Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions

Mihir Parmar[1*] **Swaroop Mishra**[1*] **Mor Geva**[2†] **Chitta Baral**[1]
[1]Arizona State University [2]Allen Institute for AI
{mparmar3, srmishr1, chitta}@asu.edu, pipek@google.com

Abstract

In recent years, progress in NLU has been driven by benchmarks. These benchmarks are typically collected by crowdsourcing, where annotators write examples based on annotation instructions crafted by dataset creators. In this work, we hypothesize that annotators pick up on patterns in the crowdsourcing instructions, which bias them to write many similar examples that are then over-represented in the collected data. We study this form of bias, termed *instruction bias*, in 14 recent NLU benchmarks, showing that instruction examples often exhibit concrete patterns, which are propagated by crowdworkers to the collected data. This extends previous work (Geva et al., 2019) and raises a new concern of whether we are modeling the *dataset creator's instructions*, rather than the task. Through a series of experiments, we show that, indeed, instruction bias can lead to overestimation of model performance, and that models struggle to generalize beyond biases originating in the crowdsourcing instructions. We further analyze the influence of instruction bias in terms of pattern frequency and model size, and derive concrete recommendations for creating future NLU benchmarks.[1]

1 Introduction

Benchmarks have been proven pivotal for driving progress in Natural Language Understanding (NLU) in recent years (Rogers et al., 2021; Bach et al., 2022; Wang et al., 2022). Nowadays, NLU benchmarks are mostly created through crowdsourcing, where crowdworkers write examples following annotation instructions crafted by dataset creators (Callison-Burch and Dredze, 2010; Zheng et al., 2018; Suhr et al., 2021). The instructions typically include a short description of the task, along with several examples (Dasigi et al., 2019; Zhou et al., 2019; Sakaguchi et al., 2020).

Despite the vast success of this method, past studies have shown that data collected through crowdsourcing often exhibit various biases that lead to overestimation of model performance (Schwartz et al., 2017; Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018; Le Bras et al., 2020; Mishra et al., 2020a; Mishra and Arunkumar, 2021; Hettiachchi et al., 2021). Such biases are often attributed to annotator-related biases, such as writing style and background knowledge (Gururangan et al., 2018; Geva et al., 2019) (see more discussion on related work in §A).

In this work, we propose that biases in crowdsourced NLU benchmarks often originate at an early stage in the data collection process of designing the annotation task. In particular, we hypothesize that task instructions provided by dataset creators, which serve as the guiding principles for annotators to complete the task, often influence crowdworkers to follow specific patterns, which are then propagated to the dataset and subsequently over-represented in the collected data. For instance, $\sim 36\%$ of the instruction examples for the QUOREF dataset (Dasigi et al., 2019) start with *"What is the name"*, and this same pattern can be observed in $\sim 59\%$ of the collected instances.

To test our hypothesis, we conduct a broad study of this form of bias, termed *instruction bias*, in 14 recent NLU benchmarks. We find that instruction bias is evident in most of these datasets, showing that $\sim 73\%$ of instruction examples on average share a few clear patterns. Moreover, we find that these patterns are propagated by annotators to the collected data, covering $\sim 61\%$ of the instances on average. This suggests that instruction examples play a critical role in the data collection process and the resulting example distribution.

It is difficult to represent a task with a few examples, and bias in instruction examples makes

*Equal Contribution
†Now at Google Research
[1]Code and data is available at https://github.com/Mihir3009/instruction-bias.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1779–1789
May 2-6, 2023 ©2023 Association for Computational Linguistics

it even more difficult since a task and its associated reasoning have a larger scope than instruction patterns. For example co-reference resolution, temporal commonsense reasoning, and numerical reasoning are much broader tasks than the prevalent patterns in QUOREF (*"what is the name..."*), MC-TACO (*"how long..."*) and DROP (*"how many field goals..."*) datasets.

We investigate the effect of instruction bias on model performance, showing that performance is overestimated by instruction bias and that models often fail to generalize beyond instruction patterns. Moreover, we observe that a higher frequency of instruction patterns in the training set often increases the model performance gap on pattern and non-pattern examples and that large models are generally less sensitive to instruction bias.

In conclusion, our work shows that instruction bias widely exists in NLU benchmarks, often leading to an overestimation of model performance. Based on our study, we derive concrete recommendations for monitoring and alleviating this bias in future data collection efforts. From a broader perspective, our findings also have implications on the recent learning-by-instructions paradigm (Efrat and Levy, 2020; Mishra et al., 2021), where crowdsourcing instructions are used in model training.

2 Instruction Bias in NLU Benchmarks

Instructions are the primary resource for educating crowdworkers on how to perform their task (Nangia et al., 2021). Bias in the instructions, dubbed *instruction bias*, could lead crowdworkers to propagate specific patterns to the collected data.

Here, we study instruction bias in NLU benchmarks[2], focusing on two research questions: (a) Do crowdsourcing instructions exhibit patterns that annotators can pick up on? and (b) Are such patterns propagated by crowdworkers to the collected data? In our study, we use the instructions of 14 recent NLU benchmarks:[3] (1) CLARIQ (Aliannejadi et al., 2020), (2) COSMOSQA (Huang et al., 2019), (3) DROP (Dua et al., 2019), (4) DUORC (Saha et al., 2018), and (5) HOTPOTQA (Yang et al., 2018) (6) HYBRIDQA (Chen et al., 2020), (7) MC-TACO (Zhou et al., 2019), (8) MULTIRC (Khashabi et al., 2018), (9) PIQA (Bisk et al., 2020), (10) QASC (Khot et al., 2020), (11) QUOREF (Dasigi et al., 2019), (12) ROPES (Lin et al., 2019), (13) SCIQA (Welbl et al., 2017), (14) WINOGRANDE (Sakaguchi et al., 2020). These benchmarks were created through different crowdsourcing protocols to evaluate diverse tasks (Mishra et al., 2021) (see dataset statistics in §B).

Tab. 2 provides the number of examples present in crowdsourcing instructions of each dataset. From Tab. 2, we can observe that our analysis

[2] All benchmarks are in English.

[3] The instructions were obtained from Mishra et al. (2021), who have collected those from the dataset authors.

Dataset	Pattern	% Ins.	% $\mathcal{S}_{\text{train}}$	% $\mathcal{S}_{\text{test}}$
CLARIQ	`[Are|Would|Do] you`	72.2	85.1	89
COSMOSQA	`What AUX`	87.5	45.1	38.4
DROP	`How many [field goals	years	yards	points	touchdowns]`	70	62.5	62.5
DUORC	`[How old	How	What	Who] AUX`	70	85.1	84
HOTPOTQA	`[In	Of	From	_] [Which	What] AUX`	87.5	53.8	54.2
HYBRIDQA	`Which AUX`	29.4	25.7	15.1
MC-TACO	`How long AUX`	100	-	87.6
	`What AUX`	100	-	90.1
	`How often AUX`	100	-	85.3
	`AUX... [still|always |by the time]`	100	-	67.3
	`When did / What time`	100	-	83.4
MULTIRC	`What AUX`	14.3	38.4	41.5
PIQA	`How [do|can]`	66.7	43.7	42.9
QASC	`What AUX`	57.1	49.3	47
QUOREF	`What AUX the [_| full|real|first |last] name`	36.4	57	60
ROPES	`Which AUX`	42.9	74.1	20.7
SCIQA	`What AUX`	100	83.6	84.5
WINOGRANDE	`[because	so	while	since	but] ... the`	73.7	63.4	63.1
Average		72.7	59	62

Table 1: Portion of patterns in instruction examples (Ins.) and in the corresponding train ($\mathcal{S}_{\text{train}}$) and test ($\mathcal{S}_{\text{test}}$) sets of NLU datasets. AUX $\in$ {am, is, are, was, were, has, have, had, do, does, did, will, would, can, could, may, might, shall, should, must}, and _ is an empty string. MC-TACO has 5 different data subsets corresponding to different types of temporal reasoning (see Tab. 2), hence, the sum of percentages for this dataset exceeds 100%.

Dataset	Task	# of Examples
CLARIQ	Clarification QA	18
COSMOSQA	Commonsense Reasoning	8
DROP	Numerical Reasoning	10
DUORC	Paraphrased RC	10
HOTPOTQA	Multi-hop QA	8
HYBRIDQA	QA	17
MC-TACO	Event Duration	3
	Event Ordering	2
	Frequency	2
	Stationary	2
	Absolute Point	2
MULTIRC	Complex QA	7
PIQA	Physical Interaction QA	6
QASC	Complex QA	7
QUOREF	Coreference QA	11
ROPES	RC	14
SCIQA	Science-based QA	6
WINOGRANDE	Commonsense Reasoning	19
Average		8.4

Table 2: Tasks of each dataset and number of examples in crowdsourcing instruction of each dataset. RC: Reading Comprehension, QA: Question Answering.

involves a wide range of different tasks. Also, we believe that the lower number of examples in crowdsourcing instructions might be limiting the imagination of annotators while creating samples, resulting in instruction bias.

2.1 Patterns in Crowdsourcing Instructions

Our goal is to quantify biases in instruction examples that propagate to collected data instances. In this study, we focus on an intuitive form of bias of recurring word patterns, which crowdworkers can easily pick up on. To find such patterns, we manually analyze the instruction examples of each dataset to find a *dominant pattern*, using the following procedure: (a) identifying repeating patterns of $n \geq 2$ words, (b) merging patterns that are semantically similar or have a substantial word overlap, and (c) selecting the most frequent pattern as the dominant pattern (an example is provided in §C).

Tab. 1 shows the dominant pattern in the instruction examples of each dataset. On average, 72.7% of the instruction examples used to create a dataset exhibit the same dominant pattern, and for 10 out of 14 datasets, the dominant pattern covers more than half of the instruction examples. This suggests that crowdsourcing instructions demonstrate a small set of repeating "shallow" patterns. Moreover, the short length of the patterns (2-4 words) and the typical low number of instruction examples (Tab. 2) make the patterns easily visible to crowdworkers, who can end up following them.

Notably, our results are an underestimation of the actual instruction bias, since (a) we only consider the dominant pattern for each dataset (b) our manual analysis over instruction examples has a preference for short patterns (c) we do not consider paraphrased patterns (beyond the shallow paraphrases which are visible in annotation instructions), and (d) datasets may include implicit patterns (e.g. writing style and biases from the annotator's background knowledge) that also contribute to instruction bias. Accounting for such patterns is expected to increase the bias percentage in Tab. 1 further.

2.2 Instruction Bias Propagation to Datasets

We now turn to investigate whether patterns in instruction examples are further propagated by crowdworkers to the collected data. We analyze the train and test sets of each benchmark[4] to find the same patterns, using simple string matching. To account for syntactic modifications in identified patterns based on some examples from dataset, we also consider synonym words where appropriate and match the paraphrased version of each pattern.

Tab. 1 shows the results. Across all datasets, instruction patterns are ubiquitous in the collected data, occurring in 60.5% of the instances on average, with similar presence in training (59%) and test (62%) examples. While the dominant pattern's frequency in the data is typically not higher than in the instructions, for CLARIQ, DUORC, MULTIRC, QUOREF and ROPES, the pattern frequency was amplified by the crowdworkers. Interestingly, these datasets used a relatively large number of instruction examples (Tab. 2), suggesting that more examples do not necessarily alleviate the propagation of instruction bias. Example data instances with instruction patterns are provided in §D.

A natural question that arises is whether patterns in collected data reflect the true task distribution rather than a bias in the instructions. We argue that this is highly unlikely. First, while the space of possible patterns for a NLU task is arguably large, the dataset patterns are imbalanced proportionately

[4]If no explicit test set exists, we use the validation set.

	Base			Large		
	$\mathcal{S}^{p}_{\text{test}}$	$\mathcal{S}^{-p}_{\text{test}}$		$\mathcal{S}^{p}_{\text{test}}$	$\mathcal{S}^{-p}_{\text{test}}$	
CLARIQ	30.5	27.4	10.2% ↓	30.3	26.3	13.2% ↓
DROP	75.7	31.8	58% ↓			
MULTIRC	40.5	33.7	16.8% ↓	42	37.4	11% ↓
PIQA	20.7	15	27.5% ↓	21.8	15.3	29.8% ↓
QUOREF	85.9	66	23.2% ↓	92.1	81.1	11.9% ↓
ROPES	57	42	26.3% ↓	57.8	58.2	0.7% ↑
SCIQA	80.7	79.7	1.2% ↓	82.8	81.9	1.1% ↓
Average	55.9	42.2	24.5% ↓	54.5	50	8.3% ↓

Table 3: Performance on $\mathcal{S}^{p}_{\text{test}}$ vs. $\mathcal{S}^{-p}_{\text{test}}$ of models trained on data instances containing instruction patterns ($\mathcal{S}^{p}_{\text{train}}$).

	Base			Large		
	$\mathcal{S}^{p}_{\text{test}}$	$\mathcal{S}^{-p}_{\text{test}}$		$\mathcal{S}^{p}_{\text{test}}$	$\mathcal{S}^{-p}_{\text{test}}$	
CLARIQ	30.5	26.9	11.8% ↓	30.7	26.9	12.4% ↓
DROP	76	78.9	3.8% ↑			
MULTIRC	40.5	39	3.7% ↓	42.4	44.5	5% ↑
PIQA	20.7	19.8	4.4% ↓	21.9	20.6	5.9% ↓
QUOREF	86.7	73.1	15.7% ↓	92.1	81.1	11.9% ↓
ROPES	59.9	48.7	18.7% ↓	59	64.2	8.8% ↑
SCIQA	80.6	80.2	0.5% ↓	82.8	82.9	0.1% ↑
Average	56.4	52.4	7.1% ↓	54.8	53.4	2.6% ↓

Table 4: Performance on $\mathcal{S}^{p}_{\text{test}}$ vs. $\mathcal{S}^{-p}_{\text{test}}$ of models trained on $\mathcal{S}_{\text{train}}$.

to the patterns in the instructions for collecting it. For example, questions for assessing temporal commonsense reasoning could have various forms, such as *"What is the duration of..."*, *"For how much time..."*, and *"How long..."*. However, MC-TACO (event duration) is heavily dominated (87.6%) by questions with the pattern *"how long"*, which appears in 100% of the instruction examples (Tab. 1). In addition, datasets of similar tasks have different dominant patterns, while each dataset's dominant pattern correlates with the pattern in the corresponding instructions; the pattern 'What is' appears in 48% of the questions in QASC and in 57% of the instruction examples, but it is entirely different from the dominant pattern of HOTPOTQA, which is another QA dataset for multi-hop reasoning.

We further validate the propagation of bias in instruction examples by comparing the pattern distributions of collected instances when the instructions include and do not include examples. We conduct this experiment for MC-TACO and QUOREF and find that, without any examples provided, the dominant pattern is substantially less frequent, showing that instruction bias is propagated during data collection. Full details are provided in §E.

Propagation of instruction bias to the test set raises concerns regarding its reliability for evaluation, which we address next.

3 Effect on Model Learning

Let $\mathcal{S}_{\text{train}}$ ($\mathcal{S}_{\text{test}}$) be the set of training (test) examples, and denote by $\mathcal{S}^{p}_{\text{train}}$ ($\mathcal{S}^{p}_{\text{test}}$) and $\mathcal{S}^{-p}_{\text{train}}$ ($\mathcal{S}^{-p}_{\text{test}}$) its disjoint subsets of examples with and without instruction patterns, respectively. We conduct two experiments where we fine-tune models on (a) $\mathcal{S}^{p}_{\text{train}}$ and (b) $\mathcal{S}^{p}_{\text{train}} \cup \mathcal{S}^{-p}_{\text{train}}$, and evaluate them on $\mathcal{S}^{-p}_{\text{test}}$ and $\mathcal{S}^{p}_{\text{test}}$. This is to assess to what extent models generalize from instruction patterns to the downstream task (a), and to compare model performance on instances with and without instruction patterns (b).

3.1 Experimental Setting

Datasets Since model training is computationally expensive, we select a subset of seven datasets from those analyzed in §2: (1) CLARIQ, (2) DROP, (3) MULTIRC, (4) PIQA, (5) QUOREF, (6) ROPES, and (7) SCIQA. These datasets cover a variety of tasks, different types and levels of instruction bias (Tab. 1), and are different in size (§B).

Models For all datasets except DROP, we evaluate T5-base and T5-large (Raffel et al., 2020), and BART-base and BART-large (Lewis et al., 2020). For DROP, we use Numnet+ (Ran et al., 2019), a RoBERTa model (Liu et al., 2019) with specialized output heads for numerical reasoning. Numnet+ has 355M parameters, which is closer to T5-base (220M) than to T5-large (770M) in size.

Evaluation We evaluate model performance using the standard F_1 evaluation score, and report the average score over three random seeds.

3.2 Results

We observe similar results for T5 and BART, and thus, present only the results for T5 in this section. Results for BART are provided in §F.

Models often fail to generalize beyond instruction patterns. Tab. 3 shows the performance on $\mathcal{S}^{p}_{\text{test}}$ and $\mathcal{S}^{-p}_{\text{test}}$ when training only on examples with instruction patterns. Across all experiments, there are large performance gaps, reaching to 58% in DROP and $> 10\%$ in both base and large models for CLARIQ, MULTIRC, PIQA, and QUOREF. This indicates that models trained only on examples with instruction patterns fail to generalize to other task examples, and stresses that instruction bias should be monitored and avoided during data collection. Notably, the gap is lower for large models than for base ones, showing that large models are less sensitive to instruction bias. This might be attributed to their larger capacity to capture knowledge and skills during pre-training.

Model performance is overestimated by instruction bias. We compare the performance on $\mathcal{S}^{p}_{\text{test}}$ and $\mathcal{S}^{-p}_{\text{test}}$ of models trained on the full training set (Tab. 4). The average performance across all datasets is higher on examples that exhibit instruction patterns by $\sim 7\%$ and $\sim 3\%$ for the base and large models, respectively. Specifically, base models perform worse on $\mathcal{S}^{-p}_{\text{test}}$ than on $\mathcal{S}^{p}_{\text{test}}$ for all datasets except DROP, in some cases by a dramatic gap of $> 15\%$ (e.g. 18.7% in ROPES and 15.7% in QUOREF). In contrast, results for the large models vary across datasets, while the performance gap is generally smaller in magnitude. This shows that model performance is often overestimated by instructions bias, and reiterates that large models are generally less sensitive to instruction patterns.

4 Conclusions and Discussion

We identify a prominent source of bias in crowdsourced NLU datasets, called *instruction bias*, which originates in annotation instructions written by dataset creators. We study this bias in 14 NLU benchmarks, showing that instruction examples used to create NLU benchmarks often exhibit clear patterns that are propagated by annotators to the collected data. In addition, we investigate the effect of instruction bias on model performance, showing that instruction patterns can lead to overestimated performance as well as limit the ability of models to generalize to other task examples.

Based on our findings, we derive three recommendations for future crowdsourced NLU benchmarks: (1) Crowdsourcing instructions should be diverse; this could be achieved, for example, by having a large number of instructive examples, rephrasing examples using neural models, or periodically sampling examples from a *diverse* set of previously collected examples. The latter could be done, for example, by maintaining a pool of diverse examples during the collection process and then presenting every annotator with a different random sample from this growing pool. (2) Word patterns in collected instances should be analyzed during data collection, as well as possible correspondence to instruction examples. Such analysis will help researchers monitor the collection process and the quality of the resulting data. (3) Correlation between model performance and input patterns should be checked during evaluation.

Limitations

This work covers 14 NLU datasets, for which annotation instructions are publicly available. However, most of these datasets are QA datasets. Our analysis can be extended to other NLU task categories, such as Natural Language Inference (NLI) and Relation Extraction (RE).

Our study reveals a concrete bias that skews the collected data distribution toward specific patterns. While the effect of instruction examples on collected data is prominent, it is hard to quantify how different the distribution of crowdsourced examples is from the natural distribution of the task. Concretely, to conduct a study that compares the distributions of crowdsourced versus natural complex reasoning questions, datasets of complex natural questions are needed. However, to the best of our knowledge, as of today, no such datasets exist.

In our analysis, we focused on shallow patterns based on word matching, however, it is known that there are other types of biases that are implicit in the text. Exploring these kinds of biases can be an interesting future direction. In addition, our analysis of model performance is based on splitting dataset instances based on the dominant pattern. However, it might be possible that there are more patterns, and the non-pattern subset might include other less frequent patterns. Hence, exploring the effect of different less frequent patterns on model learning can be a future work.

Last, our work studied the effect of instruction bias on widely used generative models (i.e., T5 and BART); it would be valuable to investigate whether our findings hold in encoder-only models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019).

Acknowledgement

We thank Daniel Khashabi, Sewon Min, and Avia Efrat for helpful feedback, and the anonymous reviewers for constructive suggestions. We acknowledge the Research Computing (RC) at Arizona State University (ASU) for providing computing resources for experiments.

References

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2020. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). *arXiv preprint arXiv:2009.11352*.

Anjana Arunkumar, Swaroop Mishra, Bhavdeep Sachdeva, Chitta Baral, and Chris Bryan. 2020. Real-time visual feedback for educative benchmark creation: A human-and-metric-in-the-loop workflow. *NeurIPS 2020 Workshop HAMLETS*.

Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. 2022. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles. Association for Computational Linguistics.

Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2334–2346.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. 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 5925–5932, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Avia Efrat and Omer Levy. 2020. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*.

Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 162–170.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. 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, China. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. 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 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Danula Hettiachchi, Mark Sanderson, Jorge Goncalves, Simo Hosio, Gabriella Kazai, Matthew Lease, Mike Schaekermann, and Emine Yilmaz. 2021. Investigating and mitigating biases in crowdsourced data. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, pages 331–334.

Xiao Hu, Haobo Wang, Anirudh Vegesana, Somesh Dube, Kaiwen Yu, Gore Kao, Shuo-Han Chen, Yung-Hsiang Lu, George K Thiruvathukal, and Ming Yin. 2020. Crowdsourcing detection of sampling biases in image datasets. In *Proceedings of The Web Conference 2020*, pages 2955–2961.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. 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 252–262.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Walter S Lasecki, Mitchell Gordon, Danai Koutra, Malte F Jung, Steven P Dow, and Jeffrey P Bigham. 2014. Glance: Rapidly coding behavioral video with the crowd. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 551–562.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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, Online. Association for Computational Linguistics.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Hong Kong, China. Association for Computational Linguistics.

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

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2021. Variational information bottleneck for effective low-resource fine-tuning. *ICLR*.

Swaroop Mishra and Anjana Arunkumar. 2021. How robust are model rankings: A leaderboard customization approach for equitable evaluation. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 35, pages 13561–13569.

Swaroop Mishra, Anjana Arunkumar, Chris Bryan, and Chitta Baral. 2020a. Our evaluation metric needs an update to encourage generalization. *ArXiv*, abs/2007.06898.

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. 2020b. Dqi: Measuring data quality in nlp. *arXiv preprint arXiv:2005.00816*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *ACL*.

Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. 2015. Deep learning applications and challenges in big data analytics. *Journal of big data*, 2(1):1–21.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks? 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 1221–1235, Online. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

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

Hossein A Rahmani and Jie Yang. 2021. Demographic biases of crowd workers in key opinion leaders finding. *arXiv preprint arXiv:2110.09248*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu. 2019. NumNet: Machine reading comprehension with numerical reasoning. In *Proceedings of EMNLP*.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *arXiv preprint arXiv:2107.12708*.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. 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. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25.

Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar, Samuel R. Bowman, and Yoav Artzi. 2021. Crowdsourcing beyond annotation: Case studies in benchmark data collection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 1–6, Punta Cana, Dominican Republic & Online. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. 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.

Feifei Zheng, Ruoling Tao, Holger R Maier, Linda See, Dragan Savic, Tuqiao Zhang, Qiuwen Chen, Thaine H Assumpção, Pan Yang, Bardia Heidari, et al. 2018. Crowdsourcing methods for data collection in geophysics: State of the art, issues, and future directions. *Reviews of Geophysics*, 56(4):698–740.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China. Association for Computational Linguistics.

A Biases in NLU Benchmarks

Crowdsourcing has been a widely adapted approach to create large scale datasets such as SQuAD 1.1(Rajpurkar et al., 2016, 2018), DROP(Dua et al., 2019), QUOREF(Dasigi et al., 2019) and many more (Najafabadi et al., 2015; Callison-Burch and Dredze, 2010; Lasecki et al., 2014; Zheng et al., 2018; Chang et al., 2017). Many past works investigate different types of bias in crowdsourcing datasets such as cognitive bias (Eickhoff, 2018), annotator bias (Gururangan et al., 2018; Geva et al., 2019), sampling bias (Hu et al., 2020), demographic bias (Rahmani and Yang, 2021) and others (Hettiachchi et al., 2021). Many works on bias in NLU benchmarks focus on biases resulting from the crowdsourcing annotations, and how annotator-specific patterns create biases in data (Geva et al., 2019).

To mitigate the bias, prior works have focused on priming crowdsourcing annotators with minimal information to increase their imagination (Geva et al., 2021; Clark et al., 2020) to avoid recurring patterns. Arunkumar et al. (2020) develops a real time feedback and metric-in-the loop (Mishra et al., 2020b) workflow to educate crowdworkers in controlling dataset biases. Nangia et al. (2021) provides an iterative protocol with expert assessments for crowdsourcing data collection to increase difficulty of instances. (Swayamdipta et al., 2020) introduces dataset map as a model-based tool to characterize and diagnose datasets. Also, Karimi Mahabadi et al. (2020); Mahabadi et al. (2021) propose learning strategies to train neural models, which are more robust to such biases and transfer better to out-of-domain datasets.

In this work, we show that biases exhibited by annotators start from the crowdsourcing instructions designed by dataset creators.

B Dataset Statistics

Tab. 5 describes the statistics of train and evaluation sets of datasets used in our experiments. Here, we can observe that each selected dataset differs in terms of number of training samples, % of instruction patterns, and tasks.

C Pattern Extraction Method

Here, we describe an example to show how we extract the dominant pattern from the crowdsourcing instructions and subsequently identify the same pattern in the dataset. We try to find recurring word patterns such as *"Are you..."*, *"how many points..."*, *"Was... still..."*, *"since... the..."*.

For example, MC-TACO (event duration) has 3 examples in crowdsourcing instructions: (1) how long did Jack play basketball?, (2) how long did he do his homework?, and (3) how long did it take for him to get the Visa? In step (a), we analyze examples manually and find *dominant pattern*. Here, we can see that all examples contain tri-gram pattern, i.e., *"how long did"*. In step (b), we try to generate more possible patterns that are semantically similar to the *dominant pattern* or have a significant word overlap. Here, *"how long did"* can be *"how long was"*, *"how long does"*, etc. (i.e, `How long AUX`). In step (c), we look for all these possible patterns in datasets using simple word-matching techniques.

D Pattern Examples

Tab. 8 provides dataset, instruction patterns and corresponding examples of data instances that exhibit the instruction patterns.

E The Effect of Instruction Examples on Pattern Frequency in Collected Data

To study the effect of bias in instruction examples on collected data, we asked NLP graduate students to write five questions for each of (1) temporal reasoning (event duration) and (2) coreference resolution, based on the crowdsourcing instructions of MC-TACO and QUOREF, respectively. For each task, we conduct two surveys, where the instructions include and do not include any examples.

We collected responses from 10 participants. The dominant patterns of MC-TACO ('how long') and QUOREF ('What is the name') only contribute to 38% and 8% of our collected data where examples are not given, in contrast to 68% (↑79%) and 32% (↑300%) in collected data where examples are given. This indicates that crowdsourcing examples bias crowdworkers to follow certain patterns, whereas showing no examples increases the creativity of crowdworkers.

In addition, our collected responses where examples are not given contain 10 and 9 unique patterns for MC-TACO (event duration) and QUOREF respectively, in contrast to only 4 and 5 unique patterns in collected data where examples are given. Our finding shows that there is substantial linguistic diversity associated with the NLP tasks, unlike

Dataset	Train						Test					
	$\mathcal{S}_{\text{train}}$	$\mathcal{S}^{p}_{\text{train}}$		$\mathcal{S}^{-p}_{\text{train}}$			$\mathcal{S}_{\text{test}}$	$\mathcal{S}^{p}_{\text{test}}$		$\mathcal{S}^{-p}_{\text{test}}$		
CLARIQ	8566	7286	85.1%	1280	14.9%		4499	4006	89%	493	11%	
DROP	77409	48422	62.5%	28987	37.5%		9536	5960	62.5%	3576	37.3%	
MULTIRC	5131	1972	38.4%	3159	61.6%		953	395	41.5%	558	58.6%	
PIQA	17171	7508	43.7%	9663	56.3%		3268	1401	42.9%	1867	57.1%	
QUOREF	19399	11052	57%	8347	43%		2418	1451	60%	967	40%	
ROPES	1412	1046	74.1%	366	25.9%		203	42	20.7%	161	79.3%	
SCIQA	11679	9765	83.61%	1914	16.4%		1000	845	84.5%	155	15.5%	
Total	140767	87051	61.8%	53716	38.2%		21877	14100	64.5%	7777	35.6%	

Table 5: Statistics of number of train and test examples with and without instruction patterns. $\mathcal{S}_{\text{train}}$: set of examples in train set, $\mathcal{S}^{p}_{\text{train}}$: set of examples in train set with instruction pattern, $\mathcal{S}^{-p}_{\text{train}}$: set of examples in train set without instruction pattern, $\mathcal{S}_{\text{test}}$: set of examples in test set, $\mathcal{S}^{p}_{\text{test}}$: set of examples in test set with instruction pattern, $\mathcal{S}^{-p}_{\text{test}}$: set of examples in test set without instruction pattern.

	Base			Large		
	$\mathcal{S}^{p}_{\text{test}}$	$\mathcal{S}^{-p}_{\text{test}}$		$\mathcal{S}^{p}_{\text{test}}$	$\mathcal{S}^{-p}_{\text{test}}$	
CLARIQ	30	26.2	12.7% ↓	29.2	25.6	12.3% ↓
MULTIRC	27.9	15.1	45.9% ↓	31.1	21	32.5% ↓
PIQA	21.9	15.3	30.1% ↓	23	16.1	30% ↓
QUOREF	78.8	45.8	41.9% ↓	87	58.6	32.6% ↓
ROPES	43.9	33.1	24.6% ↓	47.2	39.1	17.2% ↓
SCIQA	76.8	66.1	13.9% ↓	77.5	69.6	10.2% ↓
Average	46.6	33.6	27.9% ↓	49.2	38.3	22.2% ↓

Table 6: Performance of BART models on $\mathcal{S}^{p}_{\text{test}}$ vs. $\mathcal{S}^{-p}_{\text{test}}$ of models trained on data instances containing instruction patterns ($\mathcal{S}^{p}_{\text{train}}$).

	Base			Large		
	$\mathcal{S}^{p}_{\text{test}}$	$\mathcal{S}^{-p}_{\text{test}}$		$\mathcal{S}^{p}_{\text{test}}$	$\mathcal{S}^{-p}_{\text{test}}$	
CLARIQ	29.8	26	12.8% ↓	29.4	26.3	10.5% ↓
MULTIRC	28.5	29.8	4.6% ↑	41.5	33.9	18.3% ↓
PIQA	22.3	20.5	8.1% ↓	23.1	21.6	6.5% ↓
QUOREF	80.5	61.2	24% ↓	87.8	73.4	16.4% ↓
ROPES	44	44.1	0.2% ↑	47.9	46.5	2.9% ↓
SCIQA	76.5	70.6	7.7% ↓	52.2	50.8	2.7% ↓
Average	46.9	42	10.5% ↓	47	42.1	10.4% ↓

Table 7: Performance of BART models on $\mathcal{S}^{p}_{\text{test}}$ vs. $\mathcal{S}^{-p}_{\text{test}}$ of models trained on $\mathcal{S}_{\text{train}}$.

the patterns covered in instruction examples that get propagated to corresponding datasets.

The task instructions and collected annotations are available at https://github.com/Mihir3009/instruction-bias/blob/main/SURVEY.md.

F Additional Results

Tab. 6 and Tab. 7 show the performance of BART on $\mathcal{S}^{p}_{\text{test}}$ and $\mathcal{S}^{-p}_{\text{test}}$ when training only on examples with instruction patterns and the full training set, respectively. From Tab. 6, there are large performance gaps reaching 45.9% in MULTIRC and $>$ 20% in both base and large models for QUOREF, and PIQA. Overall, the average performance across all datasets is 27.9% and 22.2% higher on $\mathcal{S}^{p}_{\text{test}}$ for the base and large models, respectively. This indicates that both base and large models often fail to generalize beyond instruction patterns.

From Tab. 7, we see that the average performance across all datasets is higher on examples that exhibit instruction patterns by $\sim$ 10.5% for both base and large models. From the results, we can conclude that the model's performance is overestimated by instruction bias.

Dataset	Pattern	Examples
CLARIQ	[Are|Would|Do] you	Are you looking for a specific web site?
		What kind of train are you looking for?
		Do you want to watch news videos or read the news?
		Would you like the location of the ritz carlton lake las vegas?
COSMOSQA	What AUX	What may happen after the young man makes his call?
		What might happen if you have him for the whole day?
		What's a possible reason the writer doesn't look disabled on the outside?
DROP	How many [field goals | years | yards | points | touchdowns]	How many touchdowns did Jones have?
		How many field goals did Kris Brown kick
		How many yards was the longest touchdown of the game?
		After Akers 32-yard field goal, how many points behind was Washington?
HOTPOTQA	[in|of|from|_] [Which|What] AUX	Which franchise was founded in 1978, Chuck E. Cheese's or Jet's Pizza?
		Busan, in the area surrounding the mountain of Geumjeongsan, is the second most populated city in which country?
		What is the name of the third album from singer Selena Quintanilla-Pérez?
MC-TACO	How long AUX	How long was his mother ill?
	What AUX	What did the government decide after the 9/11 attack?
	How often AUX	How often would one family be able to do something like this?
	AUX... [still|always]	Will electronic espionage always be happening in the U.S.?
		Is she still gone?
	When did / What time	What time did the planes crash into the World Trade Center?
		When did Durer die?
MULTIRC	What AUX	What was Poe's first published work?
		What is the full name of the person described?
		What kind of career does Christie Brinkley have?
PIQA	How [do|can]	How do I make orange icing if I have store-bought white frosting?
		How can I make popsicles for dogs?
		Are you nervous about giving a speech or doing something? How can you calm yourself?
QUOREF	What AUX the [full | real | first |last] name	What is the first name of the person who purchases a revolver?
		What is the full name of the person who is calmly asked to leave?
		What was the name of the house where Appleton Water Tower was built?
		What is the last name of the person who convinces the girls to help him look for the treasure?
ROPES	Which AUX	Which area would be less likely to experience a drought and have better chance at a new growth?
		Which hair spray brand should Greg buy to be environmentally friendly?
		Which markalong was produced asexually?
SCIQA	What AUX	What are by far the most common type of invertebrate?
		What do waves deposit to form sandbars and barrier islands?
		What is the term for the total kinetic energy of moving particles of matter?
WINO-GRANDE	[because | so | while | since | but] ... the	The dog didn't like its collar but was okay with its leash because the _ was loose on it.
		Hunter took Benjamin's clothes to the laundromat, since _ had the day off that day.
		James sang his song at the top of his voice so as to be heard over the noise but the _ is louder.

Table 8: Examples of data instances from original dataset that contain instruction patterns. AUX $\in$ {am, is, are, was, were, has, have, had, do, does, did, will, would, can, could, may, might, shall, should, must}. _ : <blank>.