothesis pairs that contain at least 10 tokens marked as *rationales* so we can observe the difference in accuracy as the amount of noise increases. A consequence of this selection is that many pairs selected have many more tokens marked as *rationales* than *non-rationales*, whereas, in sentiment analysis this is the opposite. Hence, in NLI when some

[5] While similar trends are observed for both feature feedback and CAD, it is less clear how to incorporate feature feedback for training effectively with deep neural networks and pre-trained transformer architectures, whereas training (or fine-tuning) models on CAD is straightforward.

percentage of *rationales* are replaced by noise, this corresponds to many more edited tokens than when a corresponding percentage of *non-rationales* are noised.

To compare human feedback to automatic feature attribution methods such as attention [Bahdanau et al., 2015] and gradient based saliency methods [Li et al., 2016], we conduct the same set of experiments assuming tokens attended to (or not) by an attention based classifier (BiLSTM with Self-Attention) or identified as highly influential by a gradient based feature attribution method (salience scores) as new *rationales* (or *non-rationales*). In this case, unlike our findings with human feedback, we observe markedly different behavior than predicted by our analysis of the toy causal model (See Figures 3.5b, 3.5c, 3.6b, and 3.6c; and Tables 3.11, 3.12, 3.14, and 3.15).

While we might not expect spurious signals to be as reliable out of domain, that does not mean that they will always fail. For example, while the associations between genre and sentiment learned from a dataset of book reviews might not hold in a dataset of kitchen appliances, but nevertheless hold in a dataset of audiobook reviews. In such settings, even though noising non-causal features would lead to models relying more on causal features, this may not result in better out-of-domain performance.

We also look at whether we really need to go through the process of collecting CAD (or human-annotated rationales) at all or if automated methods for generating "counterfactuals" might obtain similar gains in out-of-domain performance, as the former could be an expensive process. We experiment with state-of-the-art style transfer methods to convert *Positive* reviews into *Negative* and vice versa. Ideally, we would expect these methods to preserve a document's "content" while modifying the attributes that relate to sentiment (if they obtain perfect disentanglement in the feature space). Sentiment classifiers trained on original and *sentiment-flipped* reviews generated using style transfer methods often give better out-of-domain performance compared to training only on original data of same size (Table 3.6). However, models trained on CAD perform even better across all datasets, hinting at the value of human feedback.

Chapter 4

Learning From Feature Feedback

4.1 Overview

Addressing various classification tasks in natural language processing (NLP), including sentiment analysis [Zaidan et al., 2007], natural language inference (NLI) [DeYoung et al., 2020], and propaganda detection [Pruthi et al., 2020], researchers have introduced resources containing additional side information by tasking humans with marking spans in the input text (called *rationales* or *feature feedback)* that provide supporting evidence for the label. For example, spans like “underwhelming”, “horrible”, or “worst film since Johnny English” might indicate negative sentiment in a movie review. Conversely, spans like “exciting”, “amazing”, or “I never thought Vin Diesel would make me cry” might indicate positive sentiment.

These works have proposed a variety of strategies for incorporating feature feedback as additional supervision [Lei et al., 2016, Zhang et al., 2016, Lehman et al., 2019, Chen et al., 2019b, Jain et al., 2020, DeYoung et al., 2020, Pruthi et al., 2020]. Other researchers have studied the learning-theoretic properties of feature feedback [Poulis and Dasgupta, 2017, Dasgupta et al., 2018, Dasgupta and Sabato, 2020]. We focus our study on the resources and practical methods developed for NLP.

Some have used this feedback to perturb instances for data augmentation [Zaidan et al., 2007], while others have explored multitask objectives for simultaneously classifying documents and extracting rationales [Pruthi et al., 2020]. A number of papers exploit feature feedback as intermediate supervision for building extract-then-classify pipelines [Chen et al., 2019b, Lehman et al., 2019, Jain et al., 2020]. One common assumption is that resulting models would learn to identify and rely more on spans relevant to the target labels, which would in turn lead to more accurate predictions.

However, despite their intuitive appeal, feature feedback methods have thus far yielded underwhelming results on independent drawn and identically distributed (iid) test sets in applications involving deep nets. While Zaidan et al. [2007] found significant gains when incorporating rationales into their SVM learning scheme, benefits have been negligible in the BERT era. For example, although Pruthi et al. [2020] and Jain et al. [2020] address a different aim—to improve extraction accuracy—their experiments show no improvement in classification accuracy by incorporating rationales.

On the other hand, Kaushik et al. [2020], introduced counterfactually augmented data (CAD) with the primary aim of showing how supplementary annotations can be incorporated to make models less sensitive to spurious patterns, and additionally demonstrated that models trained on CAD degraded less in a collection of out-of-domain tests than their vanilla counterparts. In followup work, they showed that for both CAD and feature feedback, although corruptions to evidence spans via random word flips result in performance degradation both in- and out-of-domain, when non-evidence spans are corrupted, out-of-domain performance often improves [Kaushik et al., 2021b]. These findings echo earlier results in computer vision [Ross et al., 2017, Ross and Doshi-Velez, 2018] where regularizing input gradients (so-called *local explanations*) to accord with expert attributions led to improved out-of-domain performance.

In this chapter, we conduct an empirical study of the out-of-domain benefits of incorporating feature feedback in NLP. We seek to address two primary research questions: (i) do models that rely on feature feedback generalize better out of domain compared to *classify-only* models (i.e., models trained without feature feedback)? and (ii) do we need to solicit feature feedback for an entire dataset or can significant benefits be realized with a modest fraction of examples annotated? Our experiments on sentiment analysis [Zaidan et al., 2007] and NLI [DeYoung et al., 2020] use both linear, BERT [Devlin et al., 2019], and ELECTRA [Clark et al., 2020b] models, using two feature feedback techniques [Pruthi et al., 2020, Jain et al., 2020].

4.2 Methodology

We focus on two techniques (classify-and-extract [Pruthi et al., 2020] and extract-then-classify [Jain et al., 2020]), two pretrained models, and one (in-domain) dataset each for sentiment analysis and NLI that contain feature feedback. For both techniques, *feature feedback* annotations provide supervision to the extractive component. The classify-and-extract model jointly predicts the (categorical) label and performs sequence tagging predict rationales. The classification head and a linear chain CRF [Lafferty et al., 2001] share an encoder, initialized with pretrained weights.

The extract-then-classify method [Jain et al., 2020] first trains a classifier (*support*) on complete examples to predict the label, using its outputs to extract continuous feature importance scores. These scores are then binarized using a second classifier (*extractor*) which is trained on the feature importance scores from *support* and makes token-level binary predictions to identify rationale tokens in the input. A binary cross-entropy term in the objective of the extractor is used to maximise agreement of the extracted tokens with human rationales. Finally, a third classifier (*predictor*) is trained to predict the target (sentiment or entailment) label based only on these extracted tokens.

For both approaches, we experiment with two pretrained models (BERT and ELECTRA). We limit the maximum sequence length to 512 tokens and train all models for 10 epochs using AdamW optimizer [Loshchilov and Hutter, 2019] with a learning rate of $2e-5$ and a batch size of 8 and early stopping based on mean of classification and extraction F1 scores on the validation set. We replicate all experiments on 5 seeds and report mean performance along with standard deviation.

To see whether results are consistent across architectures, we also use a linear SVM [Zaidan et al., 2007] with a modified objective function on top of the ordinary soft-margin SVM, i.e.,

$$\frac{1}{2}||w||^2 + C(\sum_i \delta_i) + C_{\text{contrast}}(\sum_{i,j} \xi_{ij})$$

subject to the constraints $\vec{w} \cdot \vec{x}_{ij} \cdot y_i \geq 1 - \xi_{ij}\ \forall i, j$ where $\vec{x}_{ij} := \frac{\vec{x}_i - \vec{v}_{ij}}{\mu}$ are *psuedoexamples*, created by subtracting *contrast-examples* ($\vec{v}_{ij}$), input sentence void of randomly chosen rationales, from the original input ($\vec{x}_i$). We use term-frequency embeddings with unigrams appearing in at least 10 reviews and set $C = C_{contrast} = \mu = 1$. For each training example, we generate 5 psuedoexamples.

Datasets For sentiment analysis, we use an IMDb movie reviews dataset [Zaidan et al., 2007]. Reviews in this dataset are labeled as having either *positive* or *negative* sentiment. Zaidan et al. [2007] also tasked annotators to mark spans in each review that were indicative of the overall sentiment. We use these spans as feature feedback. Overall, the dataset has 1800 reviews in the training set (with feature feedback) and 200 in test (without feature feedback). Since the test set does not include ground truth labels for evidence extraction, we construct a test set out of the 1800 examples in the original training set. This leaves 1200 reviews for a new training set, 300 for validation, and 300 for test. For NLI, we use a subsample of the E-SNLI dataset [DeYoung et al., 2020] used in Kaushik et al. [2021b]. In this dataset, there are 6318 premise-hypothesis pairs, equally divided across *entailment* and *contradiction* categories. Examples of feature feedback are shown in Table 4.1.

4.3 Results

We first fine-tune BERT and ELECTRA on a sentiment analysis dataset [Zaidan et al., 2007] following both classify-and-extract and extract-then-classify approaches. We evaluate resulting models on both iid test set as well as various naturally occurring out-of-domain datasets for sentiment analysis and compare resulting performance with classify-only models (Table 4.2). We find that both approaches lead to significant gains in out-of-domain performance compared to the classify-only method. For instance, ELECTRA fine-tuned using the extract-then-classify framework leads to $\approx 15.7\%$ gain in accuracy when evaluated on Yelp.

As Pruthi et al. [2020] demonstrate better performance on evidence extraction for sentiment analysis compared to Jain et al. [2020], we use their method for additional analysis. For both sentiment analysis and NLI, we fine-tune models with varying proportion of samples with rationales and report iid and out-of-domain performance (Tables 4.5 and 4.4). Training with no feature feedback recovers the classify-only baseline. We evaluate on CRD [Kaushik et al., 2020], SST-2 [Socher et al., 2013], Amazon reviews [Ni et al., 2019b], Tweets [Rosenthal et al., 2017] and Yelp reviews [Kaushik et al., 2021b] for sentiment analysis, and Revised Premise (RP), Revised Hypothesis (RH) [Kaushik et al., 2020], MNLI matched (MNLI-M) and mismatched (MNLI-MM) [Williams et al., 2018b] for NLI.

On sentiment analysis, we find feature feedback to improve BERT's iid performance but find ELECTRA's performance comparable with and without feature feedback. Feature feedback

Task	Examples
Sentiment Analysis (Positive)	. . . characters are portrayed with such saddening realism that you can't help but love them , as pathetic as they really are . although levy stands out , guest , willard , o'hara , and posey are all wonderful and definitely should be commended for their performances ! if there was an oscar for an ensemble performance , this is the group that should sweep it . . .
Sentiment Analysis (Negative)	. . . then , as it's been threatening all along , the film explodes into violence . and just when you think it's finally over , schumacher tags on a ridiculous self-righteous finale that drags the whole unpleasant experience down even further . trust me . there are better ways to waste two hours of your life . . .
NLI (Entailment)	**P:** a white dog drinks water on a mountainside. **H:** there is a dog drinking water right now.
NLI (Contradiction)	**P:** a dog leaping off a boat **H:** dogs drinking water from pond

Table 4.1: Examples of documents (and true label) with feature feedback (highlighted in yellow).

Test set	Classify-only	Pruthi et al.	Jain et al.
	BERT		
In-domain	$85.9_{0.7}$	$\mathbf{89.9_{2.3}}$	$\mathbf{90.4_{0.3}}$
CRD	$89.3_{0.7}$	$\mathbf{91.6_{0.7}}$	$87.5_{0.8}$
SST2	$77.6_{4.1}$	$79.3_{3.6}$	$75.6_{1.2}$
Amazon	$78.1_{4.9}$	$83.5_{3.1}$	$\mathbf{92.3_{1.2}}$
Semeval	$70.6_{5.7}$	$73.2_{2.6}$	$68.6_{2.2}$
Yelp	$86.8_{1.7}$	$85.7_{1.6}$	$\mathbf{91.6_{0.1}}$
	ELECTRA		
In-domain	$93.2_{0.3}$	$91.8_{1.4}$	$93.1_{0.3}$
CRD	$91.6_{0.4}$	$\mathbf{93.7_{0.9}}$	$91.5_{0.7}$
SST2	$73.2_{1.3}$	$74.0_{1.2}$	$\mathbf{77.2_{1.4}}$
Amazon	$72.8_{2.0}$	$75.5_{2.1}$	$\mathbf{84.2_{1.6}}$
Semeval	$67.5_{4.5}$	$72.5_{1.8}$	$66.7_{3.0}$
Yelp	$79.0_{3.6}$	$\mathbf{84.6_{1.8}}$	$\mathbf{94.7_{0.2}}$

Table 4.2: Mean and standard deviation (in subscript) of accuracy scores of classify-only models, and models proposed by Pruthi et al. [2020] and Jain et al. [2020], fined-tuned for sentiment analysis. Results highlighted in bold are significant with $p < 0.05$.

Test set	Classify-only	Zaidan et al.
In-domain	$75.2_{3.5}$	$79.1_{3.4}$
CRD	$48.3_{2.0}$	$\mathbf{58.2_{2.4}}$
SST-2	$49.7_{0.3}$	$\mathbf{65.6_{1.5}}$
Amazon	$50.9_{0.3}$	$\mathbf{68.7_{3.1}}$
Semeval	$49.8_{0.1}$	$\mathbf{58.0_{1.5}}$
Yelp	$55.7_{2.8}$	$\mathbf{74.8_{2.7}}$

Table 4.3: Mean and standard deviation (in subscript) of accuracy scores of classify-only SVM model versus SVM trained with feature feedback for sentiment analysis using Zaidan et al. [2007]'s method. Results highlighted in bold are significant with $p < 0.05$.

Test set	Classify-only	Pruthi et al.	Jain et al.
BERT			
In-domain	$88.7_{2.0}$	$89.8_{0.8}$	$77.7_{0.1}$
RP	$62.9_{3.9}$	$66.6_{0.6}$	$57.9_{0.1}$
RH	$76.9_{3.5}$	$80.5_{1.9}$	$70.7_{0.2}$
MNLI-M	$69.7_{2.6}$	$68.1_{1.9}$	$69.8_{0.1}$
MNLI-MM	$71.5_{2.7}$	$69.2_{2.3}$	$66.2_{0.1}$
ELECTRA			
In-domain	$\mathbf{96.0_{0.2}}$	$95.0_{0.3}$	$85.4_{0.04}$
RP	$80.8_{1.0}$	$78.0_{0.6}$	$72.2_{0.1}$
RH	$88.9_{1.0}$	$88.7_{0.9}$	$79.7_{0.1}$
MNLI-M	$86.5_{0.9}$	$81.9_{2.1}$	$77.1_{0.1}$
MNLI-MM	$86.6_{0.8}$	$82.1_{2.0}$	$75.7_{0.1}$

Table 4.4: Mean and standard deviation (in subscript) of F1 scores of models fine-tuned for NLI with increasing number of examples with feature feedback. Results highlighted in bold are significantly better than classify-only performance ($p < 0.05$).

leads to an increase in performance out-of-domain on both BERT and ELECTRA. For instance, with feature feedback, ELECTRA's classification accuracy increases from 91.6% to 93.7% on CRD and 79% to 84.6% on Yelp. Similar trends are also observed when we fine-tune BERT with feature feedback. Interestingly, when evaluated on the SemEval dataset (Tweets), we observe that BERT fine-tuned with feature feedback on all training examples achieves comparable performance to fine-tuning without feature feedback. However, fine-tuning with feature feedback on just 25% of training examples leads to a significant improvement in classification accuracy. We speculate that this might be a result of implicit hyperparameter tuning when combining prediction and extraction losses, and a more extensive hyperparameter search could provide comparable (if not better) gains with 100% data. Similarly, SVM trained with feature feedback [Zaidan et al., 2007] consistently outperformed SVM trained without feature feedback, when evaluated out-of-domain despite obtaining similar accuracy in-domain (Tables 4.3 and 4.6). For instance, SVM

	Fraction of Training Data with Rationales				
Evaluation set	No rationales	25%	50%	75%	100%
		BERT			
In-domain	$85.9_{0.7}$	$\mathbf{87.7_{1.1}}$	$88.1_{2.4}$	$\mathbf{90.2_{1.5}}$	$\mathbf{89.9_{2.3}}$
CRD	$89.3_{0.7}$	$\mathbf{91.7_{0.6}}$	$\mathbf{92.3_{0.9}}$	$\mathbf{92.3_{0.3}}$	$\mathbf{91.6_{0.7}}$
SST2	$77.6_{4.1}$	$81.2_{0.6}$	$81.3_{0.7}$	$81.8_{0.6}$	$79.3_{3.6}$
Amazon	$78.1_{4.9}$	$\mathbf{85.3_{1.2}}$	$\mathbf{84.6_{1.7}}$	$\mathbf{84.0_{0.5}}$	$83.5_{3.1}$
Semeval	$70.6_{5.7}$	$\mathbf{77.8_{1.0}}$	$75.5_{0.8}$	$74.9_{0.8}$	$73.2_{2.6}$
Yelp	$86.8_{1.7}$	$86.9_{1.1}$	$85.8_{1.5}$	$85.4_{0.7}$	$85.7_{1.6}$
		ELECTRA			
In-domain	$93.2_{0.3}$	$92.4_{0.9}$	$92.8_{1.2}$	$93.7_{1.9}$	$91.8_{1.4}$
CRD	$91.6_{0.4}$	$92.1_{0.8}$	$\mathbf{93.0_{0.6}}$	$\mathbf{93.1_{0.3}}$	$\mathbf{93.7_{0.9}}$
SST2	$73.2_{1.3}$	$73.1_{1.8}$	$72.3_{1.6}$	$72.3_{1.1}$	$74.0_{1.2}$
Amazon	$72.8_{2.0}$	$\mathbf{79.0_{1.8}}$	$\mathbf{75.7_{1.2}}$	$\mathbf{76.6_{1.8}}$	$75.5_{2.1}$
Semeval	$67.5_{4.5}$	$70.5_{1.5}$	$66.2_{1.5}$	$67.1_{2.2}$	$72.5_{1.8}$
Yelp	$79.0_{3.6}$	$\mathbf{84.5_{1.1}}$	$\mathbf{84.2_{1.7}}$	$\mathbf{84.3_{1.2}}$	$\mathbf{84.6_{1.8}}$

Table 4.5: Mean and standard deviation (in subscript) of accuracy scores of models fine-tuned for sentiment analysis using the method proposed by Pruthi et al. [2020] with different base models (BERT and ELECTRA) and increasing proportion of examples with feature feedback. Results highlighted in bold are significant difference with $p < 0.05$.

	Dataset size			
Evaluation Set	300	600	900	1200
In-domain	$77.0_{3.9}/77.6_{2.2}$	$78.5_{3.2}/82.3_{2.0}$	$80.5_{1.7}/\mathbf{84.9_{1.6}}$	$75.2_{3.5}/79.1_{3.4}$
CRD	$48.0_{2.9}/\mathbf{56.4_{1.3}}$	$48.3_{2.5}/\mathbf{58.0_{2.7}}$	$48.4_{2.3}/\mathbf{58.7_{1.8}}$	$48.3_{2.0}/\mathbf{58.2_{2.4}}$
SST-2	$52.2_{1.6}/\mathbf{62.9_{1.0}}$	$50.9_{3.0}/\mathbf{64.0_{0.9}}$	$51.3_{3.1}/\mathbf{64.9_{0.9}}$	$49.7_{0.3}/\mathbf{65.6_{1.5}}$
Amazon	$51.8_{1.5}/\mathbf{65.9_{1.9}}$	$52.4_{2.0}/\mathbf{66.5_{1.2}}$	$52.0_{2.9}/\mathbf{69.9_{0.4}}$	$50.9_{0.3}/\mathbf{68.7_{3.1}}$
Semeval	$50.3_{1.4}/\mathbf{56.7_{1.1}}$	$50.3_{1.2}/\mathbf{56.4_{0.8}}$	$50.1_{0.5}/\mathbf{58.8_{1.3}}$	$49.8_{0.1}/\mathbf{58.0_{1.5}}$
Yelp	$60.2_{4.0}/\mathbf{72.0_{2.4}}$	$57.3_{7.1}/\mathbf{74.5_{1.5}}$	$61.2_{4.6}/\mathbf{74.8_{2.5}}$	$55.7_{2.8}/\mathbf{74.8_{2.7}}$

Table 4.6: Mean and standard deviation (in subscript) of accuracy scores of classify-only SVM models (left) presented alongside accuracy scores of models trained with feature feedback (right), with increasing number of training-samples for sentiment analysis using the method proposed by Zaidan et al. [2007]. Results highlighted in bold show statistically significant difference with $p < 0.05$.

trained on just label information achieved $75.2\% \pm 3.5\%$ accuracy on the in-domain test set, which was comparable to the accuracy of $79.1\% \pm 3.4\%$ achieved by SVM trained with feature feedback. But the classifier trained with feature feedback led to $\approx 19\%$ and $\approx 18\%$ improvement in classification accuracy on Yelp reviews and Amazon reviews, respectively, compared to the

	Fraction of Training Data with Rationales				
Evaluation set	No rationales	25%	50%	75%	100%
		BERT			
In-domain	$88.7_{2.0}$	$89.6_{0.4}$	$89.9_{0.4}$	$89.7_{0.4}$	$89.8_{0.8}$
RP	$62.9_{3.9}$	$67.6_{2.0}$	$67.4_{1.2}$	$68.6_{0.6}$	$66.6_{0.6}$
RH	$76.9_{3.5}$	$80.4_{1.1}$	$81.7_{1.6}$	$81.4_{0.7}$	$80.5_{1.9}$
MNLI-M	$69.7_{2.6}$	$67.6_{3.4}$	$68.1_{4.6}$	$68.8_{2.0}$	$68.1_{1.9}$
MNLI-MM	$71.5_{2.7}$	$68.8_{4.5}$	$69.2_{5.9}$	$69.8_{2.7}$	$69.2_{2.3}$
		ELECTRA			
In-domain	$\mathbf{96.0_{0.2}}$	$95.1_{0.3}$	$95.0_{0.3}$	$95.0_{0.3}$	$95.0_{0.3}$
RP	$80.8_{1.0}$	$78.2_{1.3}$	$79.2_{1.1}$	$77.2_{1.3}$	$78.0_{0.6}$
RH	$88.9_{1.0}$	$88.0_{1.2}$	$88.4_{0.3}$	$87.9_{0.4}$	$88.7_{0.9}$
MNLI-M	$86.5_{0.9}$	$82.0_{2.8}$	$82.4_{1.6}$	$82.3_{0.9}$	$81.9_{2.1}$
MNLI-MM	$86.6_{0.8}$	$82.6_{2.8}$	$83.5_{1.4}$	$82.6_{0.8}$	$82.1_{2.0}$

Table 4.7: Mean and standard deviation (in subscript) of F-1 scores of models fine-tuned for NLI using the method proposed by Pruthi et al. [2020] with different base models (BERT and ELECTRA) and increasing proportion of examples with feature feedback. Results highlighted in bold are significant difference with $p < 0.05$.

classifier trained without feature feedback.

For NLI, it appears that feature feedback provides no added benefit compared to a classify-only BERT model, whereas, ELECTRA's iid performance decreases with feature feedback. Furthermore, models fine-tuned with feature feedback generally perform no better than classify-only models when trained with varying proportions of rationales (Table 4.7) while classify-only models perform significantly better than the models trained with rationales when trained with varying dataset size. (Table 4.8). These results are in line with observations in prior work on counterfactually augmented data [Huang et al., 2020]. Table 4.9 shows examples of correct and incorrect predictions made on out of domain examples by models trained with feature feedback.

4.4 Analysis and Discussion

We find that 21.37% of tokens in the vocabulary of Zaidan et al. [2007] are marked as rationales in at least one movie review. Interestingly, this fraction is 79.54% for NLI (Table 4.10). While for movie reviews, certain words or phrases might generally denote positive or negative sentiment (e.g., "amazing movie"), for NLI tasks, it is not clear that any individual phrase should suggest entailment or contradiction generally. A word or a phrase might be marked as indicating entailment in one NLI example but as contradiction in another.

We further construct vocabulary of unigrams and bigrams from phrases marked as feature feedback in examples from the sentiment analysis training set ($V_{\text{rationale}}$). We compute the fraction of unigrams (and bigrams) that occur in this vocabulary and also occur in each out-of-domain

	Dataset size			
Evaluation Set	1500	3000	4500	6318
		BERT		
In-domain	$85.9_{6.0}$/$84.5_{2.0}$	$87.9_{0.4}$/$87.7_{1.0}$	$89.1_{0.4}$/$89.2_{0.2}$	$88.7_{2.0}$/$89.8_{0.8}$
RP	$61.8_{0.9}$/$62.8_{1.8}$	$63.3_{1.6}$/$64.2_{1.8}$	$63.7_{1.8}$/$\mathbf{66.8_{1.4}}$	$62.9_{3.9}$/$66.4_{1.7}$
RH	$74.5_{1.6}$/$71.8_{3.4}$	$77.0_{1.4}$/$77.3_{2.1}$	$78.3_{1.1}$/$80.4_{1.8}$	$76.9_{3.5}$/$80.5_{1.9}$
MNLI-M	$63.7_{3.1}$/$60.8_{3.2}$	$69.2_{1.8}$/$66.3_{2.2}$	$70.2_{0.9}$/$67.5_{3.1}$	$69.7_{2.6}$/$68.1_{1.9}$
MNLI-MM	$64.8_{4.3}$/$61.8_{4.3}$	$71.3_{2.3}$/$67.5_{2.8}$	$72.1_{1.2}$/$68.9_{4.2}$	$73.1_{1.9}$/$71.4_{1.1}$
		ELECTRA		
In-domain	$\mathbf{94.6_{0.2}}$/$92.7_{0.5}$	$\mathbf{95.1_{0.4}}$/$94.2_{0.3}$	$\mathbf{95.7_{0.2}}$/$94.4_{0.2}$	$\mathbf{96.0_{0.2}}$/$95.1_{0.3}$
RP	$78.4_{1.2}$/$75.2_{2.5}$	$78.5_{1.8}$/$77.2_{0.9}$	$\mathbf{81.2_{0.6}}$/$76.2_{1.2}$	$\mathbf{80.8_{1.0}}$/$78.0_{0.6}$
RH	$\mathbf{87.7_{0.7}}$/$85.2_{1.4}$	$88.1_{1.3}$/$87.3_{0.6}$	$\mathbf{89.4_{0.6}}$/$87.1_{1.0}$	$88.9_{1.0}$/$88.7_{0.9}$
MNLI-M	$\mathbf{82.8_{2.2}}$/$77.0_{1.8}$	$\mathbf{85.4_{1.8}}$/$78.9_{1.7}$	$\mathbf{86.0_{1.6}}$/$80.4_{2.1}$	$\mathbf{86.5_{0.9}}$/$81.9_{2.1}$
MNLI-MM	$\mathbf{83.6_{2.5}}$/$77.9_{2.1}$	$\mathbf{86.2_{2.1}}$/$79.9_{1.9}$	$\mathbf{86.1_{1.8}}$/$80.8_{2.2}$	$\mathbf{86.6_{0.8}}$/$82.1_{2.0}$

Table 4.8: Mean and standard deviation (in subscript) of F-1 scores of classify-only models/models trained with feature feedback, with increasing number of training-samples for NLI using the method proposed by Pruthi et al. [2020]. Results highlighted in bold are statistically significant difference with $p < 0.05$.

dataset. We find that a large fraction of unigrams from $V_{\text{rationale}}$ also exist in CRD ($\approx 60\%$), SST2 ($\approx 64\%$), and Yelp ($\approx 78\%$) data. However, this overlap is much smaller for SemEval ($\approx 30\%$) and Amazon ($\approx 45\%$). For these overlapping unigrams, we observe a relatively large percentage (50–65%) preserve their associated majority training set label in the out-of-domain datasets. Similar trends hold for bigrams, though fewer $V_{\text{rationale}}$ bigrams are present out-of-domain (Table 4.12).

For each pair in the NLI training set, we compute Jaccard similarity between the premise and hypothesis sentence (Table 4.11). We compute the mean of these example-level values over the entire dataset, finding that it is common for examples in our training set to have overlap between premise and hypothesis sentences, regardless of the label. However, when we compute mean Jaccard similarity between premise and hypothesis rationales, we find higher overlap for entailment examples versus contradiction. Thus, models trained with feature feedback might learn to identify word overlap as predictive of entailment even when the true label is contradiction. While this may not improve an NLI model's performance, it could be useful in tasks like Question Answering, where answers often lie in sentences that have high word overlap with the question [Lamm et al., 2020, Majumder et al., 2021]. Interestingly, our results on NLI are in conflict with recent findings where models trained with rationales showed significant improvement over classify-only models in both iid and out-of-domain (MNLI-M and MNLI-MM) settings [Stacey et al., 2021]. This could be due to the different modeling strategy employed in their work, as they use rationales to guide the training of the classifier's attention module. Investigating this difference is left for future work.

Task	Examples
Sentiment Analysis (Positive, Correct)	everyone should adapt a tom robbins book for screen . while the movie is fine and the performances are good , the dialogue , which works well reading it , is beautiful when spoken .
Sentiment Analysis (Positive, Wrong)	... very uncaptivating yet one gets the feeling that their is some serious exploitation going on here ...
Sentiment Analysis (Negative, Correct)	... using quicken is a frustrating experience each time i fire it up ...
Sentiment Analysis (Negative, Wrong)	... with many cringe-worthy 'surprises', which happen around 10 minutes after you see exactly what's going to happen ...
NLI (Entailment, Correct)	**P:** a woman cook in an apron is smiling at the camera with two other cooks in the background . **H:** a woman looking at the camera .
NLI (Entailment, Wrong)	**P:** a woman in a brown dress looking at papers in front of a class . **H:** a woman looking at papers in front of a class is not wearing a blue dress .
NLI (Contradiction, Correct)	**P:** the woman in the white dress looks very uncomfortable in the busy surroundings **H:** the dress is black .
NLI (Contradiction, Wrong)	**P:** a man , wearing a cap , is pushing a cart , on which large display boards are kept , on a road . **H:** the person is pulling large display boards on a cart .

Table 4.9: Examples (from out-of-domain evaluation sets; with true label and model prediction) of explanations highlighted by feature feedback models (highlighted in yellow).

Task	Unigram	Bigram
Sentiment Analysis	21.37	11.20
NLI	79.54	35.49

Table 4.10: Percentage of unigram and bigram vocabularies that are marked as feature feedback at least once.

	Entailment	Contradiction
D_{all}	0.25	0.16
$D_{\text{rationale}}$	0.30	0.09

Table 4.11: Mean Jaccard index of premise-hypothesis word overlap (D_{all}) and rationale overlap ($D_{\text{rationale}}$) in the training set.

Dataset	% Overlap	Label Agreement
Unigram		
CRD	60.3	51.3
SST2	64.6	66.5
Amazon	45.6	47.6
Semeval	30.9	60.3
Yelp	78.3	65.1
Bigram		
CRD	28.2	51.9
SST2	28.5	64.5
Amazon	19.6	49.9
Semeval	10.2	58.5
Yelp	46.8	65.3

Table 4.12: Rationale vocabulary overlap and label agreement between in-sample and OOD datasets.

Chapter 5

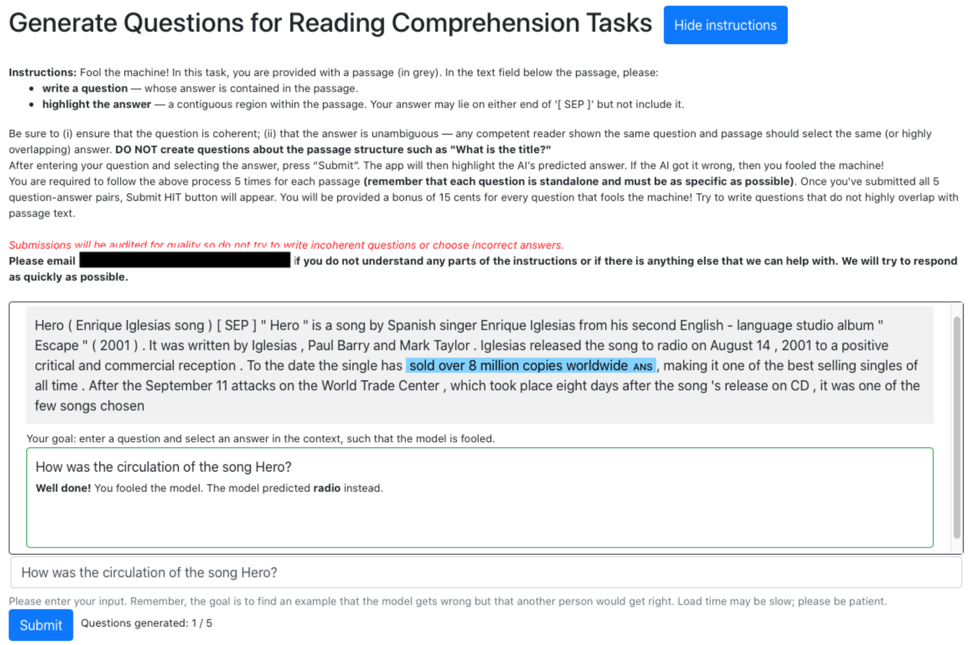
Adversarial Data Collection

5.1 Overview

Across such diverse natural language processing (NLP) tasks as natural language inference [NLI; Poliak et al., 2018, Gururangan et al., 2018], question answering [QA; Kaushik and Lipton, 2018], and sentiment analysis [Kaushik et al., 2020], researchers have discovered that models can succeed on popular benchmarks by exploiting spurious associations that characterize a particular dataset but do not hold more widely. Despite performing well on independent and identically distributed (i.i.d.) data, these models are liable under plausible domain shifts. With the goal of providing more challenging benchmarks that require this stronger form of generalization, an emerging line of research has investigated *adversarial data collection* (ADC), a scheme in which a worker interacts with a model (in real time), attempting to produce examples that elicit incorrect predictions [e.g., Dua et al., 2019, Nie et al., 2020]. The hope is that by identifying parts of the input domain where the model fails one might make the model more robust. Researchers have shown that models trained on ADC perform better on such adversarially collected data and that with successive rounds of ADC, crowdworkers are less able to fool the models [Dinan et al., 2019].

While adversarial data may indeed provide more challenging benchmarks, the process and its actual benefits vis-a-vis tasks of interest remain poorly understood, raising several key questions: (i) do the resulting models typically generalize better out of distribution compared to standard data collection (SDC)?; (ii) how much can differences between ADC and SDC be attributed to the way workers behave when attempting to fool models, regardless of whether they are successful? and (iii) what is the impact of training models on adversarial data only, versus using it as a data augmentation strategy?

In this chapter, we describe a large-scale randomized controlled study to address these questions. Focusing our study on span-based question answering and a variant of the Natural Questions dataset [NQ; Lee et al., 2019, Karpukhin et al., 2020], we work with two popular pretrained transformer architectures—BERT$_{\text{large}}$ [Devlin et al., 2019] and ELECTRA$_{\text{large}}$ [Clark et al., 2020a]—each fine-tuned on $23.1k$ examples. To eliminate confounding factors when assessing the impact of ADC, we randomly assign the crowdworkers tasked with generating questions to one of three groups: (i) with an incentive to fool the BERT model; (ii) with an in-



Generate Questions for Reading Comprehension Tasks

Instructions: Fool the machine! In this task, you are provided with a passage (in grey). In the text field below the passage, please:

- **write a question** — whose answer is contained in the passage.
- **highlight the answer** — a contiguous region within the passage. Your answer may lie on either end of '[SEP]' but not include it.

Be sure to (i) ensure that the question is coherent; (ii) that the answer is unambiguous — any competent reader shown the same question and passage should select the same (or highly overlapping) answer. **DO NOT create questions about the passage structure such as "What is the title?"**
After entering your question and selecting the answer, press "Submit". The app will then highlight the AI's predicted answer. If the AI got it wrong, then you fooled the machine!
You are required to follow the above process 5 times for each passage **(remember that each question is standalone and must be as specific as possible)**. Once you've submitted all 5 question-answer pairs, Submit HIT button will appear. You will be provided a bonus of 15 cents for every question that fools the machine! Try to write questions that do not highly overlap with passage text.

Submissions will be audited for quality so do not try to write incoherent questions or choose incorrect answers.
Please email [redacted] if you do not understand any parts of the instructions or if there is anything else that we can help with. We will try to respond as quickly as possible.

Hero (Enrique Iglesias song) [SEP] " Hero " is a song by Spanish singer Enrique Iglesias from his second English - language studio album " Escape " (2001) . It was written by Iglesias , Paul Barry and Mark Taylor . Iglesias released the song to radio on August 14 , 2001 to a positive critical and commercial reception . To the date the single has sold over 8 million copies worldwide ANS , making it one of the best selling singles of all time . After the September 11 attacks on the World Trade Center , which took place eight days after the song 's release on CD , it was one of the few songs chosen

Your goal: enter a question and select an answer in the context, such that the model is fooled.

How was the circulation of the song Hero?

Well done! You fooled the model. The model predicted **radio** instead.

How was the circulation of the song Hero?

Please enter your input. Remember, the goal is to find an example that the model gets wrong but that another person would get right. Load time may be slow; please be patient.

Questions generated: 1 / 5

Figure 5.1: Platform shown to workers generating questions in the ADC setting.

centive to fool the ELECTRA model; and (iii) a standard, non-adversarial setting (no model in the loop). The pool of contexts is the same for each group and each worker is asked to generate five questions for each context that they see. Workers are shown similar instructions (with minimal changes), and paid the same base amount.

We fine-tune three models (BERT, RoBERTa, and ELECTRA) on resulting datasets and evaluate them on held-out test sets, adversarial test sets from prior work [Bartolo et al., 2020], and 12 MRQA [Fisch et al., 2019] datasets. For all models, we find that while fine-tuning on adversarial data usually leads to better performance on (previously collected) adversarial data, it typically leads to worse performance on a large, diverse collection of out-of-domain datasets (compared to fine-tuning on standard data). We observe a similar pattern when augmenting the existing dataset with the adversarial data. Results on an extensive collection of out-of-domain evaluation sets suggest that ADC training data does not offer clear benefits vis-à-vis robustness under distribution shift.

To study the differences between adversarial and standard data, we perform a qualitative analysis, categorizing questions based on a taxonomy [Hovy et al., 2000]. We notice that more questions in the ADC dataset require numerical reasoning compared to the SDC sample. These qualitative insights may offer additional guidance to future researchers.

5.2 Related Work

In an early example of model-in-the-loop data collection, Zweig and Burges [2012] use n-gram language models to suggest candidate incorrect answers for a fill-in-the-blank task. Richardson et al. [2013] suggested ADC for QA as proposed future work, speculating that it might challenge state-of-the-art models. In the *Build It Break It, The Language Edition* shared task [Ettinger et al., 2017], teams worked as *builders* (training models) and *breakers* (creating challenging examples for subsequent training) for sentiment analysis and QA-SRL.

Research on ADC has picked up recently, with Chen et al. [2019a] tasking crowdworkers to construct multiple-choice questions to fool a BERT model and Wallace et al. [2019b] employing Quizbowl community members to write Jeopardy-style questions to compete against QA models. Zhang et al. [2018] automatically generated questions from news articles, keeping only those questions that were incorrectly answered by a QA model. Dua et al. [2019] and Dasigi et al. [2019] required crowdworkers to submit only questions that QA models answered incorrectly. To construct FEVER 2.0 [Thorne et al., 2019], crowdworkers were required to fool a fact-verification system trained on the FEVER [Thorne et al., 2018] dataset. Some works explore ADC over multiple rounds, with adversarial data from one round used to train models in the subsequent round. Yang et al. [2018b] ask workers to generate challenging datasets working first as adversaries and later as collaborators. Dinan et al. [2019] build on their work, employing ADC to address offensive language identification. They find that over successive rounds of training, models trained on ADC data are harder for humans to fool than those trained on standard data. Nie et al. [2020] applied ADC for an NLI task over three rounds, finding that training for more rounds improves model performance on adversarial data, and observing improvements on the original evaluations set when training on a mixture of original and adversarial training data. Williams et al. [2020] conducted an error analysis of model predictions on the datasets collected by Nie et al. [2020]. Bartolo et al. [2020] studied the empirical efficacy of ADC for SQuAD [Rajpurkar et al., 2016], observing improved performance on adversarial test sets but noting that trends vary depending on the models used to collect data and to train. Previously, Lowell et al. [2019] observed similar issues in active learning, when the models used to acquire data and for subsequent training differ. Yang et al. [2018a], Zellers et al. [2018, 2019] first collect datasets and then filter examples based on predictions from a model. Paperno et al. [2016] apply a similar procedure to generate a language modeling dataset (LAMBADA). Kaushik et al. [2020, 2021b] collect counterfactually augmented data (CAD) by asking crowdworkers to edit existing documents to make counterfactual labels applicable, showing that models trained on CAD generalize better out-of-domain.

Absent further assumptions, learning classifiers robust to distribution shift is impossible [Ben-David et al., 2010]. While few NLP papers on the matter make their assumptions explicit, they typically proceed under the implicit assumptions that the labeling function is deterministic (there is one right answer), and that *covariate shift* [Shimodaira, 2000] applies (the labeling function $p(y|x)$ is invariant across domains). Note that neither condition is generally true of prediction problems. For example, faced with label shift [Schölkopf et al., 2012, Lipton et al., 2018] $p(y|x)$ can change across distributions, requiring one to adapt the predictor to each environment.

5.3 Study Design

In our study of ADC for QA, each crowdworker is shown a short passage and asked to create 5 questions and highlight answers (spans in the passage, see Fig. 5.1). We provide all workers with the same base pay and for those assigned to ADC, pay out an additional bonus for each question that fools the QA model. Finally, we field a different set of workers to validate the generated examples.

Context passages For context passages, we use the first 100 words of Wikipedia articles. Truncating the articles keeps the task of generating questions from growing unwieldy. These segments typically contain an overview, providing ample material for factoid questions. We restrict the pool of candidate contexts by leveraging a variant of the Natural Questions dataset Kwiatkowski et al. [2019], Lee et al. [2019]. We first keep only a subset of $23.1k$ question/answer pairs for which the context passages are the first 100 words of Wikipedia articles[1]. From these passages, we sample $10k$ at random for our study.

Models in the loop We use BERT$_{\text{large}}$ [Devlin et al., 2019] and ELECTRA$_{\text{large}}$ [Clark et al., 2020a] models as our adversarial models in the loop, using the implementations provided by Wolf et al. [2020]. We fine-tune these models for span-based question-answering, using the $23.1k$ training examples (subsampled previously) for 20 epochs, with early-stopping based on word-overlap F1[2] over the validation set. Our BERT model achieves an EM score of 73.1 and an F1 score of 80.5 on an i.i.d. validation set. The ELECTRA model performs slightly better, obtaining an 74.2 EM and 81.2 F1 on the same set.

Crowdsourcing protocol We build our crowdsourcing platform on the Dynabench interface [Kiela et al., 2021] and use Amazon's Mechanical Turk to recruit workers to write questions. To ensure high quality, we restricted the pool to U.S. residents who had already completed at least 1000 HITs and had over 98% HIT approval rate. For each task, we conducted several pilot studies to gather feedback from crowdworkers on the task and interface. We identified median time taken by workers to complete the task in our pilot studies and used that to design the incentive structure for the main task. We also conducted multiple studies with different variants of instructions to observe trends in the quality of questions and refined our instructions based on feedback from crowdworkers. Feedback from the pilots also guided improvements to our crowdsourcing interface. In total, 984 workers took part in the study, with 741 creating questions. In our final study, we randomly assigned workers to generate questions in the following ways: (i) to fool the BERT baseline; (ii) to fool the ELECTRA baseline; or (iii) without a model in the loop. Before beginning the task, each worker completes an onboarding process to familiarize them with the platform. We present the same set of passages to workers regardless of which group they are assigned to, tasking them with generating 5 questions for each passage.

Incentive structure During our pilot studies, we found that workers spend $\approx$ 2–3 minutes to generate 5 questions. We provide workers with the same base pay—$\$0.75$ per HIT—(to ensure compensation at a $\$15$/hour rate). For tasks involving a model in the loop, we define a model prediction to be *incorrect* if its F1 score is less than 40%, following the threshold set by Bartolo

[1] We used the data prepared by Karpukhin et al. [2020], available at https://www.github.com/facebookresearch/DPR.

[2] Word-overlap F1 and Exact Match (EM) metrics introduced in Rajpurkar et al. [2016] are commonly used to evaluate performance of passage-based QA systems, where the correct answer is a span in the given passage.

Resource	Num. Passages			Num. QA Pairs		
	Train	Val	Test	Train	Val	Test
BERT	3,412	992	1,056	11,330	1,130	1,130
ELECTRA	3,925	1,352	1,352	14,556	1,456	1,456

Table 5.1: Number of unique passages and question-answer pairs for each data resource.

et al. [2020]. Workers tasked with fooling the model receive bonus pay of $\$0.15$ for every question that leads to an incorrect model prediction. This way, a worker can double their pay if all 5 of their generated questions induce incorrect model predictions.

Quality control Upon completion of each batch of our data collection process, we presented $\approx 20\%$ of the collected questions to a fourth group of crowdworkers who were tasked with validating whether the questions were answerable and the answers were correctly labeled. In addition, we manually verified a small fraction of the collected question-answer pairs. If validations of at least 20% of the examples generated by a particular worker were incorrect, their work was discarded in its entirety. The entire process, including the pilot studies cost $\approx \$50k$ and spanned a period of seven months. Through this process, we collected over $150k$ question-answer pairs corresponding to the $10k$ contexts ($50k$ from each group) but the final datasets are much smaller, as we explain below.

5.4 Experiments

Our study allows us to answer three questions: (i) how well do models fine-tuned on ADC data generalize to unseen distributions compared to fine-tuning on SDC? (ii) Among the differences between ADC and SDC, how many are due to workers trying to fool the model regardless of whether they are successful? and (iii) what is the impact of training on adversarial data only versus using it as a data augmentation strategy?

Datasets For both BERT and ELECTRA, we first identify contexts for which at least one question elicited an incorrect model prediction. Note that this set of contexts is different for BERT and ELECTRA. For each such context c, we identify the number of questions k_c (out of 5) that successfully fooled the model. We then create 3 datasets per model by, for each context, (i) choosing precisely those k_c questions that fooled the model (BERT$_{\text{fooled}}$ and ELECTRA$_{\text{fooled}}$); (ii) randomly choosing k_c questions (out of 5) from ADC data without replacement (BERT$_{\text{random}}$ and ELECTRA$_{\text{random}}$)—regardless of whether they fooled the model; and (iii) randomly choosing k_c questions (out of 5) from the SDC data without replacement. Thus, we create 6 datasets, where all 3 BERT datasets have the same number of questions per context (and $11.3k$ total training examples), while all 3 ELECTRA datasets likewise share the same number of questions per context (and $14.7k$ total training examples). See Table 5.1 for details on the number of passages and question-answer pairs used in the different splits.

Evaluation set →	$\text{BERT}_{\text{fooled}}$		$\text{BERT}_{\text{random}}$		SDC		Original Dev.	
Training set ↓	EM	F1	EM	F1	EM	F1	EM	F1
Finetuned model: $\text{BERT}_{\text{large}}$								
Original (O; 23.1k)	0.0	17.1	29.6	45.2	32.5	49.1	73.3	80.5
Original (11.3k)	$8.4_{0.9}$	$18.7_{0.6}$	$28.8_{0.5}$	$42.7_{0.9}$	$33.1_{0.7}$	$48.6_{1.1}$	$66.1_{0.3}$	$74.2_{0.4}$
$\text{BERT}_{\text{fooled}}$ (F; 11.3k)	$34.4_{5.1}$	$57.0_{5.7}$	$44.0_{8.8}$	$61.7_{8.2}$	$47.5_{10.0}$	$66.8_{8.6}$	$34.5_{2.6}$	$47.9_{3.3}$
$\text{BERT}_{\text{random}}$ (R; 11.3k)	$37.7_{2.7}$	$58.9_{2.5}$	$57.0_{4.5}$	$73.9_{3.5}$	$62.4_{4.5}$	$79.7_{3.1}$	$46.4_{3.1}$	$60.6_{3.8}$
SDC (11.3k)	$33.6_{0.3}$	$54.4_{0.4}$	$57.6_{0.6}$	$74.5_{0.4}$	$\mathbf{68.6_{0.5}}$	$\mathbf{84.2_{0.3}}$	$48.6_{1.6}$	$62.3_{1.9}$
O + F (34.4k)	$\mathbf{39.9_{0.8}}$	$\mathbf{61.7_{0.5}}$	$50.6_{0.9}$	$68.5_{0.9}$	$52.6_{1.4}$	$71.8_{1.1}$	$72.2_{0.4}$	$79.8_{0.6}$
O + R (34.4k)	$38.1_{0.5}$	$58.8_{0.6}$	$57.9_{1.0}$	$74.8_{0.5}$	$62.6_{0.5}$	$80.2_{0.3}$	$72.5_{0.5}$	$80.2_{0.3}$
O + SDC (34.4k)	$33.4_{0.4}$	$54.5_{0.6}$	$60.6_{4.4}$	$77.2_{3.6}$	$\mathbf{69.0_{0.3}}$	$\mathbf{84.3_{0.3}}$	$72.1_{0.2}$	$79.8_{0.2}$
Finetuned model: $\text{RoBERTa}_{\text{large}}$								
Original (O; 23.1k)	7.3	16.7	28.6	44.5	32.7	50.1	73.5	80.5
Original (11.3k)	$4.5_{0.4}$	$10.8_{1.1}$	$17.5_{0.9}$	$26.7_{2.0}$	$19.5_{2.1}$	$30.0_{3.2}$	$70.6_{0.3}$	$78.5_{0.4}$
$\text{BERT}_{\text{fooled}}$ (F; 11.3k)	$\mathbf{49.2_{0.5}}$	$\mathbf{71.2_{0.7}}$	$64.9_{1.3}$	$81.3_{1.1}$	$67.9_{1.5}$	$84.8_{1.0}$	$41.4_{1.0}$	$55.1_{1.1}$
$\text{BERT}_{\text{random}}$ (R; 11.3k)	$48.0_{0.4}$	$69.8_{0.4}$	$\mathbf{70.3_{0.7}}$	$\mathbf{85.3_{0.4}}$	$72.5_{0.4}$	$87.8_{0.1}$	$50.6_{0.8}$	$\mathbf{64.9_{1.0}}$
SDC (11.3k)	$42.9_{0.9}$	$65.3_{0.8}$	$67.0_{0.6}$	$83.6_{0.5}$	$\mathbf{74.4_{0.5}}$	$\mathbf{88.9_{0.3}}$	$\mathbf{51.0_{0.5}}$	$62.8_{0.6}$
O + F (34.4k)	$\mathbf{49.5_{0.5}}$	$\mathbf{71.1_{0.6}}$	$61.6_{0.8}$	$79.5_{0.6}$	$58.3_{2.0}$	$78.5_{1.2}$	$72.6_{0.4}$	$80.0_{0.4}$
O + R (34.4k)	$47.6_{0.7}$	$69.5_{0.5}$	$\mathbf{69.2_{0.5}}$	$\mathbf{84.6_{0.5}}$	$71.1_{0.7}$	$86.8_{0.3}$	$72.8_{0.6}$	$80.3_{0.5}$
O + SDC (34.4k)	$41.5_{0.4}$	$64.2_{0.4}$	$67.3_{0.6}$	$84.3_{0.4}$	$\mathbf{75.0_{0.6}}$	$\mathbf{88.9_{0.2}}$	$73.0_{0.2}$	$80.4_{0.1}$
Finetuned model: $\text{ELECTRA}_{\text{large}}$								
Original (O; 23.1k)	7.5	17.1	29.6	45.2	32.5	49.1	74.2	81.2
Original (11.3k)	$8.4_{0.9}$	$18.7_{0.6}$	$28.8_{0.5}$	$42.7_{0.9}$	$33.1_{0.7}$	$48.6_{1.1}$	$71.8_{0.1}$	$79.6_{0.1}$
$\text{BERT}_{\text{fooled}}$ (F; 11.3k)	$40.2_{4.6}$	$63.4_{3.2}$	$50.7_{4.7}$	$68.5_{4.8}$	$56.1_{4.4}$	$75.6_{3.0}$	$41.0_{4.8}$	$56.6_{4.2}$
$\text{BERT}_{\text{random}}$ (R; 11.3k)	$42.1_{2.7}$	$63.5_{2.1}$	$58.8_{2.2}$	$76.0_{1.5}$	$65.8_{1.9}$	$81.7_{1.3}$	$52.6_{1.9}$	$67.5_{1.4}$
SDC (11.3k)	$39.2_{0.3}$	$40.3_{0.4}$	$59.6_{0.7}$	$76.1_{0.6}$	$\mathbf{69.3_{0.7}}$	$\mathbf{84.2_{0.5}}$	$\mathbf{55.7_{0.7}}$	$\mathbf{69.5_{0.5}}$
O + F (34.4k)	$40.9_{3.4}$	$63.7_{2.3}$	$52.6_{2.5}$	$70.8_{2.1}$	$55.4_{4.5}$	$74.4_{4.1}$	$72.7_{1.2}$	$80.5_{1.0}$
O + R (34.4k)	$41.5_{5.6}$	$61.9_{5.7}$	$58.6_{4.6}$	$75.0_{4.4}$	$64.4_{4.1}$	$80.4_{3.3}$	$72.6_{2.0}$	$80.3_{2.1}$
O + SDC (34.4k)	$38.0_{0.6}$	$58.7_{0.6}$	$59.4_{0.6}$	$76.1_{0.4}$	$\mathbf{70.9_{0.4}}$	$\mathbf{85.1_{0.3}}$	$73.6_{0.7}$	$81.2_{0.4}$

Table 5.2: EM and F1 scores of various models evaluated on adversarial and non-adversarial datasets. Adversarial results in bold are statistically significant compared to SDC setting and vice versa with $p < 0.05$.

Models For our empirical analysis, we fine-tune BERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019], and ELECTRA [Clark et al., 2020a] models on all six datasets generated as part of our study (four datasets via ADC: $\text{BERT}_{\text{fooled}}$, $\text{BERT}_{\text{random}}$, $\text{ELECTRA}_{\text{fooled}}$, $\text{ELECTRA}_{\text{random}}$, and the two datasets via SDC). We also fine-tune these models after augmenting the original data to collected datasets. We report the means and standard deviations (in subscript) of EM and F1 scores following 10 runs of each experiment. Models fine-tuned on all ADC datasets typically

Evaluation set →	ELECTRA$_{\text{fooled}}$		ELECTRA$_{\text{random}}$		SDC		Original Dev.	
Training set ↓	EM	F1	EM	F1	EM	F1	EM	F1
Finetuned model: BERT$_{\text{large}}$								
Original (O; 23.1k)	23.3	31.9	56.7	72.6	63.8	78.5	73.3	80.5
Original (14.6k)	$36.7_{0.4}$	$50.7_{0.3}$	$48.2_{0.4}$	$64.4_{0.2}$	$55.7_{0.1}$	$70.5_{0.3}$	$67.1_{0.2}$	$75.2_{0.1}$
ELECTRA$_{\text{fooled}}$ (F; 14.6k)	$\mathbf{25.1_{1.0}}$	$\mathbf{42.4_{1.0}}$	$35.4_{1.5}$	$54.3_{1.1}$	$39.1_{2.4}$	$59.3_{1.7}$	$31.9_{7.9}$	$45.0_{9.2}$
ELECTRA$_{\text{random}}$ (R; 14.6k)	$25.4_{1.1}$	$42.0_{1.0}$	$\mathbf{38.4_{0.9}}$	$56.8_{0.8}$	$42.0_{1.4}$	$61.7_{1.3}$	$46.4_{3.1}$	$60.6_{3.8}$
SDC (14.6k)	$23.1_{1.0}$	$40.8_{1.3}$	$36.3_{1.3}$	$56.3_{1.3}$	$\mathbf{45.2_{1.8}}$	$\mathbf{65.4_{1.5}}$	$48.6_{1.6}$	$62.3_{1.9}$
O + F (37.7k)	$\mathbf{26.7_{1.7}}$	$\mathbf{43.1_{0.9}}$	$40.1_{1.3}$	$58.7_{1.5}$	$44.6_{0.9}$	$64.2_{1.2}$	$72.1_{0.5}$	$79.7_{0.7}$
O + R (37.7k)	$26.0_{0.8}$	$42.9_{0.6}$	$41.7_{0.5}$	$60.3_{0.6}$	$47.1_{1.4}$	$66.5_{1.3}$	$\mathbf{73.0_{0.5}}$	$\mathbf{80.5_{0.2}}$
O + SDC (37.7k)	$24.5_{0.7}$	$41.7_{0.7}$	$41.4_{0.9}$	$60.7_{0.4}$	$\mathbf{50.9_{1.0}}$	$\mathbf{69.7_{0.3}}$	$72.0_{0.1}$	$79.7_{0.1}$
Finetuned model: RoBERTa$_{\text{large}}$								
Original (O; 23.1k)	49.2	64.4	59.1	75.8	64.5	79.8	73.5	80.5
Original (14.6k)	$48.3_{0.9}$	$63.3_{1.4}$	$58.7_{0.9}$	$74.9_{1.0}$	$62.7_{0.4}$	$79.0_{0.7}$	$71.5_{0.5}$	$79.3_{0.6}$
ELECTRA$_{\text{fooled}}$ (F; 14.6k)	$\mathbf{65.3_{0.5}}$	$\mathbf{79.9_{0.5}}$	$69.4_{0.6}$	$84.6_{0.5}$	$75.8_{0.6}$	$89.0_{0.3}$	$55.9_{1.2}$	$67.5_{1.0}$
ELECTRA$_{\text{random}}$ (R; 14.6k)	$64.6_{0.5}$	$79.4_{0.4}$	$\mathbf{70.4_{0.5}}$	$\mathbf{85.4_{0.3}}$	$76.5_{0.5}$	$89.4_{0.3}$	$\mathbf{59.8_{1.2}}$	$\mathbf{70.6_{0.9}}$
SDC (14.6k)	$61.0_{0.2}$	$77.1_{0.3}$	$67.9_{0.4}$	$84.1_{0.4}$	$\mathbf{77.3_{0.5}}$	$\mathbf{89.9_{0.3}}$	$55.7_{1.0}$	$68.8_{0.8}$
O + F (37.7k)	$\mathbf{65.0_{0.3}}$	$\mathbf{79.9_{0.3}}$	$70.1_{0.5}$	$85.2_{0.4}$	$76.2_{0.3}$	$89.7_{0.2}$	$73.3_{0.3}$	$80.7_{0.2}$
O + R (37.7k)	$64.3_{0.3}$	$78.8_{0.3}$	$\mathbf{70.7_{0.2}}$	$\mathbf{85.8_{0.2}}$	$76.5_{0.6}$	$89.7_{0.3}$	$73.4_{0.5}$	$80.8_{0.3}$
O + SDC (37.7k)	$61.5_{0.5}$	$77.2_{0.3}$	$69.0_{0.4}$	$84.7_{0.4}$	$\mathbf{77.6_{0.4}}$	$\mathbf{90.5_{0.2}}$	$73.6_{0.5}$	$80.9_{0.4}$
Finetuned model: ELECTRA$_{\text{large}}$								
Original (O; 23.1k)	0	10.8	40.2	57.8	44.8	60.9	74.2	81.2
Original (14.6k)	$25.9_{0.2}$	$40.9_{0.4}$	$37.3_{0.6}$	$63.9_{0.7}$	$53.6_{1.3}$	$74.7_{1.1}$	$71.9_{0.3}$	$79.5_{0.3}$
ELECTRA$_{\text{fooled}}$ (F; 14.6k)	$26.4_{1.5}$	$44.0_{1.6}$	$41.2_{1.5}$	$60.8_{1.3}$	$42.7_{4.0}$	$63.5_{3.2}$	$57.5_{0.9}$	$68.8_{0.7}$
ELECTRA$_{\text{random}}$ (R; 14.6k)	$23.4_{4.9}$	$40.5_{5.6}$	$42.3_{6.9}$	$62.3_{7.0}$	$42.1_{8.0}$	$62.9_{7.5}$	$57.6_{0.8}$	$69.3_{1.0}$
SDC (14.6k)	$24.5_{2.4}$	$43.7_{3.5}$	$40.6_{3.5}$	$61.5_{3.8}$	$46.9_{5.4}$	$68.2_{4.7}$	$54.9_{1.8}$	$68.3_{1.2}$
O + F (37.7k)	$25.3_{1.9}$	$43.7_{2.0}$	$40.2_{1.9}$	$60.6_{1.9}$	$41.7_{3.9}$	$63.4_{3.6}$	$73.6_{0.5}$	$81.1_{0.4}$
O + R (37.7k)	$21.7_{1.1}$	$40.1_{1.1}$	$42.2_{2.3}$	$64.8_{1.9}$	$38.0_{3.6}$	$60.8_{2.9}$	$74.4_{0.3}$	$\mathbf{81.7_{0.1}}$
O + SDC (37.7k)	$24.5_{1.8}$	$43.4_{1.6}$	$42.8_{1.5}$	$63.5_{1.0}$	$\mathbf{49.6_{1.9}}$	$\mathbf{70.3_{1.5}}$	$74.2_{0.2}$	$81.5_{0.1}$

Table 5.3: EM and F1 scores of various models evaluated on adversarial datasets collected with an ELECTRA$_{\text{large}}$ model and non-adversarial datasets. Adversarial results in bold are statistically significant compared to SDC setting and vice versa with $p < 0.05$.

perform better on their held-out test sets than those trained on SDC data and vice-versa (Tables 5.2 and 5.3). RoBERTa fine-tuned on the BERT$_{\text{fooled}}$ training set obtains EM and F1 scores of 49.2 and 71.2, respectively, on the BERT$_{\text{fooled}}$ test set, outperforming RoBERTa models fine-tuned on BERT$_{\text{random}}$ (EM: 48.0, F1: 69.8) and SDC (EM: 42.0, F1: 65.3). Performance on the original dev set [Karpukhin et al., 2020] is generally comparable across all models.

Evaluation set →	D_{RoBERTa}		D_{BERT}		D_{BiDAF}	
Training set ↓	EM	F1	EM	F1	EM	F1
Finetuned model: $\text{BERT}_{\text{large}}$						
Original (23.1k)	6.0	13.5	8.1	14.2	12.6	21.4
Original (14.6k)	$5.3_{0.2}$	$11.4_{0.2}$	$6.8_{0.8}$	$13.9_{0.5}$	$12.1_{0.4}$	$20.6_{0.2}$
$\text{ELECTRA}_{\text{fooled}}$14.6k)	$3.8_{0.5}$	$13.3_{0.7}$	$6.2_{0.7}$	$16.4_{0.5}$	$12.6_{1.2}$	$26.2_{1.0}$
$\text{ELECTRA}_{\text{random}}$14.6k)	$\mathbf{4.3_{0.5}}$	$13.7_{0.7}$	$\mathbf{6.4_{0.4}}$	$\mathbf{16.4_{0.8}}$	$\mathbf{13.6_{0.8}}$	$\mathbf{27.1_{1.2}}$
SDC (14.6k)	$3.9_{0.4}$	$13.2_{0.4}$	$5.4_{0.4}$	$15.1_{0.5}$	$10.8_{0.7}$	$23.8_{0.8}$
Orig + $\text{ELECTRA}_{\text{fooled}}$ (37.7k)	$6.4_{0.5}$	$16.1_{0.3}$	$7.8_{0.8}$	$18.0_{0.6}$	$17.0_{0.2}$	$31.0_{0.6}$
Orig + $\text{ELECTRA}_{\text{random}}$ (37.7k)	$\mathbf{6.6_{0.6}}$	$\mathbf{16.1_{0.3}}$	$8.5_{0.6}$	$18.4_{0.5}$	$16.9_{0.3}$	$30.8_{0.4}$
Orig + SDC (37.7k)	$5.8_{0.2}$	$15.6_{0.4}$	$8.7_{0.5}$	$18.7_{0.6}$	$17.4_{0.7}$	$30.0_{0.8}$
Finetuned model: $\text{RoBERTa}_{\text{large}}$						
Original (23.1k)	15.7	25.0	26.5	37.0	37.9	50.4
Original (14.6k)	$14.3_{0.2}$	$23.7_{0.3}$	$25.1_{0.3}$	$35.4_{0.7}$	$37.4_{0.7}$	$50.2_{0.5}$
$\text{ELECTRA}_{\text{fooled}}$14.6k)	$\mathbf{16.4_{0.9}}$	$\mathbf{27.7_{1.2}}$	$27.4_{1.3}$	$40.8_{1.5}$	$46.8_{1.1}$	$62.4_{1.1}$
$\text{ELECTRA}_{\text{random}}$14.6k)	$15.8_{1.4}$	$27.2_{1.4}$	$\mathbf{28.1_{1.6}}$	$\mathbf{41.5_{1.8}}$	$\mathbf{48.0_{0.9}}$	$\mathbf{63.0_{0.6}}$
SDC (14.6k)	$12.1_{1.0}$	$23.9_{1.3}$	$22.7_{1.1}$	$35.4_{1.5}$	$40.5_{1.3}$	$56.8_{1.3}$
Orig + $\text{ELECTRA}_{\text{fooled}}$ (37.7k)	$18.9_{0.8}$	$30.4_{0.9}$	$\mathbf{33.2_{0.8}}$	$\mathbf{46.4_{0.6}}$	$\mathbf{49.2_{0.9}}$	$\mathbf{65.1_{0.8}}$
Orig + $\text{ELECTRA}_{\text{random}}$ (37.7k)	$18.0_{0.4}$	$29.6_{0.3}$	$32.3_{0.6}$	$45.1_{1.2}$	$48.2_{0.8}$	$63.5_{0.6}$
Orig + SDC (37.7k)	$18.2_{1.0}$	$29.7_{0.9}$	$28.2_{0.3}$	$41.4_{0.5}$	$45.0_{0.9}$	$60.9_{0.6}$
Finetuned model: $\text{ELECTRA}_{\text{large}}$						
Original (23.1k)	8.2	17.4	15.7	24.2	22.4	34.3
Original (14.6k)	$9.5_{0.2}$	$18.0_{0.5}$	$15.4_{0.5}$	$24.2_{0.6}$	$21.7_{0.2}$	$33.1_{0.1}$
$\text{ELECTRA}_{\text{fooled}}$14.6k)	$10.2_{0.3}$	$21.7_{0.5}$	$17.0_{0.7}$	$29.7_{0.6}$	$21.7_{1.7}$	$36.6_{1.1}$
$\text{ELECTRA}_{\text{random}}$14.6k)	$10.4_{0.5}$	$21.3_{0.5}$	$16.5_{0.2}$	$28.6_{0.8}$	$19.9_{5.0}$	$34.4_{5.9}$
SDC (14.6k)	$10.3_{0.8}$	$21.6_{0.7}$	$15.8_{1.1}$	$28.5_{1.2}$	$19.3_{4.8}$	$33.3_{7.8}$
Orig + $\text{ELECTRA}_{\text{fooled}}$ (37.7k)	$10.2_{0.3}$	$21.7_{0.5}$	$17.0_{0.7}$	$29.7_{0.6}$	$24.0_{0.7}$	$39.2_{0.7}$
Orig + $\text{ELECTRA}_{\text{random}}$ (37.7k)	$10.4_{0.5}$	$21.3_{0.5}$	$16.5_{0.2}$	$28.6_{0.8}$	$23.5_{0.5}$	$38.4_{0.4}$
Orig + SDC (37.7k)	$10.3_{0.8}$	$21.6_{0.7}$	$15.8_{1.1}$	$28.5_{1.2}$	$\mathbf{24.5_{0.6}}$	$\mathbf{39.9_{0.6}}$

Table 5.4: EM and F1 scores of various models evaluated on dev datasets of Bartolo et al. [2020]. Adversarial results in bold are statistically significant compared to SDC setting and vice versa with $p < 0.05$.

Out-of-domain generalization to adversarial data We evaluate these models on adversarial test sets constructed with BiDAF (D_{BiDAF}), BERT (D_{BERT}) and RoBERTa (D_{RoBERTa}) in the loop [Bartolo et al., 2020]. Prior work suggests that training on ADC data leads to models that perform better on similarly constructed adversarial evaluation sets. Both BERT and RoBERTa models fine-tuned on adversarial data generally outperform models fine-tuned on SDC data (or when either datasets are augmented to the original data) on all three evaluation sets (Tables 5.5

Evaluation set →	D_{RoBERTa}		D_{BERT}		D_{BiDAF}	
Training set ↓	EM	F1	EM	F1	EM	F1
Finetuned model: $\text{BERT}_{\text{large}}$						
Original (23.1k)	6.0	13.5	8.1	14.2	12.6	21.4
Original (11.3k)	$5.4_{0.3}$	$12.2_{0.1}$	$7.0_{0.6}$	$13.6_{0.8}$	$11.0_{0.9}$	$19.4_{0.7}$
$\text{BERT}_{\text{fooled}}$ (11.3k)	$11.0_{2.6}$	$21.0_{3.0}$	$14.6_{3.7}$	$24.7_{4.0}$	$25.1_{6.5}$	$39.1_{6.9}$
$\text{BERT}_{\text{random}}$ (11.3k)	$\mathbf{12.4_{1.6}}$	$22.1_{2.2}$	$16.4_{3.0}$	$26.2_{2.7}$	$29.6_{3.7}$	$43.7_{4.0}$
SDC (11.3k)	$9.1_{0.7}$	$20.4_{0.7}$	$14.0_{1.0}$	$24.6_{0.7}$	$30.1_{1.2}$	$43.8_{1.2}$
Orig + $\text{BERT}_{\text{fooled}}$ (34.4k)	$15.2_{0.8}$	$25.1_{0.6}$	$20.4_{0.4}$	$31.0_{0.4}$	$32.4_{0.6}$	$47.0_{0.6}$
Orig + $\text{BERT}_{\text{random}}$ (34.4k)	$\mathbf{16.9_{0.5}}$	$\mathbf{23.9_{0.5}}$	$\mathbf{20.5_{0.6}}$	$\mathbf{31.2_{0.9}}$	$\mathbf{34.1_{0.4}}$	$47.8_{0.7}$
Orig + SDC (34.4k)	$9.4_{0.6}$	$20.2_{0.5}$	$15.3_{1.0}$	$25.8_{1.1}$	$32.7_{1.2}$	$47.2_{1.0}$
Finetuned model: $\text{RoBERTa}_{\text{large}}$						
Original (23.1k)	15.7	25.0	26.5	37.0	37.9	50.4
Original (11.3k)	$14.6_{0.3}$	$23.8_{0.5}$	$22.5_{1.2}$	$32.6_{1.5}$	$36.0_{1.1}$	$48.9_{1.2}$
$\text{BERT}_{\text{fooled}}$ (11.3k)	$\mathbf{21.9_{1.6}}$	$\mathbf{32.2_{1.6}}$	$30.2_{1.6}$	$42.5_{1.6}$	$46.3_{1.6}$	$61.9_{1.5}$
$\text{BERT}_{\text{random}}$ (11.3k)	$21.3_{1.3}$	$31.6_{1.5}$	$\mathbf{31.3_{2.2}}$	$\mathbf{43.6_{2.3}}$	$\mathbf{48.0_{1.4}}$	$\mathbf{63.4_{1.3}}$
SDC (11.3k)	$12.8_{1.2}$	$23.4_{1.3}$	$20.0_{1.8}$	$32.1_{2.2}$	$40.0_{2.0}$	$55.0_{1.8}$
Orig + $\text{BERT}_{\text{fooled}}$ (34.4k)	$\mathbf{25.2_{0.9}}$	$\mathbf{36.4_{1.0}}$	$\mathbf{35.9_{0.9}}$	$\mathbf{48.5_{0.8}}$	$49.6_{0.7}$	$65.1_{1.1}$
Orig + $\text{BERT}_{\text{random}}$ (34.4k)	$24.6_{1.5}$	$35.2_{1.5}$	$35.7_{1.0}$	$48.0_{1.2}$	$\mathbf{50.6_{1.5}}$	$\mathbf{65.8_{1.2}}$
Orig + SDC (34.4k)	$16.1_{0.8}$	$27.6_{1.1}$	$26.6_{0.8}$	$39.7_{0.6}$	$43.4_{0.4}$	$59.4_{0.3}$
Finetuned model: $\text{ELECTRA}_{\text{large}}$						
Original (23.1k)	8.2	17.4	15.7	24.2	22.4	34.3
Original (11.3k)	$8.5_{0.4}$	$16.7_{0.5}$	$14.3_{1.0}$	$23.0_{0.9}$	$20.7_{1.4}$	$32.0_{1.3}$
$\text{BERT}_{\text{fooled}}$ (11.3k)	$13.8_{3.7}$	$24.3_{5.6}$	$18.8_{6.0}$	$31.1_{8.1}$	$29.1_{9.0}$	$44.3_{11.0}$
$\text{BERT}_{\text{random}}$ (11.3k)	$\mathbf{14.8_{1.8}}$	$\mathbf{25.9_{1.1}}$	$\mathbf{22.3_{2.9}}$	$\mathbf{34.6_{2.5}}$	$34.8_{3.4}$	$50.5_{2.7}$
SDC (11.3k)	$11.6_{0.6}$	$22.7_{0.7}$	$17.8_{1.2}$	$30.4_{1.3}$	$32.5_{1.8}$	$49.3_{1.6}$
Orig + $\text{BERT}_{\text{fooled}}$ (34.4k)	$16.5_{3.8}$	$28.0_{3.8}$	$23.1_{3.9}$	$35.6_{4.2}$	$34.8_{5.1}$	$50.2_{5.7}$
Orig + $\text{BERT}_{\text{random}}$ (34.4k)	$18.4_{4.2}$	$28.9_{5.0}$	$25.9_{5.9}$	$37.2_{6.9}$	$37.2_{7.5}$	$51.1_{9.1}$
Orig + SDC (34.4k)	$15.6_{1.1}$	$27.0_{1.1}$	$22.7_{0.6}$	$36.0_{0.8}$	$34.5_{0.9}$	$49.5_{1.2}$

Table 5.5: EM and F1 scores of various models evaluated on dev datasets of Bartolo et al. [2020]. Adversarial results in bold are statistically significant compared to SDC setting and vice versa with $p < 0.05$.

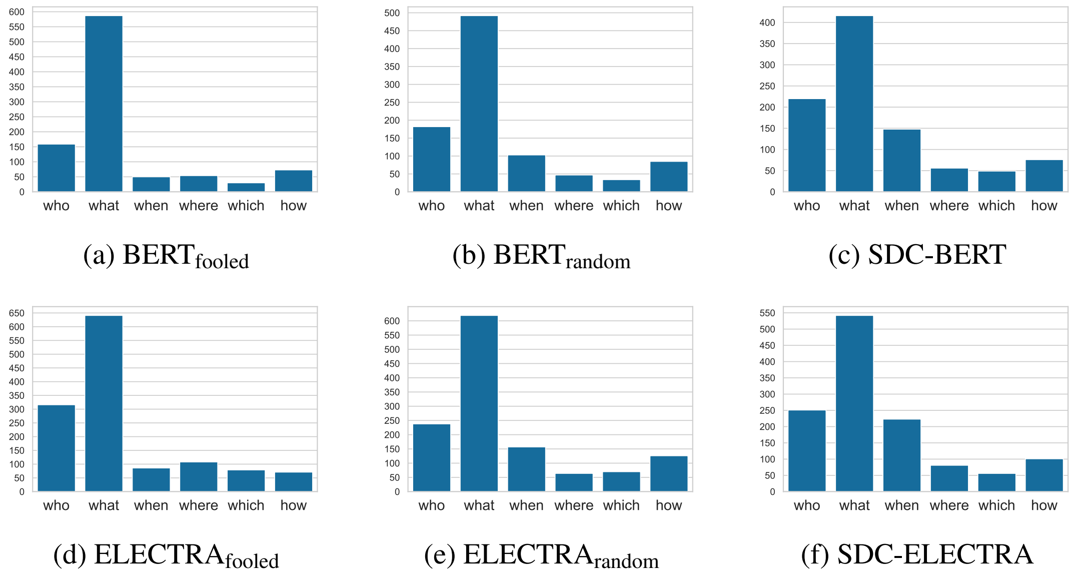
and 5.4). A RoBERTa model fine-tuned on $\text{BERT}_{\text{fooled}}$ outperforms a RoBERTa model fine-tuned on SDC by 9.1, 9.3, and 6.2 EM points on D_{RoBERTa}, D_{BERT}, and D_{BiDAF}, respectively. We observe similar trends on ELECTRA models fine-tuned on ADC data versus SDC data, but these gains disappear when the same models are finetuned on augmented data. For instance, while ELECTRA fine-tuned on $\text{BERT}_{\text{random}}$ obtains an EM score of 14.8 on D_{RoBERTa}, outperforming an ELECTRA fine-tuned on SDC data by ≈ 3 pts, the difference is no longer significant when respective models are fine-tuned after original data is augmented to these datasets. ELECTRA models fine-tuned on ADC data with ELECTRA in the loop perform no better than those trained on SDC. Fine-tuning ELECTRA on SDC augmented to original data leads to an ≈ 1 pt improvement on both metrics compared to augmenting ADC. Overall, we find that models fine-tuned on ADC data typically generalize better to out-of-domain adversarial test sets than models fine-tuned on SDC data, confirming the findings by Dinan et al. [2019].

Out-of-domain generalization to MRQA We further evaluate these models on 12 out-of-domain datasets used in the 2019 MRQA shared task[3] (Tables 5.6 and 5.7).[4] Notably, for BERT, fine-tuning on SDC data leads to significantly better performance (as compared to fine-tuning on ADC data collected with BERT) on 9 out of 12 MRQA datasets, with gains of more than 10 EM pts on 6 of them. On BioASQ, BERT fine-tuned on BERT$_{\text{fooled}}$ obtains EM and F1 scores of 23.5 and 30.3, respectively. By comparison, fine-tuning on SDC data yields markedly higher EM and F1 scores of 35.1 and 55.7, respectively. Similar trends hold across models and datasets. Interestingly, ADC fine-tuning often improves performance on DROP compared to SDC. For instance, RoBERTa fine-tuned on ELECTRA$_{\text{random}}$ outperforms RoBERTa fine-tuned on SDC by ≈ 7 pts. Note that DROP itself was adversarially constructed. On Natural Questions, models fine-tuned on ADC data generally perform comparably to those fine-tuned on SDC data. RoBERTa fine-tuned on BERT$_{\text{random}}$ obtains EM and F1 scores of 48.1 and 62.6, respectively, whereas RoBERTa fine-tuned on SDC data obtains scores of 47.9 and 61.7, respectively. It is worth noting that passages sourced to construct both ADC and SDC datasets come from the Natural Questions dataset, which could be one reason why models fine-tuned on ADC datasets perform similar to those fine-tuned on SDC datasets when evaluated on Natural Questions.

On the the adversarial process versus adversarial success We notice that models fine-tuned on BERT$_{\text{random}}$ and ELECTRA$_{\text{random}}$ typically outperform models fine-tuned on BERT$_{\text{fooled}}$ and ELECTRA$_{\text{fooled}}$, respectively, on adversarial test data collected in prior work [Bartolo et al., 2020], as well as on MRQA. Similar observation can be made when the ADC data is augmented with the original training data. These trends suggest that the ADC process (regardless of the outcome) explains our results more than successfully fooling a model. Furthermore, models fine-tuned only on SDC data tend to outperform ADC-only fine-tuned models; however, following augmentation, ADC fine-tuning achieves comparable performance on more datasets than before, showcasing generalization following augmentation. Notice that augmenting ADC data to original data may not always help. BERT fine-tuned on original $23.1k$ examples achieves an EM 11.3 on SearchQA. When fine-tuned on BERT$_{\text{fooled}}$ augmented to the original data, this

[3] The MRQA 2019 shared task includes HotpotQA [Yang et al., 2018a], Natural Questions [Kwiatkowski et al., 2019], SearchQA [Dunn et al., 2017], SQuAD [Rajpurkar et al., 2016], TriviaQA [Joshi et al., 2017], BioASQ [Tsatsaronis et al., 2015], DROP [Dua et al., 2019], DuoRC [Saha et al., 2018], RelationExtraction [Levy et al., 2017], RACE [Lai et al., 2017], and TextbookQA [Kembhavi et al., 2017].

[4] Interestingly, RoBERTa appears to perform better compared to BERT and ELECTRA. Prior works have hypothesized that the bigger size and increased diversity of the pre-training corpus of RoBERTa (compared to those of BERT and ELECTRA) might somehow be responsible for RoBERTa's better out-of-domain generalization, [Baevski et al., 2019, Hendrycks et al., 2020, Tu et al., 2020].



(a) $\text{BERT}_{\text{fooled}}$

(b) $\text{BERT}_{\text{random}}$

(c) SDC-BERT

(d) $\text{ELECTRA}_{\text{fooled}}$

(e) $\text{ELECTRA}_{\text{random}}$

(f) SDC-ELECTRA

Figure 5.2: Frequency of wh-questions generated.

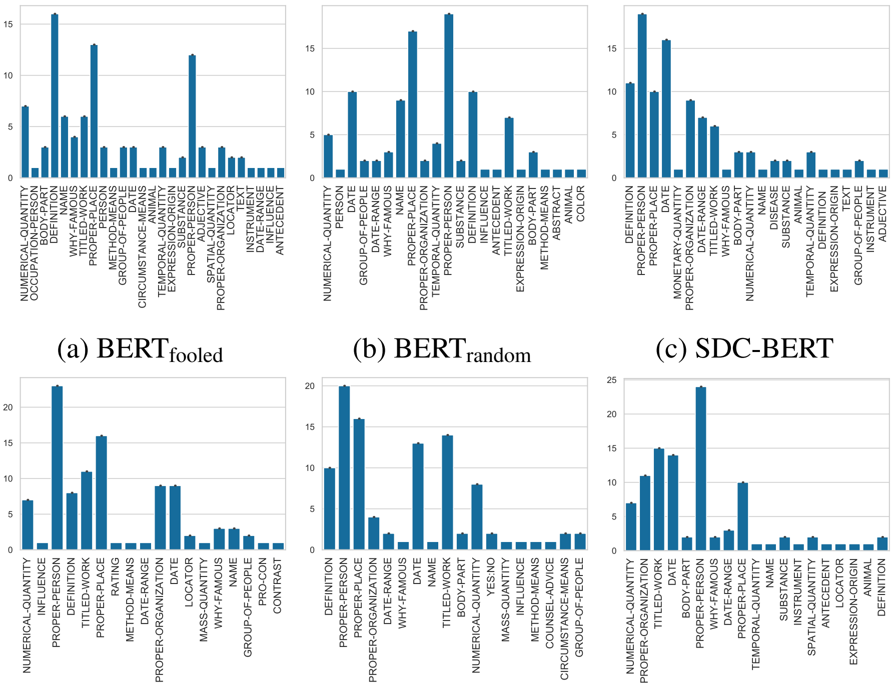
drops to 8.7, and when fine-tuned on $\text{BERT}_{\text{random}}$ augmented to the original data, it drops to 11.2. Fine-tuning on SDC augmented to the original data, however, results in EM of 13.6.

5.5 Qualitative Analysis

Finally, we perform a qualitative analysis over the collected data, revealing profound differences with models in (versus out of) the loop. Recall that because these datasets were constructed in a randomized study, any observed differences are attributable to the model-in-the loop collection scheme.

To begin, we analyze 100 questions from each dataset and categorize them using the taxonomy introduced by Hovy et al. [2000]. We also look at the first word of the *wh*-type questions in each dev set (Fig. 5.3) and observe key qualitative differences between data via ADC and SDC for both models.

In case of ADC with BERT (and associated SDC), while we observe that most questions in the dev sets start with *what*, ADC has a higher proportion compared to SDC (587 in $\text{BERT}_{\text{fooled}}$ and 492 in $\text{BERT}_{\text{random}}$ versus 416 in SDC). Furthermore, we notice that compared to $\text{BERT}_{\text{fooled}}$ dev set, SDC has more *when*- (148) and *who*-type (220) questions, the answers to which typically refer to dates, places and people (or organizations), respectively. This is also reflected in the taxonomy categorization. Interestingly, the $\text{BERT}_{\text{random}}$ dev set has more *when*- and *who*-type questions than $\text{BERT}_{\text{fooled}}$ (103 and 182 versus 50 and 159, respectively). This indicates that the BERT model could have been better at answering questions related to dates and people (or organizations), which could have further incentivized workers not to generate such questions upon observing these patterns. Similarly, in the 100-question samples, we find that a larger proportion of questions in ADC are categorized as requiring numerical reasoning (11 and 18 in



(a) BERT$_{\text{fooled}}$ (b) BERT$_{\text{random}}$ (c) SDC-BERT

(d) ELECTRA$_{\text{fooled}}$ (e) ELECTRA$_{\text{random}}$ (f) SDC-ELECTRA

Figure 5.3: Frequency of question types based on the taxonomy introduced by Hovy et al. [2000].

BERT$_{\text{fooled}}$ and BERT$_{\text{random}}$, respectively) compared to SDC (7). It is possible that the model's performance on numerical reasoning (as also demonstrated by its lower performance on DROP compared to fine-tuning on ADC or SDC) would have incentivized workers to generate more questions requiring numerical reasoning and as a result, skewed the distribution towards such questions.

Similarly, with ELECTRA, we observe that *what*-type questions constitute most of the questions in the development sets for both ADC and SDC, although data collected via ADC has a higher proportion of these (641 in ELECTRA$_{\text{fooled}}$ and 619 in ELECTRA$_{\text{random}}$ versus 542 in SDC). We also notice more *how*-type questions in ADC (126 in ELECTRA$_{\text{random}}$) vs 101 in SDC, and that the SDC sample has more questions that relate to dates (223) but the number is lower in the ADC samples (157 and 86 in ELECTRA$_{\text{random}}$ and ELECTRA$_{\text{fooled}}$, respectively). As with BERT, the ELECTRA model was likely better at identifying answers about dates or years which could have further incentivized workers to generate less questions of such types. However, unlike with BERT, we observe that the ELECTRA ADC and SDC 100-question samples contain similar numbers of questions involving numerical answers (8, 9 and 10 in ELECTRA$_{\text{fooled}}$, ELECTRA$_{\text{random}}$ and SDC respectively).

Lastly, despite explicit instructions not to generate questions about passage structure (Fig. 5.1), a small number of workers nevertheless created such questions. For instance, one worker wrote, "*What is the number in the passage that is one digit less than the largest number in the passage?*" While most such questions were discarded during validation, some of these are present in the final data. Overall, we notice considerable differences between ADC and SDC data, particularly vis-a-vis what kind of questions workers generate. Our qualitative analysis offers additional insights that suggest that ADC would skew the distribution of questions workers create, as the incentives align with quickly creating more questions that can fool the model. This is reflected in all our ADC datasets. One remedy could be to provide workers with initial questions, asking them to edit those questions to elicit incorrect model predictions. Similar strategies were employed in Ettinger et al. [2017], where *breakers* minimally edited original data to elicit incorrect predictions from the models built by *builders*, as well as in recently introduced adversarial benchmarks for sentiment analysis [Potts et al., 2020].

Finetuned model: BERT$_{\text{large}}$

Evaluation set →	BioASQ		DROP		DuoRC		Relation Extraction		RACE		TextbookQA	
Training set ↓	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Original (23.1k)	19.4	32.5	7.8	16.2	14.5	22.8	32.0	47.1	11.4	18.8	25.0	33.4
Original (11.3k)	$20.8_{1.7}$	$36.0_{3.4}$	$6.2_{1.4}$	$12.7_{1.8}$	$13.1_{1.1}$	$19.8_{1.6}$	$42.4_{0.4}$	$55.9_{0.1}$	$10.3_{0.6}$	$18.3_{0.4}$	$20.0_{0.9}$	$27.9_{0.7}$
BERT$_{\text{fooled}}$ (11.3k)	$23.5_{6.0}$	$30.3_{3.5}$	$11.5_{3.2}$	$22.2_{3.4}$	$20.3_{4.5}$	$28.2_{5.0}$	$51.5_{8.2}$	$68.9_{6.6}$	$15.1_{3.1}$	$26.1_{4.3}$	$16.7_{3.8}$	$24.7_{4.6}$
BERT$_{\text{random}}$ (11.3k)	$30.3_{3.5}$	$46.8_{2.8}$	$14.4_{2.0}$	$25.1_{2.5}$	$26.7_{3.3}$	$35.3_{3.0}$	$61.3_{5.8}$	$75.9_{4.5}$	$18.4_{1.8}$	$29.9_{2.0}$	$21.9_{3.1}$	$30.9_{3.8}$
SDC (11.3k)	$\mathbf{35.1_{2.1}}$	$\mathbf{55.7_{1.1}}$	$14.6_{0.4}$	$24.7_{0.6}$	$\mathbf{31.7_{0.7}}$	$\mathbf{41.2_{0.7}}$	$63.2_{1.2}$	$77.7_{0.7}$	$\mathbf{19.7_{0.6}}$	$31.0_{0.6}$	$\mathbf{26.0_{4.3}}$	$\mathbf{35.5_{4.7}}$
Orig + Fooled (34.4k)	$31.7_{1.2}$	$48.2_{1.2}$	$19.9_{0.9}$	$31.0_{0.8}$	$24.4_{0.9}$	$33.1_{1.4}$	$55.0_{1.7}$	$71.5_{1.2}$	$19.2_{1.3}$	$31.0_{1.1}$	$22.2_{4.7}$	$30.9_{5.4}$
Orig + Random (34.4k)	$34.9_{1.2}$	$51.8_{0.9}$	$\mathbf{21.4_{0.6}}$	$\mathbf{33.1_{0.4}}$	$27.1_{1.2}$	$36.1_{1.2}$	$62.3_{0.9}$	$77.1_{0.7}$	$21.0_{1.4}$	$33.0_{1.3}$	$27.7_{3.9}$	$37.1_{4.0}$
Orig + SDC (34.4k)	$\mathbf{38.8_{1.5}}$	$\mathbf{56.0_{1.3}}$	$19.4_{0.9}$	$31.1_{1.0}$	$\mathbf{31.9_{0.4}}$	$\mathbf{41.6_{0.6}}$	$62.4_{0.7}$	$77.8_{0.2}$	$20.7_{1.4}$	$32.7_{1.2}$	$29.0_{2.4}$	$38.8_{3.1}$

	HotpotQA		Natural Questions		NewsQA		SearchQA		SQuAD		TriviaQA	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Original (23.1k)	19.4	33.9	36.3	48.7	16.2	25.6	11.3	19.3	32.5	46.0	16.8	25.3
Original (11.3k)	$20.1_{0.3}$	$32.6_{0.6}$	$38.4_{0.5}$	$50.6_{0.6}$	$15.0_{1.0}$	$24.9_{1.7}$	$11.1_{0.7}$	$18.6_{1.2}$	$29.6_{0.4}$	$43.0_{0.7}$	$15.3_{1.0}$	$23.9_{1.4}$
BERT$_{\text{fooled}}$ (11.3k)	$27.2_{6.4}$	$43.2_{7.5}$	$28.0_{5.7}$	$42.8_{6.5}$	$22.7_{4.7}$	$37.5_{6.4}$	$6.1_{1.7}$	$11.8_{2.2}$	$42.6_{7.6}$	$60.6_{7.9}$	$16.1_{4.6}$	$24.3_{5.4}$
BERT$_{\text{random}}$ (11.3k)	$37.5_{3.1}$	$54.4_{3.1}$	$36.7_{3.9}$	$51.2_{3.5}$	$29.6_{1.9}$	$44.9_{1.9}$	$8.6_{1.4}$	$14.6_{1.8}$	$51.9_{2.6}$	$69.3_{2.1}$	$24.7_{2.8}$	$34.4_{3.0}$
SDC (11.3k)	$\mathbf{41.2_{0.9}}$	$\mathbf{57.9_{1.0}}$	$39.3_{1.2}$	$53.6_{1.1}$	$\mathbf{32.0_{0.8}}$	$\mathbf{48.0_{1.1}}$	$\mathbf{10.6_{1.4}}$	$\mathbf{18.0_{1.3}}$	$\mathbf{56.4_{0.4}}$	$\mathbf{72.5_{0.4}}$	$\mathbf{28.6_{0.8}}$	$\mathbf{39.9_{0.9}}$
Orig + Fooled (34.4k)	$34.4_{1.0}$	$51.1_{0.8}$	$39.9_{1.3}$	$54.1_{0.8}$	$26.3_{0.9}$	$42.8_{1.1}$	$8.7_{1.5}$	$14.5_{1.7}$	$47.6_{0.5}$	$66.3_{0.5}$	$21.9_{0.7}$	$30.9_{0.8}$
Orig + Random (34.4k)	$41.0_{0.7}$	$57.3_{0.7}$	$44.5_{0.4}$	$58.2_{0.2}$	$30.0_{0.5}$	$45.9_{0.6}$	$11.2_{0.7}$	$17.7_{0.9}$	$53.4_{0.4}$	$70.8_{0.4}$	$28.6_{1.3}$	$38.6_{1.4}$
Orig + SDC (34.4k)	$\mathbf{43.3_{0.2}}$	$\mathbf{60.0_{0.3}}$	$45.6_{0.9}$	$58.7_{1.1}$	$\mathbf{32.0_{0.8}}$	$\mathbf{48.6_{1.1}}$	$\mathbf{13.6_{0.4}}$	$\mathbf{22.2_{0.5}}$	$\mathbf{57.0_{0.3}}$	$\mathbf{73.2_{0.3}}$	$\mathbf{30.9_{1.0}}$	$\mathbf{42.4_{0.9}}$

Finetuned model: RoBERTa$_{\text{large}}$

Evaluation set →	BioASQ		DROP		DuoRC		Relation Extraction		RACE		TextbookQA	
Training set ↓	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Original (23.1k)	47.7	63.5	37.2	48.1	38.6	49.1	74.4	85.9	33.7	44.9	36.4	46
Original (11.3k)	$46.3_{0.1}$	$62.7_{1.0}$	$34.7_{0.3}$	$46.5_{0.8}$	$36.6_{1.8}$	$46.9_{2.1}$	$72.3_{0.8}$	$84.5_{0.3}$	$30.7_{0.2}$	$42.2_{0.3}$	$34.9_{0.4}$	$44.4_{0.2}$
BERT$_{\text{fooled}}$ (11.3k)	$35.6_{1.3}$	$51.0_{1.2}$	$34.1_{2.5}$	$46.8_{2.4}$	$31.4_{2.5}$	$39.7_{3.0}$	$67.0_{1.0}$	$81.9_{0.5}$	$28.2_{1.3}$	$41.4_{1.1}$	$25.4_{2.4}$	$35.1_{2.4}$
BERT$_{\text{random}}$ (11.3k)	$40.4_{1.2}$	$57.4_{1.2}$	$\mathbf{38.1_{2.2}}$	$\mathbf{51.2_{2.0}}$	$36.7_{1.6}$	$45.5_{1.7}$	$71.0_{0.5}$	$84.4_{0.3}$	$\mathbf{31.6_{1.3}}$	$\mathbf{45.3_{1.1}}$	$29.8_{1.4}$	$39.3_{1.6}$
SDC (11.3k)	$\mathbf{41.3_{1.0}}$	$\mathbf{59.7_{1.0}}$	$24.4_{2.2}$	$38.9_{2.9}$	$\mathbf{41.1_{0.8}}$	$\mathbf{51.8_{0.5}}$	$\mathbf{72.6_{0.6}}$	$84.6_{0.3}$	$29.5_{1.1}$	$43.3_{1.2}$	$\mathbf{35.6_{1.8}}$	$\mathbf{46.1_{1.7}}$
Orig + Fooled (34.4k)	$41.2_{1.2}$	$56.7_{0.9}$	$43.3_{1.4}$	$54.7_{1.6}$	$32.0_{0.7}$	$41.5_{1.0}$	$61.3_{2.3}$	$78.3_{1.2}$	$31.7_{0.6}$	$45.7_{1.0}$	$37.6_{2.5}$	$48.0_{2.6}$
Orig + Random (34.4k)	$\mathbf{45.7_{1.0}}$	$\mathbf{62.2_{0.8}}$	$\mathbf{46.5_{1.4}}$	$\mathbf{58.0_{1.2}}$	$38.9_{0.9}$	$48.9_{0.8}$	$67.6_{1.2}$	$82.6_{0.9}$	$33.6_{1.1}$	$\mathbf{47.1_{0.7}}$	$40.0_{1.6}$	$50.3_{1.7}$
Orig + SDC (34.4k)	$43.1_{0.8}$	$60.9_{0.4}$	$40.2_{1.4}$	$53.8_{0.8}$	$\mathbf{40.0_{1.4}}$	$\mathbf{51.9_{1.5}}$	$\mathbf{70.9_{0.4}}$	$\mathbf{83.3_{0.4}}$	$32.9_{0.8}$	$45.7_{0.7}$	$40.9_{1.1}$	$51.9_{1.3}$

	HotpotQA		Natural Questions		NewsQA		SearchQA		SQuAD		TriviaQA	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Original (23.1k)	48.1	63.5	55.3	67.6	38.6	54.4	39.7	49.3	61.9	76.7	47.5	59.6
Original (11.3k)	$46.6_{0.3}$	$63.2_{0.3}$	$54.6_{0.4}$	$66.9_{0.4}$	$36.3_{1.0}$	$51.6_{1.2}$	$33.8_{0.8}$	$43.0_{0.6}$	$60.1_{0.4}$	$75.3_{0.3}$	$44.9_{0.6}$	$57.2_{0.7}$
BERT$_{\text{fooled}}$ (11.3k)	$46.5_{0.8}$	$63.3_{0.8}$	$41.6_{1.2}$	$56.6_{1.1}$	$33.8_{1.2}$	$50.7_{1.6}$	$15.3_{1.9}$	$21.5_{1.9}$	$60.0_{0.6}$	$77.6_{0.5}$	$37.0_{1.7}$	$45.9_{2.1}$
BERT$_{\text{random}}$ (11.3k)	$50.7_{0.6}$	$67.7_{0.7}$	$48.1_{0.9}$	$62.6_{0.8}$	$39.5_{0.8}$	$56.1_{1.1}$	$17.0_{1.7}$	$23.6_{1.8}$	$65.4_{0.4}$	$81.4_{0.3}$	$43.3_{1.1}$	$52.5_{1.2}$
SDC (11.3k)	$\mathbf{52.0_{1.3}}$	$68.7_{1.4}$	$47.9_{1.2}$	$61.7_{1.3}$	$\mathbf{44.0_{0.9}}$	$\mathbf{61.9_{0.7}}$	$\mathbf{24.9_{2.0}}$	$\mathbf{33.0_{2.0}}$	$\mathbf{66.4_{0.6}}$	$\mathbf{82.2_{0.5}}$	$\mathbf{47.0_{0.6}}$	$\mathbf{58.3_{0.7}}$
Orig + Fooled (34.4k)	$47.2_{1.1}$	$64.7_{1.1}$	$53.2_{0.7}$	$66.8_{0.6}$	$33.9_{0.7}$	$52.0_{0.7}$	$28.2_{2.1}$	$35.3_{2.5}$	$58.2_{0.8}$	$76.9_{0.6}$	$38.8_{0.9}$	$48.6_{1.0}$
Orig + Random (34.4k)	$53.2_{0.5}$	$70.1_{0.5}$	$54.8_{0.4}$	$68.2_{0.3}$	$41.6_{0.6}$	$58.9_{0.7}$	$30.6_{1.9}$	$38.3_{2.0}$	$65.3_{0.5}$	$81.8_{0.3}$	$46.7_{1.0}$	$57.1_{0.9}$
Orig + SDC (34.4k)	$53.9_{0.9}$	$70.7_{0.9}$	$\mathbf{55.9_{0.4}}$	$\mathbf{68.7_{0.5}}$	$\mathbf{44.2_{0.3}}$	$\mathbf{62.5_{0.4}}$	$\mathbf{36.0_{1.3}}$	$\mathbf{45.2_{1.6}}$	$\mathbf{66.6_{0.4}}$	$\mathbf{82.7_{0.2}}$	$\mathbf{48.0_{0.8}}$	$\mathbf{59.8_{0.7}}$

Finetuned model: ELECTRA$_{\text{large}}$

Evaluation set →	BioASQ		DROP		DuoRC		Relation Extraction		RACE		TextbookQA	
Training set ↓	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Original (23.1k)	29.1	42.8	17.6	26.9	18.9	27.1	53.4	67.4	19.6	28.5	32.5	41.8
Original (11.3k)	$33.1_{1.4}$	$49.4_{2.5}$	$15.5_{1.8}$	$26.5_{1.1}$	$21.2_{0.8}$	$29.4_{0.6}$	$54.9_{0.9}$	$69.4_{1.1}$	$18.0_{0.8}$	$28.4_{0.7}$	$29.2_{0.5}$	$37.8_{0.3}$
BERT$_{\text{fooled}}$ (11.3k)	$32.4_{4.6}$	$50.2_{3.6}$	$19.9_{4.3}$	$33.4_{3.5}$	$25.2_{4.2}$	$35.1_{3.7}$	$57.0_{4.9}$	$74.6_{3.1}$	$20.6_{2.5}$	$34.0_{2.5}$	$19.5_{3.3}$	$28.5_{4.0}$
BERT$_{\text{random}}$ (11.3k)	$37.1_{2.9}$	$55.1_{2.1}$	$\mathbf{21.1_{1.9}}$	$\mathbf{35.0_{1.6}}$	$30.5_{2.1}$	$40.3_{1.6}$	$64.3_{2.9}$	$78.7_{1.3}$	$23.3_{1.5}$	$36.5_{1.5}$	$25.7_{3.3}$	$35.1_{3.5}$
SDC (11.3k)	$\mathbf{40.6_{1.7}}$	$\mathbf{59.2_{1.4}}$	$17.5_{0.9}$	$30.7_{1.1}$	$\mathbf{33.3_{2.1}}$	$\mathbf{43.6_{1.9}}$	$65.9_{1.4}$	$79.6_{0.8}$	$23.4_{1.1}$	$35.5_{1.0}$	$27.4_{2.7}$	$36.8_{2.9}$
Orig + Fooled (34.4k)	$31.7_{1.3}$	$48.2_{1.3}$	$19.9_{0.9}$	$31.0_{0.8}$	$24.5_{0.9}$	$33.1_{1.4}$	$55.0_{1.7}$	$71.5_{1.2}$	$19.2_{1.3}$	$31.0_{1.1}$	$22.2_{4.7}$	$30.9_{5.4}$
Orig + Random (34.4k)	$37.8_{5.2}$	$54.4_{5.4}$	$\mathbf{27.6_{6.8}}$	$\mathbf{39.4_{8.1}}$	$28.4_{5.1}$	$38.2_{5.7}$	$62.9_{6.8}$	$77.2_{5.2}$	$\mathbf{24.3_{4.6}}$	$\mathbf{37.4_{5.3}}$	$\mathbf{34.0_{6.1}}$	$\mathbf{43.5_{6.2}}$
Orig + SDC (34.4k)	$\mathbf{40.0_{0.9}}$	$\mathbf{57.6_{0.9}}$	$19.4_{0.9}$	$31.1_{1.0}$	$\mathbf{31.9_{0.4}}$	$\mathbf{41.6_{0.6}}$	$62.4_{0.7}$	$76.8_{0.2}$	$19.5_{1.4}$	$31.7_{1.2}$	$29.0_{2.4}$	$38.8_{3.1}$

	HotpotQA		Natural Questions		NewsQA		SearchQA		SQuAD		TriviaQA	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Original (23.1k)	29.6	43	40.9	55.3	20.4	32.2	21.5	30.3	39.9	54.8	21	31.2
Original (11.3k)	$26.8_{0.2}$	$39.7_{0.2}$	$38.7_{0.9}$	$54.2_{0.9}$	$21.0_{1.0}$	$33.2_{1.1}$	$17.2_{1.5}$	$24.8_{1.6}$	$40.5_{1.2}$	$55.9_{1.2}$	$23.9_{1.8}$	$33.5_{1.8}$
BERT$_{\text{fooled}}$ (11.3k)	$36.7_{4.0}$	$54.2_{2.9}$	$35.1_{3.8}$	$51.7_{3.1}$	$28.5_{2.4}$	$45.1_{2.4}$	$7.0_{1.3}$	$13.9_{1.7}$	$48.3_{4.2}$	$67.5_{3.4}$	$23.8_{2.9}$	$34.5_{2.3}$
BERT$_{\text{random}}$ (11.3k)	$41.4_{2.4}$	$58.4_{1.6}$	$43.2_{1.7}$	$58.5_{1.3}$	$33.3_{1.6}$	$49.8_{1.6}$	$9.2_{1.5}$	$16.8_{2.1}$	$55.4_{2.3}$	$72.9_{1.7}$	$28.9_{1.4}$	$39.9_{1.0}$
SDC (11.3k)	$43.0_{1.4}$	$59.6_{1.1}$	$\mathbf{46.1_{1.0}}$	$\mathbf{60.4_{0.8}}$	$\mathbf{35.3_{1.1}}$	$\mathbf{51.9_{1.1}}$	$10.5_{1.4}$	$\mathbf{19.0_{1.6}}$	$\mathbf{58.6_{1.4}}$	$\mathbf{74.9_{1.0}}$	$29.0_{1.6}$	$60.7_{1.3}$
Orig + Fooled (34.4k)	$34.4_{1.0}$	$51.1_{0.8}$	$45.4_{2.9}$	$59.9_{2.6}$	$26.3_{0.9}$	$42.8_{1.1}$	$8.7_{1.5}$	$14.5_{1.7}$	$47.6_{0.5}$	$66.3_{0.5}$	$21.9_{0.7}$	$30.9_{0.8}$
Orig + Random (34.4k)	$41.4_{4.7}$	$57.4_{4.5}$	$46.2_{3.8}$	$60.0_{3.5}$	$31.7_{4.2}$	$47.5_{5.2}$	$14.9_{2.2}$	$23.1_{2.2}$	$55.2_{4.6}$	$72.1_{4.6}$	$29.8_{5.2}$	$40.2_{5.2}$
Orig + SDC (34.4k)	$\mathbf{43.9_{0.5}}$	$\mathbf{60.4_{0.3}}$	$49.4_{0.5}$	$63.0_{0.7}$	$\mathbf{32.4_{0.7}}$	$\mathbf{49.0_{0.8}}$	$13.6_{0.4}$	$22.2_{0.5}$	$\mathbf{57.6_{1.0}}$	$\mathbf{74.0_{1.0}}$	$\mathbf{31.7_{0.8}}$	$\mathbf{43.4_{0.6}}$

Table 5.6: EM and F1 scores of various models evaluated on MRQA dev and test sets. Adversarial results in bold are statistically significant compared to SDC setting and vice versa.

Finetuned model: BERT$_{\text{large}}$

Evaluation set →	BioASQ		DROP		DuoRC		Relation Extraction		RACE		TextbookQA	
Training set ↓	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Original (23.1k)	19.4	32.5	7.8	16.2	14.5	22.8	32.0	47.1	11.4	18.8	25.0	33.4
Original (14.6k)	$20.4_{0.3}$	$35.9_{0.7}$	$5.1_{0.3}$	$12.4_{0.3}$	$11.6_{0.4}$	$17.8_{0.6}$	$33.0_{0.9}$	$44.2_{2.0}$	$10.4_{0.6}$	$17.7_{0.9}$	$19.5_{0.6}$	$27.3_{0.7}$
ELECTRA$_{\text{fooled}}$ (14.6k)	$13.6_{0.9}$	$29.1_{1.1}$	$3.2_{0.4}$	$11.9_{0.7}$	$11.0_{0.9}$	$19.3_{0.6}$	$33.6_{2.2}$	$52.5_{2.3}$	$7.9_{0.7}$	$17.7_{0.8}$	$12.2_{1.7}$	$21.2_{1.8}$
ELECTRA$_{\text{random}}$ (14.6k)	$15.9_{0.8}$	$32.0_{1.7}$	$3.1_{0.4}$	$10.5_{0.9}$	$12.1_{0.9}$	$20.4_{1.4}$	$35.7_{3.1}$	$55.6_{3.7}$	$9.5_{0.7}$	$19.1_{0.8}$	$14.6_{1.8}$	$23.9_{1.8}$
SDC (14.6k)	$\mathbf{17.1_{0.7}}$	$\mathbf{34.5_{1.0}}$	$2.6_{0.3}$	$10.1_{0.9}$	$11.9_{0.8}$	$21.2_{1.2}$	$34.2_{3.4}$	$53.7_{4.1}$	$9.2_{1.0}$	$19.0_{0.7}$	$\mathbf{17.5_{1.1}}$	$\mathbf{27.4_{1.3}}$
Orig + Fooled (37.7k)	$17.8_{1.0}$	$33.5_{2.0}$	$6.1_{1.1}$	$16.1_{1.7}$	$14.2_{1.4}$	$22.9_{1.9}$	$42.0_{2.2}$	$59.6_{2.5}$	$12.0_{0.9}$	$22.2_{0.9}$	$24.6_{1.0}$	$33.7_{1.2}$
Orig + Random (37.7k)	$20.0_{1.1}$	$36.4_{1.6}$	$6.8_{0.9}$	$17.1_{1.0}$	$14.6_{1.0}$	$23.5_{1.5}$	$44.0_{1.3}$	$61.8_{1.3}$	$12.0_{0.9}$	$22.0_{0.9}$	$23.9_{0.8}$	$33.5_{1.0}$
Orig + SDC (37.7k)	$\mathbf{21.8_{0.6}}$	$\mathbf{39.2_{1.1}}$	$6.1_{0.5}$	$16.1_{0.7}$	$\mathbf{16.7_{0.9}}$	$\mathbf{25.9_{1.0}}$	$\mathbf{43.4_{0.7}}$	$61.0_{1.1}$	$11.9_{0.7}$	$22.5_{0.7}$	$\mathbf{25.4_{0.5}}$	$\mathbf{35.5_{0.6}}$

	HotpotQA		Natural Questions		NewsQA		SearchQA		SQuAD		TriviaQA	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Original (23.1k)	19.4	33.9	36.3	48.7	16.2	25.6	11.3	19.3	32.5	46.0	16.8	25.3
Original (14.6k)	$17.4_{0.9}$	$28.7_{1.2}$	$35.0_{0.7}$	$47.7_{0.7}$	$12.8_{0.2}$	$22.6_{0.1}$	$9.0_{0.1}$	$13.8_{0.4}$	$26.0_{0.3}$	$39.2_{0.7}$	$11.8_{0.5}$	$18.2_{0.7}$
ELECTRA$_{\text{fooled}}$ (14.6k)	$19.1_{0.7}$	$33.4_{0.8}$	$28.0_{1.4}$	$43.1_{1.4}$	$12.9_{0.8}$	$25.9_{0.8}$	$4.0_{0.3}$	$9.1_{0.5}$	$26.9_{1.4}$	$46.4_{1.4}$	$9.2_{0.8}$	$16.3_{1.1}$
ELECTRA$_{\text{random}}$ (14.6k)	$21.2_{1.0}$	$35.5_{1.3}$	$29.0_{2.3}$	$43.8_{2.3}$	$13.8_{0.8}$	$27.1_{1.3}$	$4.2_{0.4}$	$9.1_{0.6}$	$29.2_{1.6}$	$48.3_{2.2}$	$10.0_{0.7}$	$17.3_{1.2}$
SDC (14.6k)	$\mathbf{23.5_{1.2}}$	$\mathbf{37.8_{1.3}}$	$28.4_{1.7}$	$43.5_{1.4}$	$\mathbf{15.6_{0.8}}$	$\mathbf{30.3_{1.0}}$	$\mathbf{5.0_{0.5}}$	$\mathbf{9.9_{0.7}}$	$\mathbf{31.5_{0.7}}$	$\mathbf{50.5_{0.8}}$	$10.0_{0.9}$	$\mathbf{19.1_{1.3}}$
Orig + Fooled (37.7k)	$25.5_{1.4}$	$40.8_{1.5}$	$38.5_{1.1}$	$52.2_{1.1}$	$17.0_{0.7}$	$30.9_{1.2}$	$9.9_{0.4}$	$15.8_{0.8}$	$32.7_{1.5}$	$51.7_{1.5}$	$14.2_{1.6}$	$22.6_{1.8}$
Orig + Random (37.7k)	$26.7_{1.2}$	$41.9_{1.2}$	$38.6_{1.0}$	$52.6_{0.7}$	$17.0_{0.4}$	$30.7_{0.7}$	$9.2_{0.9}$	$14.6_{1.2}$	$34.3_{0.6}$	$53.3_{0.8}$	$14.1_{0.7}$	$22.7_{1.1}$
Orig + SDC (37.7k)	$29.0_{1.0}$	$42.6_{0.8}$	$38.7_{0.3}$	$52.4_{0.1}$	$\mathbf{18.7_{0.6}}$	$\mathbf{33.9_{0.5}}$	$\mathbf{11.1_{0.7}}$	$\mathbf{16.6_{0.9}}$	$\mathbf{36.1_{0.7}}$	$\mathbf{54.9_{0.5}}$	$\mathbf{15.1_{0.3}}$	$\mathbf{24.2_{0.2}}$

Finetuned model: RoBERTa$_{\text{large}}$

Evaluation set →	BioASQ		DROP		DuoRC		Relation Extraction		RACE		TextbookQA	
Training set ↓	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Original (23.1k)	47.7	63.5	37.2	48.1	38.6	49.1	74.4	85.9	33.7	44.9	36.4	46
Original (14.6k)	$45.4_{1.7}$	$61.8_{1.0}$	$37.5_{1.7}$	$48.7_{2.0}$	$37.8_{0.7}$	$48.7_{0.8}$	$75.0_{0.6}$	$86.0_{0.2}$	$32.4_{0.7}$	$43.4_{0.9}$	$36.8_{1.1}$	$46.2_{1.3}$
ELECTRA$_{\text{fooled}}$ (14.6k)	$41.2_{1.4}$	$57.2_{1.1}$	$30.3_{1.7}$	$44.9_{1.8}$	$37.9_{2.1}$	$47.2_{2.3}$	$74.1_{0.8}$	$86.0_{0.4}$	$31.7_{1.3}$	$45.4_{1.0}$	$30.8_{1.7}$	$40.5_{1.8}$
ELECTRA$_{\text{random}}$ (14.6k)	$43.3_{1.4}$	$60.0_{1.5}$	$\mathbf{34.1_{2.4}}$	$\mathbf{48.8_{2.0}}$	$39.2_{1.5}$	$48.8_{1.6}$	$75.5_{0.5}$	$85.9_{0.2}$	$\mathbf{32.6_{0.7}}$	$46.3_{0.5}$	$32.2_{1.2}$	$42.2_{1.4}$
SDC (14.6k)	$43.7_{1.0}$	$\mathbf{62.5_{0.7}}$	$27.5_{2.6}$	$43.4_{2.9}$	$\mathbf{42.3_{0.9}}$	$\mathbf{53.5_{1.1}}$	$74.9_{0.8}$	$85.3_{0.7}$	$31.5_{0.9}$	$46.0_{1.0}$	$\mathbf{36.3_{2.0}}$	$\mathbf{47.2_{2.0}}$
Orig + Fooled (37.7k)	$45.0_{1.2}$	$61.2_{1.0}$	$\mathbf{45.9_{1.6}}$	$\mathbf{58.1_{1.3}}$	$36.8_{1.4}$	$47.2_{1.7}$	$73.9_{0.4}$	$86.3_{0.3}$	$33.7_{0.9}$	$47.3_{0.9}$	$38.5_{0.9}$	$48.3_{1.2}$
Orig + Random (37.7k)	$46.3_{1.0}$	$62.6_{0.8}$	$45.5_{1.2}$	$57.8_{0.8}$	$39.1_{1.3}$	$49.3_{1.3}$	$74.7_{0.5}$	$86.6_{0.2}$	$34.1_{0.2}$	$47.2_{0.4}$	$39.9_{1.5}$	$49.9_{1.9}$
Orig + SDC (37.7k)	$\mathbf{47.5_{0.5}}$	$\mathbf{64.0_{0.5}}$	$42.7_{1.1}$	$55.5_{1.0}$	$\mathbf{42.1_{1.3}}$	$\mathbf{53.7_{1.1}}$	$74.7_{0.9}$	$86.9_{0.5}$	$33.9_{1.2}$	$47.3_{1.0}$	$\mathbf{41.9_{0.4}}$	$\mathbf{52.5_{0.3}}$

	HotpotQA		Natural Questions		NewsQA		SearchQA		SQuAD		TriviaQA	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Original (23.1k)	19.4	33.9	36.3	48.7	16.2	25.6	11.3	19.3	32.5	46.0	16.8	25.3
Original (14.6k)	$47.0_{0.3}$	$62.7_{0.3}$	$55.6_{0.4}$	$67.5_{0.5}$	$38.2_{0.2}$	$53.6_{0.3}$	$34.5_{0.8}$	$43.8_{0.6}$	$60.5_{0.4}$	$75.6_{0.5}$	$46.5_{0.5}$	$58.5_{0.7}$
ELECTRA$_{\text{fooled}}$ (14.6k)	$51.9_{0.9}$	$67.9_{1.0}$	$49.6_{0.6}$	$64.1_{0.7}$	$37.8_{0.9}$	$54.9_{1.0}$	$24.0_{2.0}$	$31.3_{2.2}$	$66.2_{0.4}$	$82.0_{0.3}$	$45.1_{1.1}$	$55.2_{1.1}$
ELECTRA$_{\text{random}}$ (14.6k)	$54.5_{0.8}$	$71.0_{0.8}$	$51.6_{0.6}$	$65.9_{0.6}$	$40.2_{1.1}$	$57.7_{1.2}$	$24.3_{2.6}$	$32.9_{2.6}$	$66.9_{0.2}$	$82.6_{0.2}$	$45.8_{0.8}$	$56.2_{1.0}$
SDC (14.6k)	$\mathbf{55.8_{0.8}}$	$71.8_{0.8}$	$51.7_{0.5}$	$65.8_{0.5}$	$\mathbf{43.9_{0.8}}$	$\mathbf{62.1_{1.0}}$	$24.4_{2.4}$	$32.9_{2.4}$	$\mathbf{68.4_{0.5}}$	$\mathbf{84.3_{0.3}}$	$\mathbf{47.3_{0.7}}$	$\mathbf{59.1_{0.7}}$
Orig + Fooled (37.7k)	$55.6_{0.8}$	$71.7_{0.9}$	$57.1_{0.3}$	$69.6_{0.3}$	$40.6_{1.5}$	$57.7_{1.8}$	$38.3_{2.4}$	$47.3_{2.7}$	$67.0_{0.5}$	$82.7_{0.4}$	$46.7_{1.0}$	$57.5_{1.0}$
Orig + Random (37.7k)	$56.0_{0.2}$	$71.9_{0.3}$	$56.5_{0.2}$	$69.1_{0.3}$	$42.3_{0.3}$	$59.3_{0.7}$	$39.4_{1.6}$	$48.5_{1.7}$	$68.0_{0.2}$	$83.3_{0.2}$	$47.8_{0.3}$	$58.8_{0.3}$
Orig + SDC (37.7k)	$\mathbf{57.5_{0.7}}$	$\mathbf{72.8_{0.6}}$	$56.9_{0.3}$	$69.4_{0.3}$	$\mathbf{44.3_{0.7}}$	$\mathbf{62.7_{0.7}}$	$39.3_{1.0}$	$48.6_{1.1}$	$\mathbf{69.9_{0.4}}$	$\mathbf{84.3_{0.2}}$	$\mathbf{48.6_{0.5}}$	$\mathbf{60.1_{0.5}}$

Finetuned model: ELECTRA$_{\text{large}}$

Evaluation set →	BioASQ		DROP		DuoRC		Relation Extraction		RACE		TextbookQA	
Training set ↓	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Original (23.1k)	29.1	42.8	17.6	26.9	18.9	27.1	53.4	67.4	19.6	28.5	32.5	41.8
Original (14.6k)	$35.4_{0.4}$	$51.0_{0.8}$	$16.2_{0.5}$	$26.6_{0.8}$	$18.8_{0.4}$	$26.7_{0.8}$	$46.2_{1.3}$	$61.1_{1.7}$	$17.3_{0.9}$	$27.9_{0.6}$	$29.6_{0.6}$	$37.8_{0.7}$
ELECTRA$_{\text{fooled}}$ (14.6k)	$25.3_{1.1}$	$41.0_{1.6}$	$7.6_{0.9}$	$18.9_{1.4}$	$12.3_{1.5}$	$20.5_{2.0}$	$42.1_{2.0}$	$61.4_{2.3}$	$13.5_{0.6}$	$25.1_{1.0}$	$20.8_{2.5}$	$29.5_{2.9}$
ELECTRA$_{\text{random}}$ (14.6k)	$25.5_{4.9}$	$41.6_{5.5}$	$7.8_{2.6}$	$19.2_{5.3}$	$12.1_{2.3}$	$19.7_{2.9}$	$40.3_{7.7}$	$57.7_{9.4}$	$13.0_{2.7}$	$24.0_{3.7}$	$20.3_{3.5}$	$28.8_{3.4}$
SDC (14.6k)	$25.0_{7.5}$	$41.0_{1.7}$	$5.9_{2.1}$	$17.9_{4.4}$	$13.2_{3.0}$	$22.5_{4.9}$	$42.7_{6.6}$	$61.9_{7.5}$	$13.4_{2.7}$	$24.7_{4.0}$	$20.8_{3.8}$	$29.5_{3.4}$
Orig + Fooled (37.7k)	$28.4_{2.0}$	$45.2_{2.6}$	$15.6_{0.8}$	$28.6_{1.0}$	$13.3_{1.0}$	$21.2_{1.7}$	$41.5_{2.8}$	$60.5_{3.3}$	$17.6_{0.7}$	$29.6_{0.9}$	$32.2_{0.9}$	$41.6_{1.1}$
Orig + Random (37.7k)	$28.6_{1.6}$	$44.9_{2.0}$	$16.3_{0.6}$	$29.0_{1.2}$	$12.8_{1.0}$	$20.9_{1.6}$	$39.4_{3.3}$	$58.8_{3.6}$	$16.6_{1.3}$	$29.0_{1.1}$	$32.4_{0.4}$	$42.2_{0.5}$
Orig + SDC (37.7k)	$29.7_{1.9}$	$47.0_{2.2}$	$15.6_{0.8}$	$29.1_{1.3}$	$\mathbf{16.4_{0.7}}$	$\mathbf{27.1_{0.8}}$	$\mathbf{48.0_{1.8}}$	$\mathbf{67.0_{1.5}}$	$\mathbf{19.0_{0.6}}$	$\mathbf{32.1_{0.8}}$	$\mathbf{33.7_{0.4}}$	$\mathbf{43.8_{0.9}}$

	HotpotQA		Natural Questions		NewsQA		SearchQA		SQuAD		TriviaQA	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Original (23.1k)	19.4	33.9	36.3	48.7	16.2	25.6	11.3	19.3	32.5	46.0	16.8	25.3
Original (14.6k)	$23.2_{1.0}$	$40.2_{1.1}$	$33.4_{0.8}$	$49.8_{0.5}$	$17.9_{0.5}$	$31.1_{0.9}$	$16.0_{0.5}$	$22.3_{1.1}$	$31.1_{0.4}$	$50.1_{0.5}$	$21.0_{0.9}$	$29.8_{1.3}$
ELECTRA$_{\text{fooled}}$ (14.6k)	$26.2_{0.9}$	$42.2_{0.9}$	$31.5_{1.4}$	$49.7_{1.1}$	$18.7_{1.2}$	$32.1_{1.6}$	$6.5_{0.7}$	$10.4_{1.0}$	$34.5_{1.3}$	$53.7_{1.5}$	$13.2_{1.0}$	$21.5_{1.3}$
ELECTRA$_{\text{random}}$ (14.6k)	$24.7_{5.5}$	$40.9_{6.9}$	$27.9_{6.8}$	$45.7_{7.6}$	$17.2_{3.1}$	$30.8_{3.8}$	$6.4_{1.6}$	$10.3_{2.1}$	$34.1_{5.8}$	$53.1_{6.2}$	$12.4_{3.4}$	$20.1_{4.5}$
SDC (14.6k)	$24.4_{3.3}$	$41.7_{5.2}$	$28.8_{6.2}$	$46.7_{8.3}$	$19.2_{3.6}$	$\mathbf{35.5_{3.2}}$	$\mathbf{8.3_{0.9}}$	$\mathbf{12.8_{1.6}}$	$34.7_{4.2}$	$54.1_{5.1}$	$13.4_{2.0}$	$22.7_{3.5}$
Orig + Fooled (37.7k)	$28.5_{0.9}$	$45.8_{1.3}$	$35.0_{0.8}$	$52.5_{1.0}$	$20.3_{0.7}$	$34.9_{1.0}$	$14.3_{1.0}$	$19.8_{1.4}$	$36.7_{1.3}$	$56.5_{1.5}$	$15.3_{1.6}$	$24.3_{2.0}$
Orig + Random (37.7k)	$28.1_{1.5}$	$45.9_{1.3}$	$34.1_{1.1}$	$51.7_{1.1}$	$19.2_{1.1}$	$34.1_{1.8}$	$14.3_{0.8}$	$20.1_{1.3}$	$35.6_{1.7}$	$55.3_{1.4}$	$15.0_{1.4}$	$24.5_{2.0}$
Orig + SDC (37.7k)	$\mathbf{30.5_{1.1}}$	$\mathbf{47.8_{0.8}}$	$35.8_{1.1}$	$53.4_{0.8}$	$\mathbf{23.0_{0.7}}$	$\mathbf{40.2_{0.7}}$	$\mathbf{16.5_{0.6}}$	$\mathbf{22.8_{1.1}}$	$\mathbf{40.6_{0.6}}$	$\mathbf{60.7_{0.4}}$	$\mathbf{18.8_{0.8}}$	$\mathbf{30.0_{0.8}}$

Table 5.7: EM and F1 scores of various models evaluated on MRQA dev and test sets. Adversarial results in bold are statistically significant compared to SDC setting and vice versa.

Chapter 6

Does It Matter Who Gives The Feedback?

6.1 Overview

So far we've seen how human feedback can enrich machine learning models. As richer forms of human feedback can help train NLP models that are reliable when deployed in the real world, ensuring feedback received for a task is of high quality is top of the mind for researchers seeking to obtain this feedback at scale. While crowdsourcing platforms like Amazon Mechanical Turk offer easy access to a large human workforce, enabling researchers to easily collect feedback at scale, they also present challenges arising from a workforce that is not reflective of the general population. For instance, while $\approx 14\%$ of the US population is Hispanic or Latino,[1] they are only $\approx 3\%$[2] of AMT workers. Similarly, while $\approx 30\%$ of the US population has obtained at least a 4-year college degree, they make up for approximately 2/3rds of the crowdworkers on AMT.

Yet, whether or not diversity of crowdworkers matters in the kind of human feedback we receive remains an understudied topic. Aside from the fact that AMT workforce is not representative of the population, several crowdsourcing studies (our previous studies included) restrict their pool of workers to those in the United States, thereby so those providing feedback into a human in the loop system are not representative of the world's (or even the United States') population. These practices are commonly employed by academic as well as industry labs (for quality or legal purposes; an industry lab I worked at once could legally not recruit crowdworkers outside the United States), even though models trained with this feedback may be deployed worldwide [Wiegreffe et al., 2021, Wallace et al., 2022]. Thus, it stands to reason that there could be some benefits to diversifying the pool of workers. After all, preferences for a worker in India might be different than for one in the US. Given humans think in ways that are specific to their environment and cultural context, it stands to reason that if we want our machine learning models to be able to accurately capture human behavior, we need a diverse set of workers who can provide different perspectives.

Intuitively, for objective tasks (such as identifying dogs versus cats or identifying nouns in a document), it shouldn't matter who is performing the task. After all, there are indisputable objective answers in those cases. However, as we've seen in previous chapters, in response to machine

[1] Per US Census Data from 2021. Available at https://www.census.gov/quickfacts/fact/table/US/RHI725221.

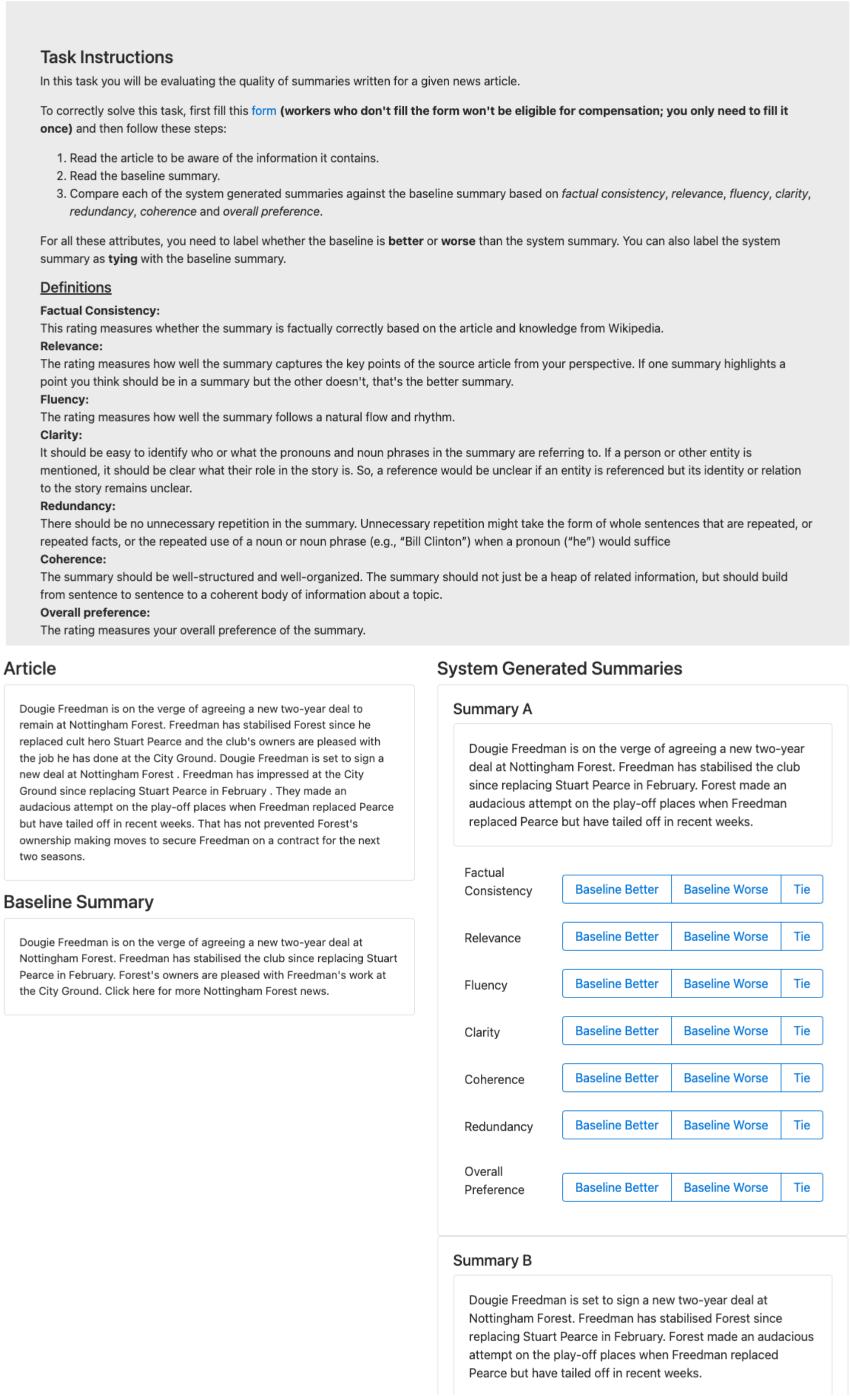
[2] Per data obtained from our self-reported demographic survey

learning (ML) growing to be a public concern, researchers have sought to employ crowdworkers in diverse ways to create datasets that might help to ensure the models are reliable under a variety of settings. Such feedback is often subjective and there could be several correct answers. Take for example our previous work on counterfactually augmented data in Chapter 3. A document can be edited in many different ways to make a counterfactual label applicable. Human editors have their own social biases and prior work has shown that people tend to edit phrases that they consider socially inappropriate, selecting counterfactuals subject to their social and functional norms [Byrne, 2016, Fazelpour, 2020]. Furthermore, in some cases, the interventions may not be on the direct causal ancestors of the label but on enabling causes [Catellani et al., 2020, Catellani and Milesi, 2001, Icard et al., 2017].

In this chapter, we investigate whether diversity among workers who provide this feedback might make a difference in what feedback is received. We design a suite of crowdsourcing studies to study diverse forms of feedback across three different tasks: (i) select one of two plausible summaries for a given article (obtaining binary feedback); (ii) converse with a dialog system in real-time and rate the conversation (obtaining ratings that could be used as reward signal); and (iii) write five questions for a given topic (obtaining free-form text) For the first task, since the objective is not to train a new model but examine the differences between human feedback, instead of training an abstractive summarizer and generating plausible summaries from it, we select 100 news articles from the CNN-DailyMail corpus [See et al., 2017] and use existing summaries generated by three recent summarization models—T5 [Raffel et al., 2020], GPT-2 [Radford et al., 2018], and BART [Lewis et al., 2020]—as plausible summaries. We examine how the choice of location of crowdworkers could impact which summaries crowdworkers select for these articles and run this study both in the United States and in India. We show each article and two plausible summaries to at least 10 crowdworkers in each country. For the interactive dialog setup, we continue to assess whether crowdworkers in the United States versus those in India offer different ratings for their interactions with a model. We collect at least 250 dialogs for three recent open-domain dialog models—BlenderBot 3B, BlenderBot 90M, and Reddit 3B [Roller et al., 2021]—from each country. In both setups we find significant differences in the feedback received from crowdworkers in India versus those in the United States. We then collect natural questions written by crowdworkers to analyze whether such differences could appear in generative feedback as well (in this case, writing questions for a question answering task), and analyze the embeddings of these questions. This analysis further appears to confirm our hypothesis, thus suggesting that diversifying the pool of workers who offer feedback to machine learning models could potentially provide insights that would be missed otherwise.

6.2 Related Work

Bhuiyan et al. [2020] investigate news credibility assessments by crowds versus experts to understand when and how ratings between them differ. They gather a dataset of thousands of credibility assessments taken from journalism students, Upwork workers, expert journalists and expert scientists, on a varied set of 50 news articles related to climate science. They examine these ratings to find differences in performance due to rater demographics, political leanings, genre of the article, and partisanship of the publication. Abebe et al. [2019] analyze Bing search



Task Instructions

In this task you will be evaluating the quality of summaries written for a given news article.

To correctly solve this task, first fill this form **(workers who don't fill the form won't be eligible for compensation; you only need to fill it once)** and then follow these steps:

1. Read the article to be aware of the information it contains.
2. Read the baseline summary.
3. Compare each of the system generated summaries against the baseline summary based on *factual consistency*, *relevance*, *fluency*, *clarity*, *redundancy*, *coherence* and *overall preference*.

For all these attributes, you need to label whether the baseline is **better** or **worse** than the system summary. You can also label the system summary as **tying** with the baseline summary.

Definitions

Factual Consistency:
This rating measures whether the summary is factually correctly based on the article and knowledge from Wikipedia.
Relevance:
The rating measures how well the summary captures the key points of the source article from your perspective. If one summary highlights a point you think should be in a summary but the other doesn't, that's the better summary.
Fluency:
The rating measures how well the summary follows a natural flow and rhythm.
Clarity:
It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.
Redundancy:
There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., "Bill Clinton") when a pronoun ("he") would suffice
Coherence:
The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.
Overall preference:
The rating measures your overall preference of the summary.

Article

Dougie Freedman is on the verge of agreeing a new two-year deal to remain at Nottingham Forest. Freedman has stabilised Forest since he replaced cult hero Stuart Pearce and the club's owners are pleased with the job he has done at the City Ground. Dougie Freedman is set to sign a new deal at Nottingham Forest . Freedman has impressed at the City Ground since replacing Stuart Pearce in February . They made an audacious attempt on the play-off places when Freedman replaced Pearce but have tailed off in recent weeks. That has not prevented Forest's ownership making moves to secure Freedman on a contract for the next two seasons.

Baseline Summary

Dougie Freedman is on the verge of agreeing a new two-year deal at Nottingham Forest. Freedman has stabilised the club since replacing Stuart Pearce in February. Forest's owners are pleased with Freedman's work at the City Ground. Click here for more Nottingham Forest news.

System Generated Summaries

Summary A

Dougie Freedman is on the verge of agreeing a new two-year deal at Nottingham Forest. Freedman has stabilised the club since replacing Stuart Pearce in February. Forest made an audacious attempt on the play-off places when Freedman replaced Pearce but have tailed off in recent weeks.

Summary B

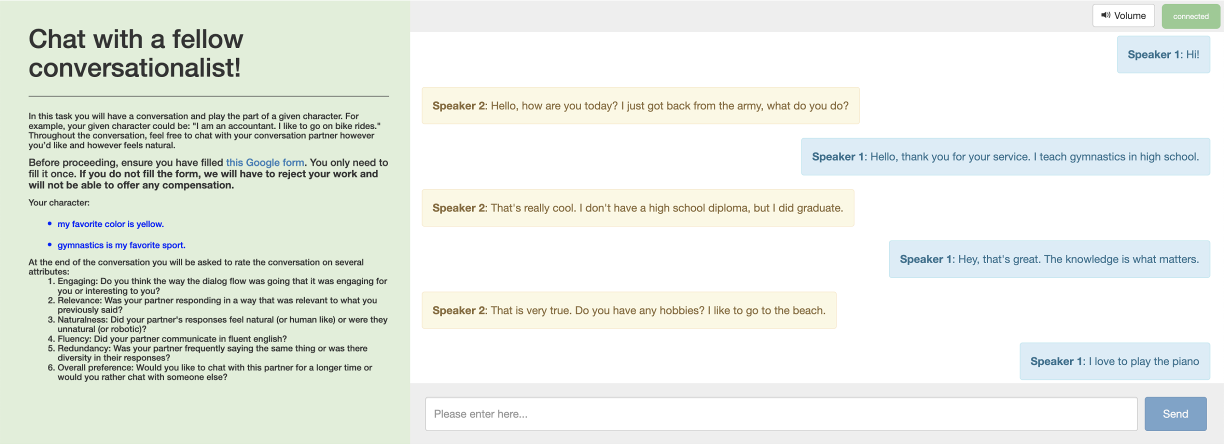
Dougie Freedman is set to sign a new two-year deal at Nottingham Forest. Freedman has stabilised Forest since replacing Stuart Pearce in February. Forest made an audacious attempt on the play-off places when Freedman replaced Pearce but have tailed off in recent weeks.

Figure 6.1: Crowdsourcing platform for summarization feedback collection

queries from several African nations and find that different patterns emerge in health information needs by demographic groups (age and gender) and country. They also find great discrepancies in the quality of content returned by search engines to users by topic, highlighting differences in user behavior and satisfaction. For instance, their analysis reveals that topics related to news on

HIV/AIDS cures are more popular among men. On the other hand, women seek information on topics related to breastfeeding, pregnancy, and family care. They also find different information needs across age groups. For instance, while topics related to the socioeconomic implications of HIV/AIDS, such as gender inequality, are more popular in the 18–24 age group, topics related to concerns about transmission to partner and child are more popular among the 25–34 age group.

In NLP, focusing on evaluating natural language generation models, Karpinska et al. [2021] run a series of story evaluation experiments with both crowdworkers and English teachers and discover that crowdworkers (unlike teachers) fail to distinguish between model-generated text and human-generated references. Sap et al. [2021] conduct two online studies with demographically and politically diverse participants to investigate the effects of annotator identities (who) and beliefs (why) when annotating hate speech datasets. They disentangle what is annotated as toxic by considering posts with three characteristics: anti-Black language, African American English (AAE) dialect, and vulgarity. They find that more conservative annotators and those who scored highly on their scale for racist beliefs were less likely to rate anti-Black language as toxic, but more likely to rate AAE as toxic, and found strong associations between annotator identity and beliefs and their ratings of toxicity generally. Similarly, in another work, Prabhakaran et al. [2021] analyze annotated data for several NLP tasks and study the impact of majority voting as an aggregation approach. They find that in the annotations for many tasks, the aggregated majority vote does not uniformly reflect the perspectives of all annotators in the annotator pool and for many tasks in their analysis, a significant proportion of the annotators had very low agreement scores with the majority vote label. They further find that these agreement scores may vary significantly across different socio-demographic groups that annotators identify with.



Chat with a fellow conversationalist!

In this task you will have a conversation and play the part of a given character. For example, your given character could be: "I am an accountant. I like to go on bike rides." Throughout the conversation, feel free to chat with your conversation partner however you'd like and however feels natural.

Before proceeding, ensure you have filled this Google form. You only need to fill it once. **If you do not fill the form, we will have to reject your work and will not be able to offer any compensation.**

Your character:

- my favorite color is yellow.
- gymnastics is my favorite sport.

At the end of the conversation you will be asked to rate the conversation on several attributes:

1. Engaging: Do you think the way the dialog flow was going that it was engaging for you or interesting to you?
2. Relevance: Was your partner responding in a way that was relevant to what you previously said?
3. Naturalness: Did your partner's responses feel natural (or human like) or were they unnatural (or robotic)?
4. Fluency: Did your partner communicate in fluent english?
5. Redundancy: Was your partner frequently saying the same thing or was there diversity in their responses?
6. Overall preference: Would you like to chat with this partner for a longer time or would you rather chat with someone else?

Speaker 1: Hi!

Speaker 2: Hello, how are you today? I just got back from the army, what do you do?

Speaker 1: Hello, thank you for your service. I teach gymnastics in high school.

Speaker 2: That's really cool. I don't have a high school diploma, but I did graduate.

Speaker 1: Hey, that's great. The knowledge is what matters.

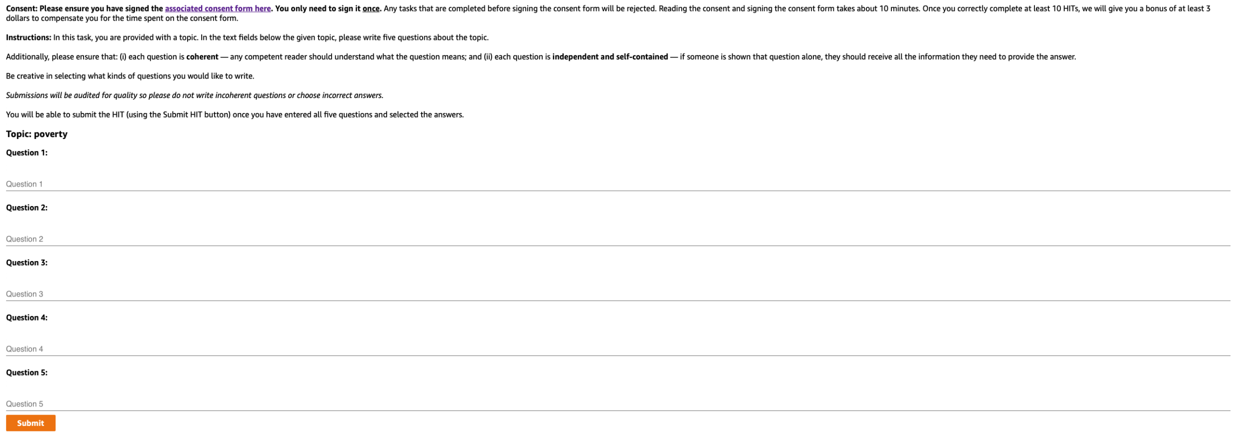
Speaker 2: That is very true. Do you have any hobbies? I like to go to the beach.

Speaker 1: I love to play the piano

Figure 6.2: Crowdsourcing platform for open domain dialog feedback collection

6.3 Data and Methods

We use Amazon's Mechanical Turk crowdsourcing platform to recruit crowdworkers for our study. To ensure high quality of the collected data, we restricted the pool to workers who had already completed at least 5000 HITs and had an over 97% HIT approval rate. For each task,



Consent: Please ensure you have signed the associated consent form here. You only need to sign it once. Any tasks that are completed before signing the consent form will be rejected. Reading the consent and signing the consent form takes about 10 minutes. Once you correctly complete at least 10 HITs, we will give you a bonus of at least 3 dollars to compensate you for the time spent on the consent form.

Instructions: In this task, you are provided with a topic. In the text fields below the given topic, please write five questions about the topic.

Additionally, please ensure that: (i) each question is **coherent** — any competent reader should understand what the question means; and (ii) each question is **independent and self-contained** — if someone is shown that question alone, they should receive all the information they need to provide the answer.

Be creative in selecting what kinds of questions you would like to write.

Submissions will be audited for quality so please do not write incoherent questions or choose incorrect answers.

You will be able to submit the HIT (using the Submit HIT button) once you have entered all five questions and selected the answers.

Topic: poverty

Question 1:

Question 1

Question 2:

Question 2

Question 3:

Question 3

Question 4:

Question 4

Question 5:

Question 5

Figure 6.3: Crowdsourcing platform for collecting questions

we conducted several pilot studies to gather feedback from crowdworkers. We identified median time taken by workers to complete each task in our pilot studies and used that to design incentive structure for the respective task. We also conducted multiple studies with different variants of instructions to observe trends in the quality of questions and refined our instructions based on feedback from crowdworkers. Feedback from the pilots also guided improvements to our crowdsourcing interface. For both summarization and dialog, we focus on the country of crowdworkers, thus collecting data from both the United States and India. Whereas, for, question generation we only collect data from workers in the United States to perform comparisons across gender, race and ethnicity, and political leaning. In total, over 3500 workers participated in this study.

News summarization We select 100 articles from the CNN-DailyMail dataset and their summaries created by three recent, high-performing abstractive summarization models—T5, GPT-2 and BART.[3] We show an article along with a baseline summary (chosen as one of the three models) and another summary (one of the remaining two models) and ask crowdworkers to select which of the two summaries *better* exhibits certain attributes (in their opinion). We focus on six criteria on which a user must provide binary feedback: which summary is more **factually consistent** w.r.t the article, which summary captures **relevant** points of the article, which summary is more **fluent**, which summary is more **clear** and easy to understand, which summary is more **coherent**, which summary has less **redundancy**. We further ask crowdworkers to also select which of the two summaries they would have preferred overall. Each crowdworker is restricted to a maximum of 5 HITs. Crowdworkers are paid $1 per HIT. To ensure workers were paying attention, we inserted attention checks where the two summaries were the same with spacing differences so they appear different in length. Workers not paying attention would likely select one summary to be better than the other rather than selecting the *Tie* option.

[3] These articles are a subset of the articles used by Fabbri et al. [2021]. This choice of articles allows us to use existing, high quality summaries since the objective is not to train a new model but examine the differences between human feedback.

Open Domain Dialog For open domain dialog feedback, we make use of the single-model, per-dialog evaluation setup from [Smith et al., 2022]. We use three recent open domain conversation models—BlenderBot 3B, BlenderBot 90M, and Reddit 3B. BlenderBot 3B and BlenderBot 90M differ in the number of parameters (3 billion and 90 million, respectively) and Reddit 3B is simply BlenderBot 3B before it was finetuned on dialog datasets. We design an interactive platform using ParlAI and Mephisto. Each interaction begins with a default *Hi!* prompt to the conversational agent. A crowdworker engages in a live interaction with a conversational agent for 8 turns following the agent's first response. Once a dialog is complete, the worker is asked to rate the whole conversation on five criteria: how **engaging** was the conversation, how **relevant** were the agent's responses, how **natural** (or robotic) the agent's responses sounded, and **fluency** of agent's responses. For each model, we deploy 280 HITs on Amazon Mechanical Turk, both in India, and in the United States (allowing for a buffer assuming $\approx 10\%$ of the HITs may result in spam). We pay workers $\$2$ per HIT. Each worker does at most 2 HITs per model (a maximum of 8 HITs per worker).

Question Answering In this setup, for a given topic, we ask a crowdworker to write 5 questions that they are genuinely curious about. We identify a mix of 51 topics, some of which have been shown to invoke polarizing sentiments amongst different demographics such as abortion, affirmative action and men's rights, and others where we do not have a reason to expect such differences to arise, such as AIDS, artificial intelligence, obesity, and sports. Following IRB approval for this study, workers were asked to consent to participate in this study and only upon consenting, they could self report their demographics in a Google Survey and start writing questions. We conducted a range of pilots to approximate the proportion of different demographic groups in our worker pool. Our objective was to ensure we collected at least 3000 questions from each demographic subgroup to perform meaningful comparisons, and this would also allow us to train QA models with ease (an insight from our experiments in Chapter 5). As Black respondents were the smallest subgroup in our pilot studies ($\approx 5\%$ of respondents), we collected a total of 61000 questions by showing each topic to 239 crowdworkers. As a result, this data contained 3080 questions created by Black respondents. We pay workers $\$0.4$ per HIT to generate these questions. While our initial goal was to collect questions, and then recruit additional crowdworkers to find relevant Wikipedia passages that could answer these questions, our initial analysis revealed several challenges with that approach. Thus, for now, we pause here and restrict our analysis to differences in the questions asked by crowdworkers across different demographic groups.

6.4 Analysis

Abstractive summarization We examine how workers recruited from India versus those recruited from the United States select summaries offered by different summarization models. We compute mean and standard error of scores obtained from crowdworkers (-1, 0, or 1) over 100 news articles. Our findings show statistically significant differences in scores obtained from these workers across several evaluation criteria across all models (Table 6.1). For instance, while crowdworkers in India preferred summaries generated by GPT-2 over summaries generated by

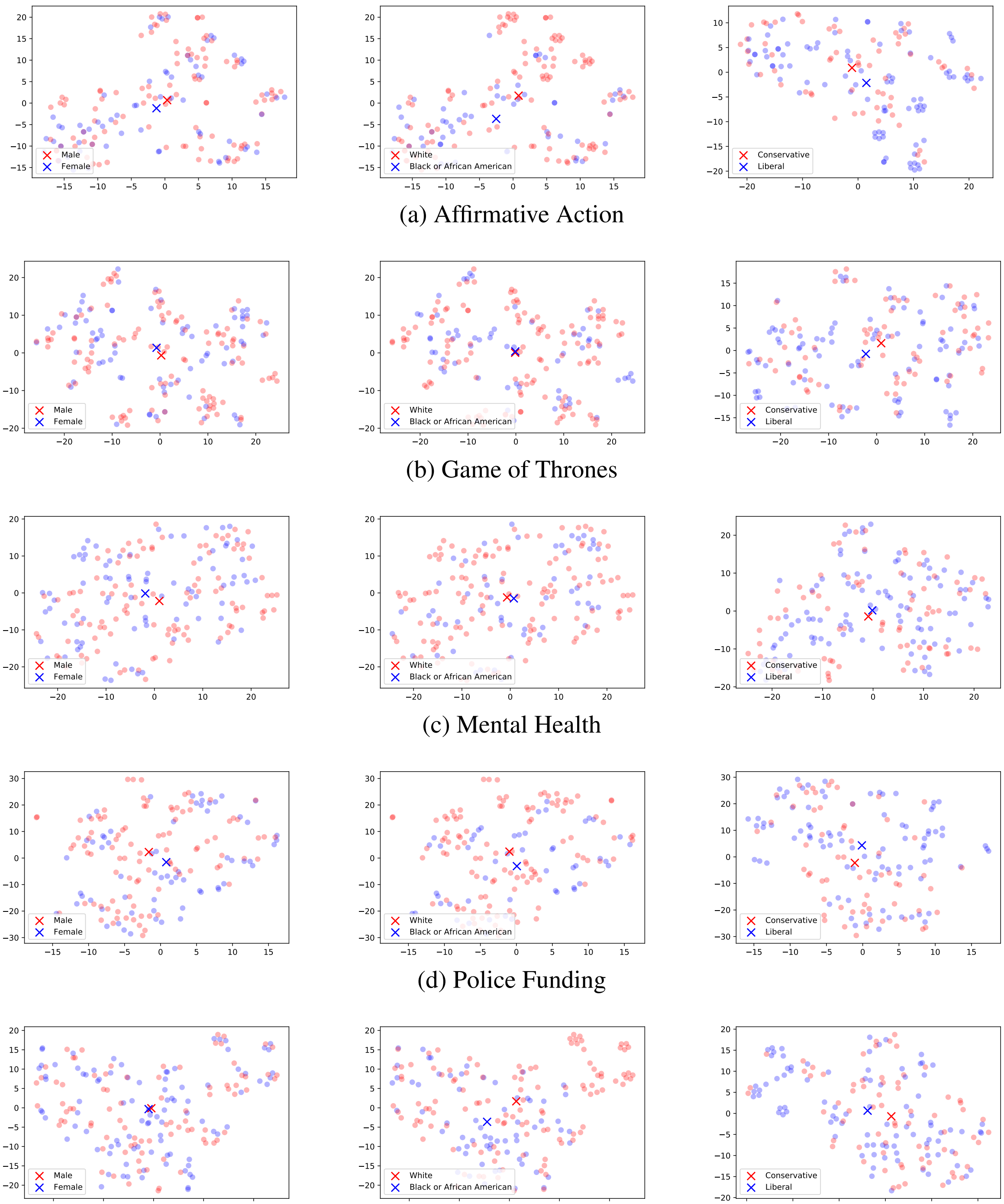
Country	Clarity	Coherence	Consistency	Fluency	Redundancy	Relevance	Overall
			T5 versus GPT2				
India	$-0.04_{0.03}$	$\mathbf{-0.39_{0.03}}$	$\mathbf{-0.72_{0.02}}$	$\mathbf{-0.25_{0.03}}$	$-0.03_{0.03}$	$0.03_{0.03}$	$-0.36_{0.03}$
US	$-0.06_{0.03}$	$\mathbf{-0.24_{0.03}}$	$\mathbf{-0.65_{0.02}}$	$\mathbf{-0.17_{0.03}}$	$-0.04_{0.03}$	$0.07_{0.03}$	$-0.29_{0.03}$
			T5 versus BART				
India	$\mathbf{-0.14_{0.03}}$	$\mathbf{-0.27_{0.03}}$	$\mathbf{-0.57_{0.02}}$	$\mathbf{-0.23_{0.03}}$	$\mathbf{-0.11_{0.03}}$	$\mathbf{-0.07_{0.03}}$	$\mathbf{-0.33_{0.03}}$
US	$\mathbf{0.05_{0.03}}$	$\mathbf{-0.15_{0.03}}$	$\mathbf{-0.40_{0.03}}$	$\mathbf{-0.09_{0.03}}$	$\mathbf{0.01_{0.03}}$	$\mathbf{0.09_{0.03}}$	$\mathbf{-0.16_{0.03}}$
			GPT2 versus BART				
India	$\mathbf{-0.03_{0.03}}$	$\mathbf{-0.25_{0.03}}$	$\mathbf{-0.49_{0.03}}$	$\mathbf{-0.29_{0.03}}$	$\mathbf{-0.07_{0.03}}$	$0.01_{0.03}$	$\mathbf{-0.27_{0.03}}$
US	$\mathbf{0.07_{0.03}}$	$\mathbf{-0.01_{0.03}}$	$\mathbf{-0.32_{0.03}}$	$\mathbf{0.00_{0.03}}$	$\mathbf{0.00_{0.03}}$	$0.16_{0.03}$	$\mathbf{0.03_{0.03}}$

Table 6.1: Scores given by workers in India and US on several criteria for a summarization task shown in Fig. A score of -1 means a worker prefers the first model more than the second, zero indicates a tie, and 1 indicates a preference for the second. Mean and standard error (over 100 article-summary pairs) of scores are reported. Results are bolded when the difference in ratings obtained from workers in India versus the United States is statistically significant with $p < 0.05$.

Location ↓	Evaluation parameters			
	Relevance	Fluency	Naturalness	Engaging
	Blender Bot 3B			
India	$\mathbf{3.96_{0.06}}$	$\mathbf{4.12_{0.06}}$	$\mathbf{4.12_{0.06}}$	$4.01_{0.06}$
United States	$\mathbf{4.38_{0.06}}$	$\mathbf{4.51_{0.05}}$	$\mathbf{4.37_{0.06}}$	$4.18_{0.06}$
	Blender Bot 90M			
India	$\mathbf{3.57_{0.07}}$	$3.82_{0.07}$	$\mathbf{3.74_{0.07}}$	$\mathbf{3.80_{0.07}}$
United States	$\mathbf{3.77_{0.07}}$	$3.81_{0.07}$	$\mathbf{3.52_{0.08}}$	$\mathbf{3.60_{0.07}}$
	Reddit 3B			
India	$3.38_{0.07}$	$3.78_{0.08}$	$3.69_{0.07}$	$\mathbf{3.74_{0.08}}$
United States	$3.30_{0.09}$	$3.61_{0.09}$	$3.51_{0.09}$	$\mathbf{3.18_{0.09}}$

Table 6.2: Likert scale ratings provided by crowdworkers following interaction with an open-domain chatbot. Mean and standard error reported over 250 conversations, where each conversation involved a total of 16 dialog utterances (8 human, 8 chatbot). In scenarios where difference between ratings provided by workers in India and those in the United States is statistically significant with $p < 0.05$ have been bolded.

BART, crowdworkers from the United States had opposite preferences on almost all evaluation criteria. Similarly, crowdworkers from India preferred T5 summaries over BART summaries across all criteria, far more than US workers. Overall, it appears workers in India rank the models as T5>GPT2>BART whereas workers in the United States find GPT2 and BART to be



(a) Affirmative Action

(b) Game of Thrones

(c) Mental Health

(d) Police Funding

(e) Vaccines

Figure 6.4: TSNE Plots of question representation (randomly selected five topics).

mostly tied.

Open domain dialog Crowdworkers in this task were asked to rate the conversation from 1–5 on several attributes. We compute mean and standard error across ratings offered by crowdworkers from India versus the United States (250 dialogs each) and found statistically significant differences across both versions of BlenderBot (Table 6.2). Crowdworkers in the United States found BlenderBot 3B to be more relevant, fluent, engaging, and natural than crowdworkers in India did. On the other hand, crowdworkers in India found BlenderBot 90M to be more natural and engaging than the workers in US did. Interestingly, for Reddit 3B, the differences were only on how engaging the crowdworkers thought it was, with crowdworkers from India finding it more engaging than crowdworkers from the US and we see no differences on relevance, fluency, or naturalness.

Question Generation From the 61000 questions, we subsample a set of 9205 questions. These 9205 questions allow us to sample 3000 questions for each demographic group we wish to study. To observe the differences in questions written by crowdworkers from different demographic groups for each topic, we compute embeddings for these questions using the text-davinci-002 model from OpenAI (the closest literature reference is InstructGPT). We then plot the embeddings for each topic using T-SNE to visualize these differences (Figure 6.4). Similar to the previous two setups, we find that the feedback offered by different demographic groups can be very different and not explicitly considering demographics of the cohort could inadvertently exclude those perspectives. For instance, while there is hardly any difference between representations of questions for the topic vaccines by those identifying and males and females, there appears to be a considerable difference across race/ethnicity, and political leaning. And while there appear to be hardly any differences on the topic of mental health or game of thrones, it is less of a case for the topic of affirmative action. We saw crowdworkers who self-identified as Female asked a lot more personal questions on the topic of abortion such as *"Will abortion affect my ability to get pregnant in the future?"*, *"Does having an abortion increase my chance of getting breast cancer?"*, and *"How far in the pregnancy can I get an abortion?"*, whereas those who self-identified as Male asked more questions about legality and their rights, such as *"What can I do if I change my mind, or wish to rearrange my wife's abortion appointment?"*, *"What is the relationship between abortion access and crime rates?"*, *"Why does the government fund planned parenthood when a lot of taxpayers are against abortion?"* and *"How will reduced access to abortion in the US affect other areas of the globe?"*.

6.5 Discussion

In this chapter we examine the idea that when seeking human in the loop feedback for ML models (such as picking one of two plausible summaries or rating a conversation), crowdworkers of different demographics could offer different perspectives. For instance, while crowdworkers in India preferred summaries generated by GPT-2 over summaries generated by BART, crowdworkers from the United States had opposite preferences for almost all attributes. Similarly, crowdworkers from India preferred T5 summaries over BART summaries across all criteria, far more than US workers. For the question generation task, we observe from both embedding analysis and additional qualitative analysis that questions written by workers from one demographic express different perspectives than others.

There are several potential implications of these findings. First, if machine learning models are trained on feedback that is not representative of the population, they may be biased. For instance, if a model is trained on feedback obtained from workers who are mostly college educated, it may be less accurate for those without a college education. Second, if different demographic groups offer different kinds of feedback, this could impact the accuracy of the models. For example, if US crowdworkers tend to rate summaries generated by GPT-2 as equally good as generated by BART, but Indian crowdworkers have the opposite preference, using feedback from US workers to train a summarization model could lead to summaries that are less acceptable to people in India. These findings suggest that it is important to consider the diversity of the workforce when collecting feedback for machine learning models. This is especially important when collecting feedback at scale, as it can help to ensure that the feedback is representative of the population and ties closely to this dissertation's core thesis of the importance of human feedback in ensuring NLP models are reliable when deployed in the real world. One shortcoming of our analysis is that due to the same amount of money paid to workers in India and workers in the US, it could be that the differences in ratings are a reflection of the increased compensation for workers in India (adjusted for purchasing power) than for workers in the US, rather than qualitative differences between populations. Additioanlly, while these findings offer evidence in support of recruiting a more diverse workforce, one key limitation of this analysis is the lack of any results showing differences on the downstream performance of an ML model trained with feedback from a diverse group of crowdworkers versus a model trained with feedback from a rather homogeneous group of crowdworkers, and is left as future work.

Chapter 7

Resolving The Human Subjects Status Of Machine Learning's Crowdworkers

7.1 Overview

As the focus of machine learning (ML)—and, in particular, natural language processing (NLP)—has shifted towards settings characterized by massive datasets, researchers have become reliant on crowdsourcing platforms [Kovashka et al., 2016, Vaughan, 2017, Sheng and Zhang, 2019, Drutsa et al., 2021]. These practices have produced hundreds of new datasets. In NLP, for the task of passage-based question answering (QA) alone, over 15 new datasets containing at least $50k$ annotations have been introduced since 2016. Prior to 2016, the available QA datasets contained at most an order of magnitude fewer human-annotated examples. The ability to construct such enormous resources derives, in large part, from the liquid market for temporary labor enabled by crowdsourcing platforms, including Amazon Mechanical Turk, Upwork, Appen, and Prolific. Over time, the relationship between the ML community and crowdworkers has evolved to encompass a wide variety of tasks and interaction mechanisms. However, the positive view of crowdsourcing as a means to produce *better* and *larger* datasets, potentially leading to technological breakthroughs, has been offset by growing concerns about the ethical and social dimensions of these one-off engagements with crowdworkers. Points of concern include (i) the low wages received by crowdworkers [Fort et al., 2011, Whiting et al., 2019, Silberman et al., 2018, Kummerfeld, 2021]; (ii) disparate access, benefits, and harms of developed applications [Adelani et al., 2021, Nekoto et al., 2020, Orife et al., 2020, Bender and Friedman, 2018, Kiritchenko and Mohammad, 2018, Rudinger et al., 2018, Bender et al., 2021, Strubell et al., 2020]; (iii) the reproducibility of proposed methods [Dodge et al., 2019, Ning et al., 2020, Freitag et al., 2021, Card et al., 2020]; and (iv) concerns about fairness and discrimination arising in the resulting technologies [Hovy and Spruit, 2016, Leidner and Plachouras, 2017, Bender et al., 2020, Blodgett et al., 2020].

Our focus here is on what ethical framework should govern the interaction of ML researchers and crowdworkers, and the unique challenges posed by ML research to regulators. While researchers in fields like NLP typically lack expertise in human subjects research, they nevertheless require practical guidance for how to classify the role played by crowdworkers in their research so that they can comply with relevant ethical and oversight requirements. Unfortunately, clear guidance is presently lacking. Reflecting the current state of confusion, some institutions and a recent paper by Shmueli et al. [2021] suggest that all ML crowdwork constitutes human subjects research, while other institutions suggest that ML crowdworkers rarely constitute human subjects Ipeirotis.

In this chapter, we address the source of confusion, arguing that difficulties in resolving the appropriate designation of ML's crowdworkers owe to several formidable challenges:

Novel relationships The ethical framework that oversight boards use to identify human subjects—the U.S. Common Rule—was developed in the wake of abuses in biomedical and behavioral research. This framework was especially influenced by dynamics in biomedical research, including the need to distinguish clinical research from medical practice London [2021]. Binning activities into these categories facilitated the goal of ensuring that these distinct relationships were governed by the relevant set of norms—the norms of clinical medicine or the norms of medical research. Because the distinction between employees on a research team and study participants is less ambiguous in medical contexts, little attention has been paid to criteria for distinguishing *research staff* from study participants.

Novel methods Compared to biomedical or social sciences, where data are collected to answer questions that have been specified in advance, ML research often involves a more dynamic workflow in which data are collected in an open-ended fashion and research questions are articulated in light of data and its analysis. Additionally, while it is typical in biomedicine for teams that gather data to analyze it, or for researchers to analyze data that was first gathered for clinical purposes, in ML research there can be a more distributed division of labor with some research groups collecting data that will serve as the foundation for future studies by a whole community of researchers.

Ambiguity Under the Common Rule, whether an individual is a human subject hinges on whether the data collected, and later analyzed, is *about* that individual. However, as Shmueli et al. [2021] have noted, crowdworkers can fill such diverse roles in ML research (even within a single study) that is becomes difficult to draw a line between which data is collected *about* the crowdworker versus merely *from* them (but about something else) Shmueli et al. [2021].

Inexperience Despite the enormous productivity in this area, crowdsourcing-intensive NLP papers seldom discuss the ethical considerations that would otherwise be central to human subjects research and rarely discuss whether an Institutional Review Board (IRB) approval or exemption was sought prior to the study—only 14 ($\approx 2\%$) of the aforementioned 703 papers described IRB review or exemption Shmueli et al. [2021][1];

[1] It is worth noting that in other computing fields such as human computer interaction, it is common practice to seek IRB review prior to collecting data from human annotators. Additionally, not all of these 703 papers came from Common Rule institutions, so the actual percentage could be higher than 2%.

Scale Currently NLP research is producing hundreds of crowdsourcing papers per year, with 703 appearing at the top venues (ACL, NAACL, EMNLP) alone from $2015 - -2020$ Shmueli et al. [2021].

Moreover, we argue that these challenges not only create confusion among stakeholders, they also open the potential for loopholes, whereby researchers can avoid IRB oversight without altering the substantive research procedures performed on participants London et al. [2020]. In particular, a single study that would be considered human subjects research could be split into two parts: one in which researchers collect data about workers and release an anonymized version to the public without analyzing information about the workers themselves; and a second in which they or another team of researchers perform analysis on information about the workers. According to some institutional policies, the latter two studies might not require research ethics oversight whereas the single study would.

To ensure that ML research is conducted according to the appropriate ethical and regulatory standards, greater clarity is required. In Section 7.2, we elaborate the criteria that define human subjects for ethical and regulatory purposes in the United States. We briefly discuss the relationship between the question of whether one or more persons satisfy these criteria and the question of whether that research must undergo review by a properly constituted IRB. In Section 7.3, we present prototypical examples from research in NLP to identify paradigmatic cases for which it is clear/unclear how a given crowdworker should be classified. We then show how the diversity of roles that crowdworkers can play in ML research poses a challenge for research ethics and provide guidance on interpreting the Common Rule to identify when crowdworkers should be classified as human subjects versus as extensions of the research team for both ethical and regulatory purposes. Finally, in Section 7.4, we offer policy solutions to address these concerns.

7.2 Regulatory Framework

In the United States, the regulations that govern the use of human participants in scientific research are set out in the Code of Federal Regulations (CFR) and are commonly referred to as the Common Rule. These regulations are promulgated by the Executive Branch and apply only to institutions that accept federal funds or that have agreed to abide by these rules. Nevertheless, the language and the requirements laid out in these rules have been adopted by, and exert a great deal of influence within, the larger literature on research ethics.

Because the Common Rule only applies to research with human participants, it sets out two important criteria to determine whether a person constitutes a research participant: those used to define *research* and those used to define a *human subject*.

First, in order to be a participant in research, research must be taking place. Research is defined, in part, as “a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge.” Second, human subjects are then defined as follows:

(e)(1) Human subject means a living individual about whom an investigator (whether professional or student) conducting research:

(i) Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or

	Studies/Analyzes	Uses
Intervention	Identifying better crowdsourcing strategies via a randomized study	Train an ML model on data collected in a gamification environment
Interaction	Analyzing data collected via surveys on Mechanical Turk	Asking crowd to annotate a dataset to train ML models

Table 7.1: Examples of research interactions with the crowd.

(ii) Obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens. (45 CFR 46.102 (e)(1))

For simplicity, we limit our discussion to the production of information, rather than to a discussion of specimens.

Two points of clarification are in order. First, we note that in (e)(1), the definition of a human subject requires that researchers obtain information *about* the individual in question. This does not imply that the researcher is conducting research about the individual, per se, since research aims to produce generalizable knowledge. In the biomedical context, for example, a study might seek to determine the effect of some intervention on blood pressure among the population of individuals who suffer from a particular disease. To answer this question, researchers might measure the blood pressure of specific individuals with that disease. That information is then analyzed to produce generalizable knowledge pertaining to the underlying population. However, as we will see, delineating precisely which information is *about* an individual can be difficult in many settings where crowdworkers are engaged by ML researchers. Second, conditions (i) and (ii) lump together a range of cases that vary in substantive ways. Condition (i) is a combination of two conjuncts. The first conjunct concerns the way that information is produced: information can arise from intervention or from interaction. These terms are defined respectively as:

(2) Intervention includes both physical procedures by which information or biospecimens are gathered (e.g., venipuncture) and manipulations of the subject or the subject's environment that are performed for research purposes.

(3) Interaction includes communication or interpersonal contact between investigator and subject.

Of these possibilities, interaction is the weaker condition. Interventions can reasonably be understood as the subset of interactions that produce a change in either the individual (e.g., administering a drug, or drawing blood) or their environment (e.g., placing an individual in an imaging device). By contrast, interactions include communication or interpersonal contact that generate information without necessarily bringing about a change to the individual or their environment. For example, a study might involve randomizing a group of participants to receive either an investigational intervention in addition to usual care, or to receive only usual care. Although the former group receives an intervention—something they would not have received outside of the context of research—the latter group is not subject to an intervention. Nevertheless, their inclusion in a group that is randomized within a study constitutes a form of social interaction necessary to generate data that controls for confounding, and so helps to produce generalizable knowledge.

The second conjunct in condition (i) requires that information that arises in one of these two ways is then used, studied, or analyzed. Of these, *use* is the broadest category, as there may be myriad ways information from a social interaction might be used in the course of research. In contrast, study and analysis seem to constitute a strict subset of uses in which data are analyzed or evaluated, presumably to generate the generalizable knowledge that defines the study in question. Table 7.1 provides a representation of the combinations of views that result from combining these modes of interaction and types of use. Among these, the *intervention analysis* condition is the most narrow and captures a paradigm of the researcher-participant relationship. Namely, a person is a human subject if, in the course of research, they are the target of an intervention from which a researcher generates information that is then the subject of an analysis that is intended to produce the generalizable information that is the focus of the research. In contrast, the *interaction use* criteria are weaker, holding that a person is human subject if, in the course of research, researchers interact with them in a way that produces information that is used to further the goals of research.

Condition (ii) deals with cases in which researchers obtain, use, study, analyze or generate private information about a living individual. This condition is intended to cover cases in which researchers might not interact with living persons, in the sense outlined in condition (i), but they nevertheless use or generate private information about a living individual in the course of their research. This condition therefore applies to research involving datasets that include private information about living individuals or to research that would generate that information from datasets that might not include private information about living individuals taken on their own.

These definitions play a key role in demarcating which set of ethical and regulatory requirements apply to an activity. A research activity that does not involve human subjects does not fall under the purview of the regulations governing research with human subjects. Consequently, if there are no human subjects in a study then the study does not need to be reviewed by an IRB. In contrast, if a researcher is carrying out research with human participants, then that researcher incurs certain moral and regulatory responsibilities. Among these regulatory responsibilities is the duty to present one's research for review by an IRB.

This last claim might come as a surprise to some who read the Common Rule, since a significant portion of ML research, and NLP research in particular, is likely to be classified as *exempt*. Per 46.104.(3)(i), research involving benign behavioral interventions in conjunction with the collection of information from an adult subject through verbal or written responses (including data entry) or audiovisual recording can qualify for *exempt* status if the subject prospectively agrees to the intervention and information collection and at least one of the following criteria is met:

(A) The information obtained is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subjects;

(B) Any disclosure of the human subjects' responses outside the research would not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, educational advancement, or reputation.

However, a researcher cannot unilaterally declare their research to be *exempt* from IRB review. Rather, *exempt* is a regulatory category whose status must be certified by an IRB (§46.109.(a)). This can seem paradoxical to some since, in order to qualify for an exemption, researchers must submit sufficient information about their research to the IRB so that the latter can determine that these, or other applicable criteria (as laid out in the Common Rule) are met.[2] Nevertheless, the work required to secure this certification is usually less than is required to submit a full protocol and the certification is usually granted in less time than it would take for an IRB to provide a review of that protocol involving the full IRB board. For present purposes, the main point is that if a researcher at an institution bound by the Common Rule carries out human subjects research without first having that research reviewed by the relevant IRB, then that researcher would be in violation of that institution's regulatory obligations, even if that research would have qualified for an exemption.

7.3 Common Rule and ML Research

Based on the preceding analysis, we can now identify a large subset of ML research in which crowdworkers are clearly human subjects. These cases fit squarely into the paradigm of research, familiar in biomedicine and social science, where researchers interact with crowdworkers to produce data *about* those individuals, and then analyze that data to produce generalizable knowledge about a population from which those individuals are considered to be representative samples.

First, we consider studies where researchers assign crowdworkers at random to interventions in order to produce data that can be analyzed to generate generalizable knowledge about best practices for utilizing crowdworkers. Here, the crowdworkers are clearly human subjects. They are the target of an intervention that was designed for the specific purpose of capturing data about them (namely, their performance at some task), that could then be analyzed qualitatively and statistically to address the central hypotheses of the study.

For instance, consider Khashabi et al. [2020], who engage crowdworkers to investigate which workflows result in higher quality question-answering datasets. They recruit one set of crowdworkers to write questions given a passage, while another group of crowdworkers are shown a passage along with a suggested question and are tasked with minimally editing this question to generate new questions. In these settings the data is about the workers themselves, as is the analysis. Investigating adversarial setups for generating question answering datasets, Kaushik et al. [2021a] conduct a large-scale controlled study focused on a question answering task. One set of workers is asked to write five questions after reading a passage, highlighting the answers to each, and are awarded a base pay of $\$0.15$ per question. Another set of crowdworkers is shown the same passages but asked to write questions that elicit incorrect predictions from an ML model trained using a different dataset to perform passage based question answering. To incentivize workers to spend more time thinking about ways to fool this existing model, workers are paid $\$0.15$ for each question that fools the model in addition to the base pay of $\$0.15$ per question. The research team later analyzed this data to identify the differences between the questions generated by both sets of workers and derive insights about how each data creation setup influences crowdworker behavior. They also trained various machine learning models on these datasets and evaluated them on several other question answering datasets to establish which

[2] This is a commonality in administrative rulemaking as well as judicial review. After all, Courts get to determine whether something is in their jurisdiction but a plaintiff has to provide information to enable a court to make that determination.

interaction mechanism produced better data (as measured by performance of models trained on the respective datasets), producing generalizable knowledge to aid future data collection efforts.

Humans subjects research in NLP is not limited to studies aimed at dataset quality. Hayati et al. [2020] paired two crowdworkers in a conversational setting and asked one crowdworker to recommend a movie to the other. They then study the resulting data to identify what recommendation and communication strategies lead to successful recommendations, and use these insights to train automated dialog systems. In another work, Pérez-Rosas and Mihalcea [2015] asked crowdworkers to each write seven truths and seven plausible lies on topics of their own choice. The authors also collected demographic attributes (such as age and gender) for each crowdworker. They then analyzed how attributes of deceptive behavior relate to gender and age. They also train classifiers using this data to predict deception, gender, and age. In these cases, the researchers interacted with crowdworkers to produce data about the crowdworkers that was then analyzed to answer research hypotheses, creating generalizable knowledge.

7.3.1 Cases Where the Human Subjects Designation is Problematic

Unlike the above, many ML crowdsourcing studies do not fit neatly within the paradigm of research that is common in biomedicine and the social sciences. For example, crowdworkers are often brought in, not as objects of study, but to perform tasks that could have been—and sometimes are—performed by members of the research team themselves. Note that in these cases, members of the research team certainly do interact with crowdworkers and that those interactions produce data that in some meaningful sense is used to produce generalizable knowledge. Moreover some of the collected data certainly is *about* the worker e.g., for the purpose of facilitating payment. However, in these cases, the data that is analyzed in order to produce generalizable knowledge are not about the crowdworkers in any meaningful sense.

In perhaps the most common category of crowdsourcing study in machine learning, researchers hire workers to label a training dataset that will be used for training ML models. For instance, Hovy et al. [2014] recruit crowdworkers to annotate parts of speech in text. They then train machine learning models on this data to predict parts of speech on test set. In another study, Taboada et al. [2011] recruit crowdworkers to create a collection of words associated with a sentiment label which is then used to produce a sentiment classification model. Countless such datasets are introduced every year. Often researchers interact with the crowdworkers and use the data generated as a result of that interaction. While it might appear that any such research satisfies the *interaction* + *use* criteria from the Common Rule, the subtle distinction is that the information used to produce generalizable knowledge is not *about* the worker.

In many of these cases, crowdworkers are performing tasks that are routinely performed by research team members themselves when working data on smaller scales. For example, Kovashka et al. [2016] describe numerous computer vision papers where researchers provide their own labels. In another example, NLP researchers often ask crowdworkers to not only provide the correct label for a document, but also to highlight *rationales*, contiguous segments in the text that provide supporting evidence. Notably, while DeYoung et al. [2020] recruit crowdworkers to annotate rationales for various classification tasks, Zaidan et al. [2007] opt to annotate the rationales themselves. In another setting, Kaushik et al. [2020] recruit crowdworkers, who given a text and associated label, were tasked to minimally edit the text to make a counterfactual label applicable. In a followup study, instead of recruiting crowdworkers, Gardner et al. [2020] opt to make these edits themselves.

How should crowdworkers in these cases be classified? On a strict reading of the claim that a human subject is a living individual "about whom" researchers obtain information that is used or analyzed to produce generalizable knowledge, then crowdworkers in these cases would not be classified as human subjects. This reading is consistent with the practice of some IRBs. For example, Whittier College's IRB states:[3]

> Information-gathering interviews with questions that focus on things, products, or policies rather than people or their thoughts about themselves may not meet the definition of human subjects research. Example: interviewing students about campus cafeteria menus or managers about travel reimbursement policy.

In contrast, other IRBs adopt a far more expansive reading of the language in the Common Rule. For instance, Loyola University's IRB says:[4]

> In making a determination about whether an activity constitutes research involving human subjects, ask yourself the following questions:
> *1) Will the data collected be publicly presented or published?*
> AND
> *2) Do my research methods involve a) direct and/or indirect interaction with participants via interviews, assessments, surveys, or observations, or b) access to identifiable private information about individuals, e.g., information that is not in the public domain?*
> If the answer to both these questions is "yes", a project is considered research with human subjects and is subject to federal regulations."

Note that this interpretation does not distinguish whether the information is about an individual or just obtained via a direct and/or indirect interaction. This view appears to be shared by other IRBs as well.[5]

How does information about versus merely from impact human subjects determination? Traditionally, research ethics has not had to worry about who is a member of the research team and who is a participant in that research. This ambiguity arises in cases of self-experimentation, but such cases are relatively rare and fit squarely into the *intervention* + *analysis* category from the Common Rule. The scope of the effort required to produce data that can be used in ML research has engendered new forms of interaction between researchers and the public. Without explicit guidance from federal authorities in the Office of Human Research Protections, individual IRBs will have to grapple with this issue on their own.

Our contention is that in the problematic cases referred to in this section, crowdworkers are best understood as augmenting the labor capacity of researchers rather than participating as human subjects in that research. This argument has two parts.

[3] Archived on February 14, 2022. https://web.archive.org/web/20220214194123/ https://www.whittier.edu/academics/researchethics/irb/need

[4] Archived on February 14, 2022. https://web.archive.org/web/20220214194036/ https://www.luc.edu/irb/irb_II.shtml

[5] Archived on February 27, 2022. https://web.archive.org/web/20220228012326/ https://www.bsc.edu/academics/irb/documents/BSC%20IRB%20Decision%20Tree.pdf

The first part is an argument from symmetry. Within a division of labor, if more than one person can carry out a portion of that division of labor, then the way that we categorize the activity in question should depend on substantive features of that activity rather than on the identity of the individual in question.[6] From this, it follows that if a task is performed by a researcher in one instance and then by one or more crowdworkers in a second instance, then our categorization of that activity should be the same in both cases. The argument from symmetry alone entails only that either the crowdworker and the researcher are both part of the research team or both human subjects.

The second part of the argument appeals to three additional considerations that support classifying both parties as part of the research team. First, when researchers perform the tasks in question it seems clear that they are not self-experimenting—they are not subjects in their own study. Second, this impression or intuition is explained by the fact that these interactions produce information that contributes to the production of generalizable knowledge, but that this information is better classified as coming from, rather than being about, these individuals. Researchers interact with other members of their team to produce information and this information is used in research, but this use involves creating or refining the instruments, materials, metrics, and other means necessary to carry out research. Its purpose is to create the means of generating new knowledge rather than to constitute that data or evidence base whose study or analysis will generate this new knowledge. Third, ignoring the distinction between data that is about a person rather than merely from them, and holding that both researchers and crowdworkers are human subjects in these cases, creates a regulatory category so broad that it would class members of every research team, including those in traditional biomedical and social science, as as human subjects. The reason is simply that those researchers routinely interact with other members of their team to create information that is used to produce generalizable knowledge. But this consequence is absurd.

7.3.2 Loopholes in Research Oversight

The analysis in the previous section illustrates one challenge that ML research poses for research ethics. Part of the ethical rationale for the oversight of research with human participants is that the interests of study participants can be put at risk when researchers interact with or intervene upon them for the purpose of generating generalizable knowledge. These risks can derive from the nature of the interaction or intervention, or from the use that is made of the resulting information. A loophole in research oversight has been defined "as the unilateral ability of a researcher to avoid an oversight requirement without altering the substantive research procedures performed on participants" London et al. [2020]. Loopholes in research oversight are morally troubling, in part, because they violate a concern about equal treatment for like cases: if researchers interact with individuals for the purpose of generating data that is about those individuals and generalizable knowledge is produced from the study or analysis of that data, then the interests of those

[6] One might argue that the way we treat unionized vs non-unionized workers or independent contractors vs employees are counterexamples where the work might be exactly the same but the identity of the individual and a feature about them makes a difference regarding workplace protections amongst other things. In these cases, although, prior agreements might shape the entitlements of agents, they do not alter the classification of the activity performed i.e., whether the task is work or research.

individuals should receive the same level of oversight regardless of how the labor in this process is organized. However, two features of ML research make the Common Rule particularly prone to loopholes: the way that labor is divided between the collection and the analysis of data and the way that research questions often arise after data collection.

Scenario 1 It is clear that the Common Rule envisions several ways in which labor might be divided in research. First, in traditional biomedical or social science research it is common for the same individuals who collect data to also analyze that data in the course of their research. This division of labor is presupposed in 45 CFR 46.102 (e)(1)(i) which says that when a researcher conducting research "[o]btains information or biospecimens through intervention or interaction with the individual, *and* uses, studies, or analyzes the information or biospecimens", that research activity would be categorized as human subjects research. In this scenario, ethics review covers two morally weighty aspects of this division of labor: whether researchers interact with participants in ways that respect their autonomy and safeguard their welfare and whether they use the information that they obtain from these interactions in a way that respects the rights and welfare of the people this information is about.

Second, it is common for data or biospecimens to be generated in the course of the provision of medical care or other health services. In these cases the interactions of medical professionals with patients are not shaped around research purposes—they are shaped by the goals and purposes of the provision of health care or other medical services. As such, those interactions are usually governed by the norms of medical or professional ethics. Research ethics review thus focuses on whether the data or specimens in question constitute or include identifiable private information and, if so, whether research with this information respects the rights and welfare of the individuals from whom the information was gathered.

It is not clear that the Common Rule envisioned a division of labor in which researchers would interact with individuals for research purposes (i.e., where the interactions are shaped by the goal of generating generalizable knowledge rather than the provision of health services) but those researchers would not use, study or analyze that information themselves. To be clear, this is different from secondary use of data that was gathered for research purposes since, in traditional biomedical or behavioral research, the initial research would likely have been subject to research oversight. That oversight would ensure that researchers interact with participants on terms that respect their rights and welfare and subsequent oversight would evaluate additional uses of that data.

In contrast, it is common for ML researchers to collect large datasets in an open-ended manner before hypotheses are formulated, often with the goal of facilitating a range of future research in broad topic areas Williams et al. [2018a], Zhang et al. [2021], Aggarwal et al. [2021], Ao et al. [2021], Le et al. [2021], Zang et al. [2021]. For example, Williams et al. [2018a] collect a large scale corpus for the task of recognizing textual entailment. They train an ML model on this dataset and release the dataset with anonymized crowdworker identifiers for future research. Similarly, Mihaylov et al. [2018] and [Talmor et al., 2019] collect large scale question answering datasets created by crowdworkers, train ML models on this data, and release these datasets with anonymized crowdworker identifiers for future research. Since these studies only involved interacting with crowdworkers, and using or analyzing data *from* crowdworkers, they may not require IRB review. Subsequently, Geva et al. [2019] took these anonymized datasets and analyzed information *about* the crowdworkers. Specifically, they looked at how ML models trained on data created by one set of crowdworkers do not generalize to data created by a disjoint set of crowdworkers. They further train ML models, which given a document as input, predict which crowdworker wrote that document. Since Geva et al. [2019] did not interact with the crowdworkers, and only analyzed existing (anonymized) datasets, their studies also may not require IRB review. However, had the researchers who collected these datasets also analyzed this information, that study would have required IRB review. As part of this review, an IRB would not only perform oversight over the questions asked, but also how the researchers interact with the crowd and whether adequate protections are in place for crowdworkers participating in these studies.

Although a significant portion of ML research poses few risks to participants, there are cases where interactions or interventions are less benign, as when researchers ask crowdworkers to write toxic comments. For example, Xu et al. [2020] recruit crowdworkers to interact with an automated chatbot with the aim of eliciting *unsafe* responses from the chatbot, using this data to train models that are better at generating *safe* responses. Crowdworkers may not be human subjects in this case, insofar as the information they provide is not about them in the relevant sense. However, in this example, the research team also created a taxonomy of offensive language types to classify human utterances citing potential use for this taxonomy in future research. From this larger data set inferences could be drawn about the proclivities to, or proficiency of, particular crowdworkers using offensive language of particular types.

In each of these cases, datasets are collected which contain information that is from crowdworkers for the purposes of producing generalizable knowledge that can include information that is about the crowdworkers. A loophole in research oversight is created because 45 CFR 46.102 (e)(1)(i) holds that individuals participating in a study are considered human subjects if researchers both obtain *and* use, study or analyze that information in a single study. To be clear, releasing such a dataset with identifiable private information for research purposes would fall under clause (ii) from 45 CFR 46.102(e)(1) (discussed in Section 7.2). Once the dataset has been created, then using it for research purposes would fall under this same clause, as long as the identifiable private information remains.

A division of labor in which one set of researchers interact with individuals specifically for the purpose of generating data necessary to produce generalizable knowledge and then release that data (with anonymized identifiers) so that another set of researchers can analyze it represents a loophole because, unlike the secondary use of data from ordinary clinical practice, this data is produced by researchers who interact with individuals for research purposes–to produce data that will help to create generalizable knowledge. But, unlike the case where the researchers themselves analyze this data, this research activity would not be subject to oversight or review aimed at providing credible social assurance that those interactions respect individual autonomy and welfare [London, 2021]. Anonymizing the data that is produced helps to shield individuals from harm that results from the way that information is used, such as uses that expose sensitive personal information. But whether the means used to gather that data respect the autonomy and wellbeing of those individuals is not subject to oversight or review.

As a result, one way to address loopholes of this type would be to amend 45 CFR 46.102 (e)(1)(i) to explicitly include the **release** of data alongside its use, study or analysis.

Scenario 2 Amending 45 CFR 46.102 (e)(1)(i) to include the release of data may not be sufficient to foreclose a second scenario in which loopholes might arise. Consider a scenario in which a research team interacts with crowdworkers to collect some data from and some that is about them and then proceeds to analyze both sets of data. This single protocol fits the mold of traditional research in biomedicine or the social sciences and so would constitute research with human subjects. Now consider a scenario in which the research team distributes this work over two separate protocols. In the first protocol they propose to gather data that is both from and about crowd workers but only use data that is from them in their analyses. This study might not require IRB approval because it does not analyze, study or use data that is *about* crowdworkers. The researchers then anonymize the full dataset and submit a second protocol in which they analyze the now-anonymized data to answer questions about the crowd workers. The second study might not require IRB approval because it does not involve obtaining information via any interaction with living individuals and it does not involve generating or using any identifiable private information.

In this scenario, a single study that would require IRB approval could be decomposed into separate studies that involve the same interactions or interventions on crowdworkers in order to answer the same set of hypotheses but in a way that avoids research ethics oversight. Because the researchers are not releasing their data publicly, the proposal in the previous section would not close this loophole. As a result, the determination of whether an ML project constitutes research with human participants might need to be made at a higher level than the individual study protocol. In the context of drug development, a trial *portfolio* has been defined as a "series of trials that are interrelated by a common set of objectives" London and Kimmelman [2019]. In ML research, the determination of whether an activity constitutes research with human participants may need to be made at the portfolio level by considering whether data to be generated and the questions to be investigated across an interrelated set of investigations are *about* the crowdworkers. For portfolio level reviews to succeed, however, researchers would need to identify ex ante the scope and nature of the data they are collecting and the full range of questions they might seek to answer from that data across multiple studies. Given the dynamic nature of ML research and the extent to which research questions are often posed after data has been collected, this may require consultation with IRBs to determine the conditions under which an envisioned portfolio of studies would or would not constitute research with humans and the steps that can be taken ex ante to facilitate the ability of researchers to pursue important questions as they arise.

7.4 Discussion

There is currently considerable confusion about when ML's crowdworkers constitute human subjects for ethical and regulatory purposes. While some sources suggest treating all crowdworkers as human subjects [Shmueli et al., 2021], our analysis makes a more nuanced proposal, identifying: (i) clear-cut cases of human subjects research: these require IRB consultation, even if only to confirm that they belong to an exemption category; (ii) crowdsourcing studies that do not constitute human subjects research because the analyses that produce generalizable information do not involve data *about* the workers; (iii) difficult cases, where the distinctive features of ML's crowdworking studies combine with ambiguities in the Common Rule to create substantial uncertainty about how to apply existing requirements; and (iv) loopholes, whereby researchers can escape the human subjects designation without making substantive changes to the research performed.

Part of the spirit of research oversight is to safeguard the rights and interests of individuals involved in research. In some cases, crowdworkers are the subjects of interventions or interactions that are designed to generate information about them which researchers intend to analyze in order to create generalizable knowledge. In these cases, the task of securing their rights and interests rightfully falls into the domain of human subjects ethics and oversight. But if researchers don't seek to either obtain or use, study, analyze or release information *about* a person (in some meaningful sense), then it is not clear that frameworks for the protection of participants in research with human subjects are applicable or appropriate. Individuals who are not research participants can still be exposed to risks to their well-being and threats to their autonomy. This is true of most social interactions. It is particularly true of employment interactions as employers often have access to sensitive, private, identifiable information (such as Social Security Number, travel records, and background check reports) about their employees. But the solution to ensuring that crowdworkers have credible public assurance that their rights and interests are protected is not to expand the definition of human subjects to include all crowdworkers.

Recommendations:

1. **ML researchers** must work proactively with IRBs to determine which, if any, information they will generate that is *about* versus merely *from* crowdworkers and whether, given the full range of questions they intend to investigate across the portfolio of studies involving this data, the anticipated set of studies constitutes human subject research. They should also recognize that as the questions they investigate change, the status of the research they are conducting may change correspondingly. Researchers should therefore work proactively with their IRB to determine when modifications to ongoing research require a new submission or the submission of a protocol modification for IRB review.
2. **IRBs** should not reflexively classify all ML research involving crowdworkers as human subjects research. At the same time, IRBs should also establish clear procedures for evaluating portfolios of research to address the possibility of loopholes in research oversight. They should also communicate with ML researchers clearly about the conditions under which the classification of research might change and the conditions under which a revised protocol would need to be submitted.
3. **The Office of Human Research Protections (OHRP)** should offer more precise guidance about what it means for information or analysis to be "about" a set of individuals. We also recommend that OHRP should revise the Common Rule so that 45 CFR 46.102(e)(1) condition (i) reads: "Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, analyzes, **or releases** the information or biospecimens." In short, this modification would require that an original investigator who collects data through interaction with humans and plans to release a dataset (even if anonymized) that could be used to ask questions *about* those individuals must secure IRB approval for the research in which those data are gathered. Subsequent studies that draw upon the resulting anonymized public resource would not be marked as human subjects research,

provided that they do not attempt to re-identify the individuals represented in the dataset. This modification would resolve the loophole identified in this paper. OHRP also has a role to play in offering guidance to ML researchers, which could be achieved by issuing an agency Dear Colleague letter or an FAQ document.

Chapter 8

Conclusion

To facilitate training of NLP models that can generalize well under plausible distribution shifts, this thesis put forth methods, datasets, and analysis of the benefits offered by several forms of human in the loop feedback. We discussed in detail three specific forms of feedback—counterfactually augmented data, feature feedback, and adversarial data collection—and demonstrate the promise of leveraging human-in-the-loop feedback to disentangle the spurious and non-spurious associations, yielding classifiers that hold up better when spurious associations do not transport out of domain. A large part of the thesis focused on counterfactually augmented data, where we present evidence that by leveraging humans not only to provide labels but also to intervene upon the data, revising documents to accord with various labels, we can elucidate the difference that makes a difference. Moreover, we can leverage the augmented data to train classifiers less dependent on spurious associations. Additionally, through our discussions on diversity impacts, we highlighted another important aspect that must be considered when designing an NLP workflow involving crowd labor—reflecting thoughtful benefit and risk assessments. Finally, although the US Common Rule provides stringent protection of human subjects in research settings involving even small amounts of money or personal information disclosure, such untailored regulations are detrimental for certain fields like natural language processing due to the wide scale use of anonymous participant contributions over digital interfaces. Thus, special considerations need to be taken when framing ethical standards for studies deploying crowdsourcing for value extraction from humans' linguistic production.

In sum, this dissertation has advanced our understanding of how human interactions with machine learning systems can be leveraged to increase model efficacy for natural language processing, across varied datasets and application contexts, and makes a compelling case for the effectiveness and potential of employing human-in-the-loop feedback in NLP models to facilitate better performances under distribution shift. Drawing from this dissertation, researchers are better equipped to understand more precisely how humans can provide critical feedback that enables better performing NLP systems while being mindful of ethical concerns pertaining to the autonomy of research participants, data privacy, and staying apprised of newer interpretations of regulations regarding human subjects research. These considerations will both ensure fair use of data within society and promote greater understanding between human view points and predictive models. By looking into counterfactually augmented data, feature feedback, and adversarial collection of data, we have demonstrated the power of human input to design efficient

NLP systems that exhibit improved generalization performance compared to traditional supervised learning approaches that rely merely on label information. This demonstrates how powerful these collaborative approaches are when it comes to improving NLP model performance while avoiding many well known pitfalls associated with automation-alone approaches. It may seem that these approaches are no longer necessary as the community has largely adopted large pretrained language models. Though large pretrained language models are widely used, counterfactual augmentation may still be necessary to break spurious correlations as such pretrained models learn the wrong associations from prior texts, for instance, a pretrained model created before the Civil Rights movement might have encoded white supremacy as the norm. Finetuning via augmented data could help combat such undesired spurious patterns. The success of these systems' ability to generalized could prove transformative across a wide range of high impact applications including healthcare systems and autonomous vehicles where reliability is essential due to the potentially life altering consequences they carry. As such, this work provides linchpin information supportive towards continued exploration within this area so as enable technology able to tackle these problems at peak performance levels.

8.1 Future Directions

This thesis is centered on building NLP models with richer forms of human feedback so they are reliable when deployed in the real world. Going forward we believe these models will play an important role towards the practical deployment and use of NLP systems. However, there are still many open questions remaining in this area. In this section, we provide possible extensions to our work, which we believe can enhance the experience of practitioners while building complex sequence systems for real-world scenarios.

8.1.1 Developing more efficient ways to collect counterfactuals from humans

In Chapter 3, we demonstrated the effectiveness of counterfactually augmented data. However, getting crowdworkers to edit documents is costly and the costs increase with the length of the documents. Furthermore, as documents become longer, it is unclear whether crowdworkers are capturing all nuances in the document in order to ensure that the counterfactual document is coherent. Several works [Wu et al., 2021, Madaan et al., 2021, Robeer et al., 2021, Paranjape et al., 2022] have since looked at automatic counterfactual creation but in their attempts to remove humans from the process (and making it cheap to generate counterfactuals), they may suffer from a lack of diversity that human editors would bring to the table. We could explore different ways of eliciting information from crowdworkers that would allow us to construct counterfactuals with less human effort while enjoying the benefits of diversity that humans could offer. For example, instead of asking workers to edit the entire document, we could offer suggestions for potential edits and ask them to now choose one of the plausible alternatives. Following this choice of selection, they could now edit this document to resolve any syntactic and semantic errors that might be in the automatically generated counterfactual.

8.1.2 Developing better ways to incorporating human feedback into machine learning models

In our work on incorporating counterfactual feedback from humans into our models, we chose the examples for which we'd seek counterfactuals through random selection. Then once we obtained these counterfactual examples, to inject causal knowledge into our classifiers, we simply augmented counterfactual examples to the original data. These point to two opportunities for future work. Are there smarter ways to select initial examples that: (i) reduce the number of examples for which counterfactuals are required; and (ii) maximize performance gains over random selection? Furthermore, once we obtain this feedback, how could we smartly incorporate this into our models other than through data augmentation? In this regard, Teney et al. [2020] have explored a gradient supervision approach to better extract signal from this data (and their results point to better improvements versus augmentation). A contrastive learning approach could also be a promising direction.

Additionally, in Chapter 7 we highlighted the role of diversity amongst our human workforce when considering human in the loop feedback. Given humans think in ways that are specific to their environment and cultural context, it stands to reason that if we want our machine learning models to be able to accurately capture human behavior, we need a diverse set of workers who can provide different perspectives. However, we have currently not thought about how to integrate distinct feedback offered for the same example by two humans. For instance, depending on your cultural context, some text might be offensive and not culturally appropriate but in another context it might be completely fine. It is not obvious right now what would be the best way to incorporate both of these into a training scheme that recognizes the cultural contexts behind the feedback offered.

8.1.3 Studying how humans and machine learning systems can work together most effectively to solve problems

Finally, as human in the loop systems are widely deployed, it is critical to understand how they impact the end users. What are the user needs other than simply higher accuracy and reliability? And what would it mean to serve those needs to the fullest extent possible? Several researchers actively work in this area and made advances in answering these questions [Yin et al., 2019, Poursabzi-Sangdeh et al., 2020, Alvarez-Melis et al., 2021, Inel et al., 2021, Lai et al., 2021, 2022, Chen et al., 2022] but a lot remains to be answered.

8.1.4 Developing a research ethics framework for governing labor interactions

In Chapter 8, we shared that if researchers don't seek to either obtain or use, study, analyze or release information *about* a person (in some meaningful sense), then it is not clear that frameworks for the protection of participants in research with human subjects are applicable or appropriate. However, individuals who are not research participants can still be exposed to risks to their well-being and threats to their autonomy. We shared that in these cases crowdworker rights and interests are safeguarded through ethical and regulatory frameworks that govern employment relationships, workplace safety, and other labor practices. However, a coherent framework to govern these researcher-crowdworker relationships does not yet exist. This would be an important direction to pursue future work in.

Bibliography

Rediet Abebe, Shawndra Hill, Jennifer Wortman Vaughan, Peter M Small, and H Andrew Schwartz. Using search queries to understand health information needs in africa. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 3–14, 2019. 6.2

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. Masakhaner: Named entity recognition for african languages. *arXiv preprint arXiv:2103.11811*, 2021. 7.1

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for commonsenseqa: New dataset and models. In *Workshop on Commonsense Reasoning and Knowledge Bases*, 2021. 7.3.2

David Alvarez-Melis, Harmanpreet Kaur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. From human explanation to model interpretability: A framework based on weight of evidence. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2021. 8.1.3

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. Pens: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92, 2021. 7.3.2

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 3.2

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. Cloze-driven pretraining of self-attention networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, November 3-7, 2019*, pages 5359–5368. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1539. URL `https://doi.org/10.18653/v1/D19-1539`. 4

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015. 3.7

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 2020. doi: 10.1162/tacl_a_00338. URL `https://www.aclweb.org/anthology/2020.tacl-1.43`. (document), 5.1, 5.2, 5.3, 5.4, 5.4, 5.5, 5.4

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020. 3.7

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Artificial Intelligence and Statistics (AISTATS)*, 2010. 5.2

Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. 7.1

Emily M Bender, Dirk Hovy, and Alexandra Schofield. Integrating ethics into the nlp curriculum. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–9, 2020. 7.1

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. 7.1

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D Manning. Modeling biological processes for reading comprehension. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 2.1

Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020. 6.2

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020. 7.1

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2015. 3.1

Eric Breck, Marc Light, Gideon Mann, Ellen Riloff, Brianne Brown, and Pranav Anand. Looking under the hood: Tools for diagnosing your question answering engine. In *Association for Computational Linguistics (ACL) Workshop on Open-Domain Question Answering*, 2001. 2.1

Ruth MJ Byrne. Counterfactual thought. *Annual review of psychology*, 67:135–157, 2016. 6.1

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, 2020. 7.1

Patrizia Catellani and Patrizia Milesi. Counterfactuals and roles: Mock victims' and perpetrators' accounts of judicial cases. *European Journal of Social Psychology*, 31(3):247–264, 2001. 6.1

Patrizia Catellani, Mauro Bertolotti, Monia Vagni, and Daniela Pajardi. How expert witnesses' counterfactuals influence causal and responsibility attributions of mock jurors and expert judges. *Applied Cognitive Psychology*, 2020. 6.1

Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. Machine explanations and human understanding. *arXiv preprint arXiv:2202.04092*, 2022. 8.1.3

Danqi Chen, Jason Bolton, and Christopher D Manning. A thorough examination of the cnn/daily mail reading comprehension task. In *Association for Computational Linguistics (ACL)*, 2016. 2.1, 2.2, 2.4, 3.2

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/W19-2008. URL `https://www.aclweb.org/anthology/W19-2008`. 5.2

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. Seeing things from a different angle: Discovering diverse perspectives about claims. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, 2019b. 4.1

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020a. URL `https://openreview.net/forum?id=r1xMH1BtvB`. 5.1, 5.3, 5.4

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*, 2020b. 4.1

Peter Clark and Oren Etzioni. My computer is an honor student but how intelligent is it? standardized tests as a measure of ai. *AI Magazine*, 37(1):5–12, 2016. 1

Sanjoy Dasgupta and Sivan Sabato. Robust learning from discriminative feature feedback. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020. 4.1

Sanjoy Dasgupta, Akansha Dey, Nicholas Roberts, and Sivan Sabato. Learning from discriminative feature feedback. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2018. 4.1

Pradeep Dasigi, Nelson F Liu, Ana Marasovic, Noah A Smith, and Matt Gardner. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. *arXiv preprint arXiv:1908.05803*, 2019. 5.2

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 3.4, 3.7, 4.1, 5.1, 5.3, 5.4

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Association for Computational Linguistics (ACL)*, 2020. (document), 3.1, 3.7, 3.17, 4.1, 4.2, 7.3.1

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. Gated-attention readers for text comprehension. In *Association for Computational Linguistics (ACL)*, 2017. 2.3

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 5.1, 5.2, 5.4

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2018. 3.2

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, 2019. 7.1

Alexey Drutsa, Dmitry Ustalov, Valentina Fedorova, Olga Megorskaya, and Daria Baidakova. Crowdsourcing natural language data at scale: A hands-on tutorial. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 25–30, 2021. 7.1

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 5.1, 5.2, 3

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017. 3

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5401. URL `https://www.aclweb.org/anthology/W17-5401`. 5.2, 5.5

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. 3

Sina Fazelpour. Norms in counterfactual selection. *Philosophy and Phenomenological Research*, 2020. 6.1

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5801.

URL `https://www.aclweb.org/anthology/D19-5801`. 5.1

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420, 2011. doi: 10.1162/COLI_a_00057. URL `https://www.aclweb.org/anthology/J11-2010`. 7.1

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *arXiv preprint arXiv:2104.14478*, 2021. 7.1

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1307–1323, 2020. 7.3.1

Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1107. URL `https://www.aclweb.org/anthology/D19-1107`. 7.3.2

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3.2

Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking nli systems with sentences that require simple lexical inferences. In *Association for Computational Linguistics (ACL)*, 2018. 2.1, 3.2

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2.1

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6), 2005. 3.4, 3.7

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018. 2.1, 3.1, 3.2, 5.1

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. Inspired: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152, 2020. 7.3

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.244. URL https://www.aclweb.org/anthology/2020.acl-main.244. 4

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2.1, 2.2, 2.4, 3.2

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The Goldilocks principle: Reading children's books with explicit memory representations. In *International Conference on Learning Representations (ICLR)*, 2016. 2.1, 2.2, 2.4

Lynette Hirschman, Marc Light, Eric Breck, and John D Burger. Deep read: A reading comprehension system. In *Association for Computational Linguistics on Computational Linguistics (ACL)*, 1999. 2.1

Dirk Hovy and Shannon L Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, 2016. 7.1

Dirk Hovy, Barbara Plank, and Anders Søgaard. Experiments with crowdsourced re-annotation of a pos tagging data set. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 377–382, 2014. 7.3.1

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Michael Junk, and Chin-Yew Lin. Question answering in webclopedia. In *TREC*, volume 52, 2000. (document), 5.1, 5.5, 5.3

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *International Conference on Machine Learning (ICML)*, 2017. 3.7, 3.16

William Huang, Haokun Liu, and Samuel Bowman. Counterfactually-augmented snli training data does not yield better generalization than unaugmented data. In *First Workshop on Insights from Negative Results in NLP*, 2020. 4.3

Thomas F Icard, Jonathan F Kominsky, and Joshua Knobe. Normality and actual causal strength. *Cognition*, 161:80–93, 2017. 6.1

Oana Inel, Tomislav Duricic, Harmanpreet Kaur, Elisabeth Lex, and Nava Tintarev. Design implications for explanations: A case study on supporting reflective assessment of potentially misleading videos. *Frontiers in artificial intelligence*, 4, 2021. 8.1.3

Panos Ipeirotis. Mechanical Turk, Human Subjects, and IRB's. URL https://www.behind-the-enemy-lines.com/2009/01/mechanical-turk-human-subjects-and-irbs.html. 7.1

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018. 3.2

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. Learning to faithfully rationalize by construction. In *Association for Computational Linguistics (ACL)*, 2020. (document), 4.1, 4.2, 4.3, 4.2

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 1, 2.5, 3.2

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Association for Computational Linguistics (ACL)*, 2017. 1, 3

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. Text understanding with the attention sum reader network. In *Association for Computational Linguistics (ACL)*, 2016. 2.3

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using mechanical turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, 2021. 6.2

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL `https://www.aclweb.org/anthology/2020.emnlp-main.550`. 5.1, 1, 5.4

Anurag Katakkar, Weiqin Wang, Clay H Yoo, Zachary C Lipton, and Divyansh Kaushik. Practical benefits of feature feedback under distribution shift. *arXiv preprint arXiv:2110.07566*, 2021. 1.1

Divyansh Kaushik and Zachary C Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 1.1, 3.1, 3.2, 5.1

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*, 2020. (document), 1.1, 3.7, 3.10, 3.11, 3.12, 3.7, 3.17, 3.5, 4, 4.1, 4.3, 5.1, 5.2, 7.3.1

Divyansh Kaushik, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih. On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021a. 1.1, 7.3

Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. Explaining the efficacy of counterfactually-augmented data. *International Conference on Learning Representations (ICLR)*, 2021b. 1.1, 4.1, 4.2, 4.3, 5.2

Divyansh Kaushik, Zachary C Lipton, and Alex John London. Resolving the human subjects status of machine learning's crowdworkers. *arXiv preprint arXiv:2206.04039*, 2022. 1.1

Aniruddha Kembhavi, Min Joon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5376–5384. IEEE

Computer Society, 2017. doi: 10.1109/CVPR.2017.571. URL https://doi.org/10.1109/CVPR.2017.571. 3

Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. More bang for your buck: Natural perturbation for robust question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 163–170, 2020. 7.3

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, 2021. URL https://www.aclweb.org/anthology/2021.naacl-main.324. 5.3

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 3.4

Svetlana Kiritchenko and Saif Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In *Joint Conference on Lexical and Computational Semantics (*SEM)*, 2018. 3.1, 3.2, 7.1

Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, and Kristen Grauman. Crowdsourcing in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 10(3):177–243, 2016. 7.1, 7.3.1

Jonathan K Kummerfeld. Quantifying and avoiding unfair qualification labour in crowdsourcing. *arXiv preprint arXiv:2105.12762*, 2021. 7.1

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 08 2019. URL https://doi.org/10.1162/tacl_a_00276. 5.3, 3

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*, 2001. 4.2

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 2.1, 1, 3

Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021. 8.1.3

Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. Human-ai collaboration via conditional delegation: A case study of content moderation. In *CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022. 8.1.3

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. Qed: A framework and dataset for explanations in question answering. *arXiv preprint arXiv:2009.06354*, 2020. 4.4

Hung Le, Chinnadhurai Sankar, Seungwhan Moon, Ahmad Beirami, Alborz Geramifard, and Satwik Kottur. Dvd: A diagnostic dataset for multi-step reasoning in video grounded dialogue. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5651–5665, 2021. 7.3.2

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1612. URL `https://www.aclweb.org/anthology/P19-1612`. 5.1, 5.3

Moontae Lee, Xiaodong He, Wen-tau Yih, Jianfeng Gao, Li Deng, and Paul Smolensky. Reasoning in vector space: An exploratory study of question answering. In *International Conference on Learning Representations (ICLR)*, 2016. 2.1, 2.2

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. Inferring which medical treatments work from reports of clinical trials. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 4.1

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 4.1

Jochen L Leidner and Vassilis Plachouras. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, 2017. 7.1

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, August 2017. doi: 10.18653/v1/K17-1034. URL `https://www.aclweb.org/anthology/K17-1034`. 3

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020. 6.1

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp. In *Proceedings of NAACL-HLT*, pages 681–691, 2016. 3.7

Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018. 3.7, 3.16

Zachary C Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. In *International Conference on Machine Learning (ICML) Machine Learning Debates Workshop*,

2018. 2.5

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3128–3136. PMLR, 2018. URL `http://proceedings.mlr.press/v80/lipton18a.html`. 1, 5.2

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 5.4

Alex John London. *For the Common Good: Philosophical Foundations of Research Ethics*. Oxford University Press, 2021. 7.1, 7.3.2

Alex John London and Jonathan Kimmelman. Clinical trial portfolios: a critical oversight in human research ethics, drug regulation, and policy. *Hastings Center Report*, 49(4):31–41, 2019. 7.3.2

Alex John London, Monica Taljaard, and Charles Weijer. Loopholes in the research ethics system? informed consent waivers in cluster randomized trials with individual-level intervention. *Ethics & human research*, 42(6):21–28, 2020. 7.1, 7.3.2

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 4.2

David Lowell, Zachary C. Lipton, and Byron C. Wallace. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 21–30, November 2019. doi: 10.18653/v1/D19-1003. URL `https://www.aclweb.org/anthology/D19-1003`. 5.2

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. *arXiv preprint arXiv:1807.11714*, 2018. 3.2

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2011. 3.1

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. Politeness transfer: A tag and generate approach. *arXiv preprint arXiv:2004.14257*, 2020. 3.7, 3.16

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13516–13524, 2021. 8.1.1

Sagnik Majumder, Chinmoy Samant, and Greg Durrett. Model agnostic answer reranking system for adversarial question answering. In *European Chapter of the Association for Computational Linguistics: Student Research Workshop (EACL SRW)*, 2021. 4.4

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. It's all in the name:

Mitigating gender bias with name-based counterfactual data substitution. *arXiv preprint arXiv:1909.00871*, 2019. 3.2

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, 2018. 7.3.2

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 2.1, 1, 2.3

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning (ICML)*, 2010. 3.4

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/S13-2052`. 3.5

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, et al. Participatory research for low-resourced machine translation: A case study in african languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2144–2160, 2020. 7.1

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016. 2.1, 1

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1018. URL `https://www.aclweb.org/anthology/D19-1018`. 3.5

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019b. 4.3

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, 2020. 5.1, 5.2

Qiang Ning, Hao Wu, Pradeep Dasigi, Dheeru Dua, Matt Gardner, Robert L Logan IV, Ana

Marasovic, and Zhen Nie. Easy, reproducible and quality-controlled data collection with crowdaq. *arXiv preprint arXiv:2010.06694*, 2020. 7.1

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. Who did what: A large-scale person-centered cloze dataset. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 2.1, 2.2, 2.4

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*, 2020. 7.1

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. The lambada dataset: Word prediction requiring a broad discourse context. In *Association for Computational Linguistics (ACL)*, 2016. 1, 5.2

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. Retrieval-guided counterfactual generation for qa. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1670–1686, 2022. 8.1.1

Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe, Caroline Sporleder, Corina Forascu, Yassine Benajiba, and Petya Osenova. Overview of qa4mre at clef 2012: Question answering for machine reading evaluation. 2012. 2.1

Judea Pearl. *Causality*. Cambridge university press, 2009. 3.6

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research (JMLR)*, 12, 2011. 3.4

Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe, Corina Forascu, and Caroline Sporleder. Overview of qa4mre at clef 2011: Question answering for machine reading evaluation. 2011. 2.1

Verónica Pérez-Rosas and Rada Mihalcea. Experiments in open domain deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1125, 2015. 7.3

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 5(78):947–1012, 2016. 3.2

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://www.aclweb.org/anthology/N18-1202`. 3.4

Jonas Pfeiffer, Aishwarya Kamath, Iryna Gurevych, and Sebastian Ruder. What do deep networks like to read? *arXiv preprint arXiv:1909.04547*, 2019. 3.2

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis Only Baselines in Natural Language Inference. In *Joint Conference on Lexical and Computational Semantics (*Sem)*, 2018. 1, 2.1, 3.1, 3.2, 3.4, 5.1

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. DynaSent: A Dynamic Benchmark for Sentiment Analysis. *arXiv preprint arXiv:2012.15349*, 2020. 5.5

Stefanos Poulis and Sanjoy Dasgupta. Learning with feature feedback: from theory to practice. In *Artificial Intelligence and Statistics (AISTATS)*, 2017. 3.2, 4.1

Forough Poursabzi-Sangdeh, Samira Samadi, Jennifer Wortman Vaughan, and Hanna Wallach. A human in the loop is not enough: The need for human-subject experiments in facial recognition. In *CHI Workshop on Human-Centered Approaches to Fair and Responsible AI*, 2020. 8.1.3

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, 2021. 6.2

Danish Pruthi, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. Weakly-and semi-supervised evidence extraction. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020. (document), 4.1, 4.2, 4.3, 4.2, 4.5, 4.7, 4.8

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009. 1

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 6.1

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 6.1

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 1, 2.1, 2.2, 5.2, 2, 3

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging nlp models. In *Association for Computational Linguistics (ACL)*, 2018. 3.2

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2013. 2.1, 5.2

Marcel Robeer, Floris Bex, and Ad Feelders. Generating realistic natural language counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3611–3625, 2021. 8.1.1

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, 2021. 6.1

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *International Workshop on Semantic Evaluation (SemEval)*, 2017. 4.3

Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *AAAI Conference on Artificial Intelligence*, 2018. 4.1

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: training differentiable models by constraining their explanations. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017. 4.1

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, 2018. 7.1

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1156. URL `https://www.aclweb.org/anthology/P18-1156`. 3

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*, 2021. 6.2

B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij. On causal and anticausal learning. In *29th International Conference on Machine Learning (ICML 2012)*. International Machine Learning Society, 2012. 5.2

D Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner's curse? on pace, progress, and empirical rigor. In *International Conference on Learning Representations (ICLR) Workshop Track*, 2018. 2.5

Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, 2017. 6.1

Judy Hanwen Shen, Lauren Fratamico, Iyad Rahwan, and Alexander M Rush. Darling or babygirl? investigating stylistic bias in sentiment analysis. *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML).*, 2018. 3.2

Victor S Sheng and Jing Zhang. Machine learning with crowdsourcing: A brief summary of the past research and future directions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9837–9843, 2019. 7.1

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

1, 5.2

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. Beyond fair pay: Ethical implications of nlp crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, 2021. 7.1, 7.1, 7.1, 7.1, 7.4

M Six Silberman, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar. Responsible research with crowds: pay crowdworkers at least minimum wage. *Communications of the ACM*, 61(3):39–41, 2018. 7.1

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, 2022. 6.3

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2013. 4.3

Joe Stacey, Yonatan Belinkov, and Marek Rei. Natural language inference with a human touch: Using human explanations to guide model attention. *arXiv preprint arXiv:2104.08142*, 2021. 4.4

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13693–13696, 2020. 7.1

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. Transforming delete, retrieve, generate approach for controlled text style transfer. *arXiv preprint arXiv:1908.09368*, 2019. 3.7, 3.16

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems (NIPS)*, 2015. 2.3

Richard Sutcliffe, Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Corina Forascu, Yassine Benajiba, and Petya Osenova. Overview of qa4mre main task at clef 2013. 2013. 2.1

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011. 7.3.1

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, 2019. 7.3.2

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*, 2020. 8.1.2

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL `https://www.aclweb.org/anthology/N18-1074`. 5.2

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6601. URL `https://www.aclweb.org/anthology/D19-6601`. 5.2

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. In *Workshop on Representation Learning for NLP (RepL4NLP)*, 2017. 2.1, 1

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1), 2015. 3

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633, 2020. doi: 10.1162/tacl_a_00335. URL `https://www.aclweb.org/anthology/2020.tacl-1.40`. 4

Jennifer Wortman Vaughan. Making better use of the crowd: How crowdsourcing can advance machine learning research. *J. Mach. Learn. Res.*, 18(1):7026–7071, 2017. 7.1

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for nlp. *arXiv preprint arXiv:1908.07125*, 2019a. 3.2

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401, 2019b. 5.2

Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela. Analyzing dynamic adversarial training data in the limit. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 202–217, 2022. 6.1

Mengqiu Wang, Noah A Smith, and Teruko Mitamura. What is the jeopardy model? a quasi-synchronous grammar for qa. In *Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007. 2.1

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. In *International Conference on Learning Representations (ICLR)*, 2016. 2.1, 2.2

Mark E Whiting, Grant Hugh, and Michael S Bernstein. Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the AAAI Conference on Human Computation and*

Crowdsourcing, volume 7, pages 197–206, 2019. 7.1

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-ai collaboration for generating free-text explanations. *arXiv preprint arXiv:2112.08674*, 2021. 6.1

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, 2018a. (document), 3.17, 7.3.2

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018b. 4.3

Adina Williams, Tristan Thrush, and Douwe Kiela. Anlizing the adversarial natural language inference dataset. *arXiv preprint arXiv:2010.12729*, 2020. 5.2

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`. 5.3

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, 2021. 8.1.1

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*, 2020. 7.3.2

Yi Yang, Wen-tau Yih, and Christopher Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2015. 2.1, 1

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, 2018a. 5.2, 3

Zhilin Yang, Saizheng Zhang, Jack Urbanek, Will Feng, Alexander H. Miller, Arthur Szlam, Douwe Kiela, and Jason Weston. Mastering the dungeon: Grounded language learning by mechanical turker descent. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018b. URL `https://openreview.net/forum?id=SJ-C6JbRW`. 5.2

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019. 8.1.3

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. In *International Conference on Learning Representations (ICLR)*, 2018. 2.3

Omar Zaidan, Jason Eisner, and Christine Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies: North American chapter of the association for computational linguistics (NAACL-HLT)*, 2007. (document), 3.1, 3.2, 3.7, 3.13, 3.14, 3.15, 3.7, 3.6, 4.1, 4.2, 4.2, 4.3, 4.3, 4.3, 4.6, 4.4, 7.3.1

Omar F Zaidan and Jason Eisner. Modeling annotators: A generative approach to learning from annotator rationales. In *Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2008. 3.2

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152, 2021. 7.3.2

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1009. URL `https://www.aclweb.org/anthology/D18-1009`. 5.2

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, July 2019. doi: 10.18653/v1/P19-1472. URL `https://www.aclweb.org/anthology/P19-1472`. 5.2

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. Multimet: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214–3225, 2021. 7.3.2

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*, 2018. 5.2

Ye Zhang, Iain Marshall, and Byron C Wallace. Rationale-augmented convolutional neural networks for text classification. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 4.1

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018. 3.2

Ran Zmigrod, Sebastian J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Association for Computational Linguistics (ACL)*, 2019. 3.2

Geoffrey Zweig and Chris J.C. Burges. A challenge set for advancing language modeling. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 29–36, Montréal, Canada, June 2012. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W12-2704`. 5.2