Does Self-Rationalization Improve Robustness to Spurious Correlations?

Alexis Ross[†*]   **Matthew E. Peters**[‡]   **Ana Marasović**[§*]

[†]Massachusetts Institute of Technology, Cambridge, MA, USA
[‡]Allen Institute for AI, Seattle, WA, USA
[§]University of Utah, Salt Lake City, UT, USA
alexisro@mit.edu   matthewp@allenai.org   ana.marasovic@utah.edu

Abstract

Rationalization is fundamental to human reasoning and learning. NLP models trained to produce rationales along with predictions, called self-rationalization models, have been investigated for their interpretability and utility to end-users. However, the extent to which training with human-written rationales facilitates learning remains an under-explored question. We ask whether training models to self-rationalize can aid in their learning to solve tasks *for the right reasons*. Specifically, we evaluate how training self-rationalization models with free-text rationales affects robustness to spurious correlations in fine-tuned encoder-decoder and decoder-only models of six different sizes. We evaluate robustness to spurious correlations by measuring performance on 1) manually annotated challenge datasets and 2) subsets of original test sets where reliance on spurious correlations would fail to produce correct answers. We find that while self-rationalization can improve robustness to spurious correlations in low-resource settings, it tends to hurt robustness in higher-resource settings. Furthermore, these effects depend on model family and size, as well as on rationale content. Together, our results suggest that explainability can come at the cost of robustness; thus, appropriate care should be taken when training self-rationalizing models with the goal of creating more trustworthy models.

1 Introduction

Rationalization—the process of explaining the reasoning used to come to a particular decision—plays a pivotal role in human inference and learning (Lombrozo, 2016). For these reasons, there has been a growing interest in producing NLP models that can output rationales[1] for their predictions. Models that output such rationales have multiple benefits: First, they are more interpretable and easier to interact with for end-users than non-rationalizing models (Alvarez-Melis and Jaakkola, 2018). Second, such intermediate rationalization can offer learning benefits, such as achieving comparable performance with less data and improving out-of-distribution generalization (Nye et al., 2021; Wei et al., 2022; Zelikman et al., 2022).

However, the question of whether training models to rationalize can help them learn how to solve tasks *for the right reasons* remains open. In particular, rationales encode information about the underlying reasoning humans use to reach answers, which raises the question: Does incorporating such rationales into training allow models to rely on human-aligned reasoning rather than spurious feature interactions? If so, training with rationales could offer a pathway towards creating more robust, trustworthy, or cognitively plausible models.

In this work, we explore this question by empirically investigating whether training models with human-written rationales can help make them more robust to spurious correlations in data. We analyze a class of models called self-rationalization models—which jointly output free-text rationales along with predictions—and focus specifically on the fine-tuning setting, in which prior work has found reliance on spurious correlations to emerge (Utama et al., 2021).

We evaluate six models of varying architectures and sizes across two tasks, natural language inference and commonsense question answering. Our main results are as follows:

1. While the effects of training with rationales are model- and task-specific, when it improves robustness to spurious correlations, it tends

---

[1]Prior work has used the terms "explanation" and "rationale" interchangeably. In this work, we use the word "rationale" for consistency with "self-rationalization" models.

to be in lower-resource settings. In higher-resource settings, training with rationales can hurt robustness (§4.1).

2. Within model families, larger models benefit more in robustness from rationales (§4.2).

3. The effects of self-rationalization on robustness are not fully explained by its effects on in-domain task performance (§4.3).

4. The content of rationales used during training influences both task performance and robustness to spurious correlations (§4.4).

Our results suggest that straightforward self-rationalization training does not always facilitate learning to solve a task for the right reasons. Instead, the effects of self-rationalization on robustness to spurious correlations depend on a multitude of factors. Thus, appropriate care should be taken when training models to self-rationalize for the goal of creating trustworthy models.

2   Related Work

Learning to rationalize   Two classes of approaches to producing models that can rationalize their predictions include *self-rationalization* models,[2] which are fully differentiable and output free-text rationales along with task predictions, and *pipeline* models, which consist of two components—one that produces rationales, and a second that makes predictions from those rationales (Wiegreffe et al., 2021).[3] Such methods are typically evaluated by the *faithfulness* and *plausibility* of their rationales, where faithfulness represents the extent to which a model actually relied on the rationale in making its prediction, and plausibility indicates human judgment of how well the rationale explains the output (DeYoung et al., 2020).

In contrast to these works which aim to improve model interpretability through new methods for rationalizing models, we ask to what extent existing methods affect model robustness to spurious correlations. We conduct our analysis on self-rationalization models, which have been found to achieve better task performance and produce higher-quality rationales than do pipeline models (Wiegreffe et al., 2021; Camburu et al., 2018).

---

[2]Such approaches have also been referred to as *explain-then-predict* (Camburu et al., 2018) and *rationalize-then-predict* (Chen et al., 2022) models.

[3]See Wiegreffe et al. (2021) for a detailed discussion of pipeline and self-rationalization approaches to rationalization.

Learning from rationales   Recent work has explored the utility of rationales for improving end-task performance in in-context learning (Wei et al., 2022; Lampinen et al., 2022; Ye and Durrett, 2022) as well as in fine-tuning (Zaidan et al., 2007; Hancock et al., 2018; Camburu et al., 2018; Narang et al., 2020; Hase and Bansal, 2021; Nye et al., 2021; Zhao and Vydiswaran, 2021). Previous work has shown that training with both human-annotated rationales (Rajani et al., 2019) and rationales generated by language models (Paranjape et al., 2021) can increase in-domain task performance, particularly in low-resource settings (Bhat et al., 2021; Pruthi et al., 2022; Zelikman et al., 2022). Unlike these prior works, which study how training with rationales affects in-domain, end-task performance, we focus specifically on evaluating impact on robustness to spurious correlations.

Improving robustness with rationales   Most closely related are recent works that study how training with rationales affects model robustness. Stacey et al. (2022) propose a method of supervising attention weights with extractive rationales and show that this method leads to both in-distribution and out-of-distribution improvements for natural language inference. Schuster et al. (2021) find that training with contrastive extractive rationales improves robustness as measured by performance on adversarial evaluation sets. Concurrent work by Chen et al. (2022) investigates to what extent training models to extract rationales through pipelines improves their robustness to adversarial attacks.

In contrast to all three of these works, we focus on freeform rationales instead of extractive rationales and explore the impact of amount of training data on robustness. In contrast to Schuster et al. (2021) and Chen et al. (2022), we analyze self-rationalization models instead of pipeline models and measure robustness to spurious correlations, rather than robustness to adversarial attacks. While Stacey et al. (2022) evaluate robustness to spurious correlations for natural language inference with some of the same test sets, they work with masked language models and evaluate the effect of supervising model attention with rationales; in contrast, we work with encoder-decoder and decoder-only models of varying sizes and evaluate the effect of outputting rationales along with predictions. In addition, their analysis is limited to natural language inference, for which evaluation datasets targeting robustness exist; in contrast, we also experiment

with commonsense question answering through new methods for evaluating robustness. In §4.1, we discuss the variance in results across different tasks and highlight the importance of cross-task evaluation.

3 Experiments

3.1 Experimental Set-Up

Models We experiment with encoder-decoder and decoder-only models of varying sizes ranging from 140 to 774 million parameters, as shown in Figures 1 and 2. Our encoder-decoder models build on pretrained **T5** (Raffel et al., 2020) and **BART** models (Lewis et al., 2020), and our decoder-only models build on pretrained **GPT2** (Radford et al., 2019) models. Our T5 models build specifically on the versions trained for an additional 100K steps on the language modeling objective after pretraining (Lester et al., 2021), as we aim to measure how the amount of training data impacts results, and the default T5 models have already been fine-tuned on the full SNLI training dataset.[4]

Tasks We evaluate self-rationalization models on two tasks—**natural language inference** (NLI), and **commonsense question answering** (CQA)—for which human-annotated rationales already exist. For NLI, we train task models on SNLI (Bowman et al., 2015) and obtain rationales from ESNLI (Camburu et al., 2018). For CQA, we train task models on CQA (Talmor et al., 2019) and obtain rationales from ECQA (Aggarwal et al., 2021). Examples of inputs and outputs for both tasks are shown in Table 2. For CQA, unless otherwise specified, we train on the "positive" freeform rationales in ECQA, which explain why the gold answer is the correct answer for a given question. In §4.4, we explore the impact of training with the different forms of rationales shown in Table 2.

Rationales For each task, we compare a baseline model trained solely to predict task labels with models trained to also self-rationalize. All self-rationalization models are trained to generate a rationale following the task label, as previous work has found that outputting rationales conditioned on labels leads to better performance than outputting

labels conditioned on rationales in the fine-tuning setting (Schuff et al., 2021).

Data We experiment with different numbers of training examples n, as we seek to understood how training data size influences the impact of self-rationalization training on robustness to spurious correlations. We experiment with $n \in \{$1K, 2.5K, 5K, 10K, 50K, 100K$\}$ for NLI and $n \in \{$1K, 5K, 7598$\}$ for CQA.[5] For each training data amount n, we create validation data for checkpointing models by randomly sampling $n/2$ instances from the original *task-only* validation dataset, such that we perform model selection based on task performance across baseline and self-rationalization models. For self-rationalization models, we create training data by concatenating original task-only training input-output pairs with their rationale-extended counterparts, such that we have $2n$ training inputs obtained from n original instances.[6]

Training For each amount of training data n, we report the average difference between task-only and self-rationalization models across multiple random seeds (5 for NLI and 10 for CQA).[7] For one random seed in each evaluation setting (where a setting is determined by the task, model family, model size, whether rationales are used, and amount of training data), we tune the learning rate from possible values $[1e-5, 3e-5, 5e-5]$ and use the best-performing learning rate for other random seeds in the same setting. We train with fixed batch size 64 and linear learning rate scheduler using Adafactor until accuracy on the validation data stops decreasing, or for a maximum of 50 epochs. We use patience values of 10 for $n < 10K$, 5 for $n >= 10K$, and 3 for $n >= 50K$.

Evaluation We decode predictions using greedy decoding and evaluate accuracy using exact match with gold labels. We evaluate robustness to spurious correlations by measuring performance on 1) manually annotated challenge datasets and 2) subsets of original test sets where reliance on spurious

---

[4]For example, when experimenting with T5-BASE, we work specifically with `t5-base-lm-adapt` available in huggingface at `https://huggingface.co/google/t5-base-lm-adapt`.

[5]The total size of original training datasets are 549,339 for SNLI and 7,598 for CQA.

[6]In initial experiments, we find that this leads to better performance/robustness measures than only using the n input-outputs for self-rationalization; we hypothesize that without including the original task-only inputs as well, self-rationalization models may be overfitting to the rationale generation part of the training objective.

[7]We experiment with more seeds for CQA because we have fewer metrics/evaluation datasets to measure robustness for CQA, and so it is harder to disentangle real effects from noise.

correlations would fail to produce correct answers. Both methods are discussed below in §3.2.

3.2 Evaluating Reliance on Spurious Features

Out-of-domain challenge datasets Our first method of evaluating reliance on spurious correlations leverages out-of-domain evaluation sets designed by experts to test for reliance on spurious features. For NLI, we evaluate on HANS (McCoy et al., 2019) and CAD (Kaushik et al., 2021). **HANS** is a controlled evaluation dataset that tests for reliance on surface-level syntactic biases present in SNLI. **CAD** is an evaluation dataset with human-annotated edits to inputs that change entailment labels. To the best of our knowledge, such evaluation datasets do not exist for CQA.

CQA		NLI		
Feature	**z**	**Feature**	**z**	
fountain	3.50	lex-overlap ≥ 0.8	140.16	(e)
music	3.18	for	93.86	(n)
welcome	3.01	to	83.30	(n)
atlas	3.00	sleeping	80.68	(c)
satisfied	3.00	there	78.11	(e)
hard	2.98	outside	77.64	(e)
stage	2.86	nobody	68.44	(c)
tale	2.86	outdoors	65.17	(e)
amusement	2.65	no	52.58	(c)
feel	2.65	cat	50.72	(c)

Table 1: Features found to have statistically significant correlations with labels in training datasets using Gardner et al. (2021)'s framework. Details in §3.2.

"Hard" subsets of original evaluation data To directly test for reliance on spurious correlations without introducing additional domain shifts, we also subset the original task test sets into subsets of varying difficulty, where difficulty is measured by the success of spurious heuristics: "Easy" subsets include instances for which heuristics that build on spurious correlations in training data would lead to correct predictions, and "hard" subsets include instances where such spurious heuristics would fail.

To create these "easy" and "hard" subsets, we build on the statistical framework for uncovering dataset-level artifacts introduced by Gardner et al. (2021). Specifically, we measure correlation between features and outputs across the CQA and SNLI training datasets and consider as artifacts any features showing statistically significant correlation, *i.e.,* with z-statistic > 2.

For SNLI, we consider tokens in inputs as features, as well as lexical overlap between premise and hypothesis. Following previous work (Wu et al., 2022), we consider an input to have high lexical overlap if the ratio of tokens in the hypothesis that are also present in the premise is at least 0.8. We use classification labels as outputs. For CQA, the feature and output spaces are less clearly defined, as it contains different output choices for each input. We take tokens in answer choices to be features and whether or not those tokens are present in the gold answers as outputs. To remove features that are very frequent or infrequent, we filter features that appear less than 10 or more than 200K times for SNLI and less than 5 or more than 10K times for CQA. Table 1 displays the 10 features with highest z-statistics for the CQA and SNLI training sets.[8]

We subset the original CQA and NLI test sets based on whether artifacts appear with the same output they showed statistically significant correlations with in the training datasets. **TEST-HARD** contains instances for which relying solely on artifacts to make predictions would fail to produce correct predictions (*i.e.,* artifacts appear with a different output than they are correlated with), and **TEST-EASY** contains instances for which relying on artifacts would lead to correct predictions.[9] For example, a CQA test instance for which an incorrect answer choice had token "fountain" would be considered "hard," as "fountain" has statistically significant correlation with being in the correct answer choice (Table 1). The sizes of TEST-EASY and TEST-HARD are 76/333 respectively for NLI and 82/372 for CQA. In addition to reporting performance values for these subsets, we measure the *spread* in performance on hard vs. easy subsets, *i.e.,* TEST-EASY − TEST-HARD, which we refer to as Δ **TEST-SUBSETS**. We take a lower value of Δ TEST-SUBSETS to indicate less reliance on artifacts.[10] For NLI, we also evaluate on TEST-

---

[8]We note that the z-statistics for SNLI artifacts are much higher than for CQA; this finding aligns with prior work showing that SNLI contains many artifacts (Poliak et al., 2018; Gururangan et al., 2018; Wallace et al., 2019). SNLI train has 3,496 total artifacts, and CQA train has 43.

[9]If an instance (x, y_i) contains *both* artifact(s) that show statistically significant correlation with y_i in the training data *and* other artifact(s) that show statistically significant correlation with $y \neq y_i$ in the training data, we exclude this instance from both TEST-EASY and TEST-HARD.

[10]We observe that TEST-EASY and TEST-HARD in fact have the expected difficulties as measured by accuracy values of different models. In particular, as shown in the last rows of Figures 1 and 2, the large Δ TEST-SUBSETS values, annotated in gray, indicate that baseline models perform noticeably

Figure 1: Effect of self-rationalization for NLI across six models (columns) and varying amounts of training data (x axis). Bar heights show mean differences between baseline task-only models and self-rationalization models for: task performance (row 1), performance on manually annotated challenge datasets (rows 2-3), performance on hard subsets of original SNLI evaluation data (rows 4-5), and Δ TEST-SUBSETS (row 6). Baseline values are shown in gray. Error bars indicate standard errors of the means. Green/red bars indicate improvement/degradation in robustness; gray bars indicate error bars intersecting 0.

HYP, a subset of the SNLI test set for which a hypothesis-only classifier was found to give incorrect predictions (Gururangan et al., 2018).[11]

4 Results

Figures 1 and 2 show, for NLI and CQA respectively, the effects of self-rationalization across multiple random seeds. Plotted are mean differences between self-rationalization models and baseline

worse on TEST-HARD than on TEST-EASY for both NLI and CQA. In addition, baseline accuracies on TEST-HARD are notably worse than accuracies on the full test sets (row 1) for NLI. While this latter trend does not hold as consistently for CQA, we observe that the baseline accuracies on original test sets are lower for CQA than for NLI. Thus, we hypothesize that for CQA, the relative lack of drop in performance on TEST-HARD compared to original test sets can be explained by the fact that ECQA contains fewer artifacts and so original test sets are already "hard" for CQA models in the sense of prevalence of artifacts to be exploited.

[11]We do not evaluate on the analogous "easy" counterpart for TEST-HYP, i.e., the subset for which a hypothesis-only classifier succeeds, as it would require re-training a hypothesis-only classifier; instead, we evaluate only on the TEST-HYP subset released by Gururangan et al. (2018).

task-only models (i.e., self-rationalization – baseline) across six models (columns) and varying amounts of training data (x axis). Improvements on TEST (row 1) reflect in-domain, task improvements, while improvements on other metrics (rows > 1) indicate robustness improvements.

4.1 Main Results

As shown in Figure 1, under our evaluation of robustness to spurious correlations, we observe that self-rationalization improves the robustness of BART- and GPT2-based NLI models in lower resource data settings. In higher resource settings, we observe a degradation in some robustness metrics, namely performance on TEST-HYP & TEST-HARD and Δ TEST-SUBSETS for all models except BART-LARGE. For BART-BASE, this degradation in higher-resource settings is also seen for performance on HANS. The T5 models (T5-BASE & T5-LARGE) show more mixed results: While self-rationalization hurts performance on
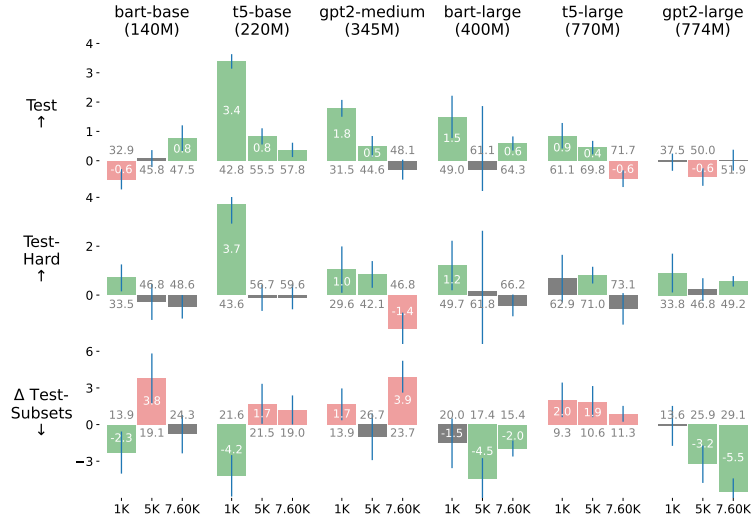
Figure 2: Effect of self-rationalization for CQA across six models (columns) and varying amounts of training data (x axis). Bar heights show mean differences between baseline task-only models and self-rationalization models for: task performance (row 1), performance on the TEST-HARD subset of original CQA evaluation data (rows 2-3), and Δ TEST-SUBSETS (row 4). Baseline values are shown in gray. Error bars indicate standard errors of the means. Green/red bars indicate improvement/degradation in robustness; gray bars indicate error bars intersecting 0.

HANS for both T5-BASE and T5-LARGE in all data regimes, it improves performance on some metrics, *i.e.,* Δ TEST-SUBSETS in higher-resource settings (n>=5k) for T5-LARGE.[12]

For CQA (Figure 2), results are more mixed, and they depend on model properties, *i.e.,* architecture and size, as well as size of the training data. For BART and GPT2 models of size LARGE, training with rationales generally leads to improvements. For models smaller than size LARGE, as well as both T5 models, the effect of training with rationales depends on the amount of training data, but rationales tend to hurt robustness in higher-resource settings (7.6K training examples) for these models.

These general trends are similar to those for NLI, with more improvements from self-rationalization in lower-resource settings and some degradation in higher-resource settings. However, unlike for NLI, the results are not always monotonic in the amount of training data, particularly for BART-BASE and GPT2-MEDIUM on Δ TEST-SUBSETS. In addition, for GPT2-LARGE, results on Δ TEST-SUBSETS improve with increasing data size, opposite to the general trend. Furthermore, improvements in TEST-HARD are similar to standard errors, except for T5-BASE and n=1K, suggesting that even in low-resource settings, self-rationalization does not notably improve robustness for CQA.

The varied results for CQA and lack of consistency between NLI and CQA may be influenced by the differing numbers of artifacts in the datasets; in particular, perhaps self-rationalization training has a larger effect on robustness to spurious correlations when there are more spurious correlations in the training data (as in SNLI but not ECQA). We leave it to future work to investigate the impact of artifacts in training data on effect of rationales. The differences between NLI and CQA also suggest that evaluations solely based on NLI may not cleanly transfer to other tasks; this finding provides further evidence that the benefits of rationales are task-dependent (Carton et al., 2020; Palaskar et al., 2022) and that evaluations on one task such as NLI alone are not comprehensive enough to draw general conclusions about the utility of rationales.

4.2 Effect of Model Size

For NLI, for the GPT2 and BART models, we find that increasing model size leads to increasing gains in robustness: Self-rationalization leads to larger improvements in robustness for BART-LARGE than for BART-BASE, and similarly for GPT2-LARGE and GPT2-MEDIUM (except for when n=2.5K);

---

[12]One distinct property of T5 models is that they were pre-trained with a denoising objective and then adapted with a language modeling (LM) objective, while BART was pretrained only with denoising and GPT2 only with LM. Thus, we speculate that an explanation for the difference in results from the T5 models could be that the objectives used to pretrain a model before fine-tuning may influence how self-rationalization affects robustness to spurious correlations, but why exactly the objectives may have such an effect remains unclear.

	NLI input:
	snli hypothesis: The family is sitting down for dinner. premise: A couple play in the tide with their young son.

SNLI	contradiction
ESNLI	contradiction explanation: The family cannot simultaneously be playing in the tide and sitting down to dinner.

	CQA input:
	Where would you get a dog if you do not have one? (A) pet store (B) outside (C) neighbor's house (D) park (E) rug

CQA	pet store
ECQA Positive	pet store explanation: Pet store is a retail business which sells different kinds of animals, variety of animal supplies and pet accessories to the public. We would get a dog at a pet store, if we do not have one.
ECQA Negative	pet store explanation: Outside means not in the living or working area. Those found outside in streets are street dogs and are not usually a pet dog. Neighbor's house is the man who lives in the house next door to your house is your neighbor and the house in which he lives is your neighbor's house. The one at neighbor's house is his pet dog and we can't get it even though we don't have one. Park is a large public garden or area of land used for recreation. Dogs found at park are street dogs not suitable for being a pet. Rug is a floor covering of thick woven material or animal skin, typically not extending over the entire floor. Rug is a floor covering and not a pet shop where we can get a dog.
ECQA Freeflow	pet store explanation: Pet store is a retail business which sells different kinds of animals, variety of animal supplies and pet accessories to the public. We would get a dog at a pet store if we do not have one. Those found outside in streets are street dogs and are not usually a pet dog. The one at neighbor's house is his pet dog and we can't get it even though we don't have one. Dogs found at park are street dogs not suitable for being a pet. Rug is a floor covering and not a pet shop where we can get a dog.

Table 2: Examples of inputs and outputs used for training baseline and self-rationalization models.

	TEST ↑		CAD ↑		HANS ↑		TEST HYP ↑		TEST HARD ↑		Δ TEST SUBSETS ↓	
1K												
no rationales	84.62	±0.31	71.20	±0.27	50.18	±0.08	73.68	±0.40	73.82	±0.85	14.70	±1.50
original	+1.05	±0.21	+1.75	±0.34	+0.44	±0.19	+1.56	±0.40	+2.04	±0.91	-2.04	±1.93
shuffled	-0.59	±0.37	-2.11	±0.45	+0.86	±0.52	-1.81	±0.50	-2.51	±1.60	+4.15	±2.85

Table 3: Comparison between training BART-LARGE on 1K training instances with original rationales in ESNLI vs. shuffled rationales across 5 random seeds. We report means, as well as standard errors of the means.

furthermore, we do not observe the same degradation in robustness for BART-LARGE in higher-resource settings that we observe for BART-BASE. For the T5 models, self-rationalization generally leads to less degradation in robustness for T5-LARGE than for T5-BASE. For CQA, we observe a similar trend: self-rationalization generally leads to larger improvements in robustness for BART-LARGE than for BART-BASE, for GPT2-LARGE than for GPT2-MEDIUM, and for T5-LARGE than for T5-BASE (except for when n=1K).

Thus, our results suggest that, within model families, increasing model size may improve effects on robustness from training with rationales. Previous work has shown that rationales improve in-domain performance only for larger models, in both fine-tuning (Nye et al., 2021) and in-context learning (Wei et al., 2022; Lampinen et al., 2022); our results can be seen as an extension of this finding to the effects of training with rationales on *robustness*. It is worth noting that the trends we observe

appear to be specific to model families, *i.e.,* increasing model size has no noticeable effect when not conditioning on model family.

4.3 Correlation between robustness metrics

To determine how results on different robustness metrics relate to each other, we compute their correlations. These correlations should indicate how much insight we can get into the overall impact of self-rationalization on a model's robustness by only looking at select metrics. For each pair of metrics in Figure 1, we aggregate the differences in performance between baseline and self-rationalization performance on those metrics in all evaluation settings (*e.g.,* model type, training data size), and compute the Pearson Correlation of these values.

As shown in Figure 3, results on the "hard" subsets of original test data (TEST-HARD & TEST-HYP) are overall correlated with the results on out-of-domain challenge datasets; the lowest correlation we observe for these subsets is between TEST-
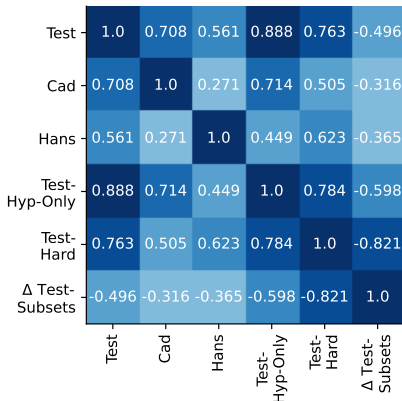
Figure 3: Pearson correlations between results on pairs of evaluation metrics. Each cell color is determined by absolute value of the correlation coefficient.

HYP and HANS, with Pearson coefficient 0.449. Furthermore, CAD and HANS, the manually annotated challenge sets, show low correlation with each other, with a Pearson coefficient of 0.271, suggesting that out-of-domain performance does not straightforwardly reflect all aspects of robustness. We also observe that in-domain test performance is not always highly correlated with robustness metrics, with Pearson coefficient magnitudes as low as 0.496; this result suggests that difference in test performance is not entirely predictive of the effect of self-rationalization on robustness. In other words, training with rationales has effects on robustness that go beyond facilitating or hurting in-domain task performance.

4.4 Effect of rationale content

Shuffled explanations One hypothesis for why training models to output rationales in addition to predictions may improve robustness is that it serves as a form of regularization; under this hypothesis, training to output even rationales with low explanatory power might improve robustness to spurious correlations by reducing overfitting.

To determine to what extent rationale *content* influences effects on robustness, we experiment with shuffling rationales during training such that the rationale for a given input no longer explains that input. Results from training BART-LARGE with shuffled rationales for NLI are shown in Table 3. We also report results for BART-BASE, GPT2-MEDIUM, and T5-LARGE, which follow a similar trend, in Table 5 in the Appendix. We find that, as expected, training with shuffled rationales leads to worse robustness compared to training with

original rationales, except on HANS.

Different ECQA rationales We also experiment with training BART-LARGE with the different rationale types in the ECQA dataset, depicted in Table 2. Results for BART-BASE, GPT2-MEDIUM, and T5-LARGE, which follow a similar trend, are shown in Table 6 in the Appendix.

"Positive" rationales explain why the gold answer is correct for a given question, "negative" rationales explain why other choices are incorrect, and "freeflow" rationales combine positive and negative rationales into a coherent and free-flowing paragraph and thus constitute freeform contrastive rationales.[13] As shown in Table 4, training with 1K positive rationales improves performance on both TEST & TEST-HARD and decreases Δ TEST-SUBSETS. In contrast, training with 1K negative or freeflow rationales hurts performance on TEST & TEST-HARD. We also observe that training with freeflow rationales generally leads to worse results than positive rationales and better results than negative rationales. In contrast to prior findings on the benefits of contrastive rationales (Paranjape et al., 2021; Schuster et al., 2021), our results suggest that contrastive rationales do not always provide more learning benefits than non-contrastive rationales, given that training with freeflow rationales hurts robustness compare to the non-contrastive positive rationales.

A possible explanation for the differences in effects from training with these different rationale types is their varying lengths. As shown in Table 2, negative and freeflow rationales are longer than positive rationales.[14] To rule out this explanation, we also train BART-LARGE with length-controlled negative and freeflow rationales, in which we truncate their lengths to 96 tokens, the maximum length used to train with positive rationales. As shown in Table 4, we still observe degradation in both task performance and robustness when using negative or freeflow rationales rather than positive rationales. These consistent results suggest that rationale content, rather than length, indeed influences learning.

Another possible explanation for these varied effects is that the topical relevance of rationales to gold labels may influence their utility in training. Positive rationales, as explanations of gold

---

[13]*Contrastive* explanations explain why answers are correct compared to alternative (incorrect) answers (Miller, 2019; Ross et al., 2021; Jacovi et al., 2021b).

[14]The mean number of rationale tokens are 60, 62, and 29 for negative, freeflow, and positive rationales respectively.

	TEST ↑		TEST HARD ↑		Δ TEST SUBSETS ↓	
no rationales	48.3	±0.8	49.0	±1.0	19.9	±1.8
positive	+2.2	±1.3	+2.2	±1.0	-1.9	±3.0
freeflow	-2.3	±1.1	-1.7	±1.2	-2.4	±2.5
freeflow*	-3.3	±0.4	-1.1	±2.7	+2.2	±4.3
negative	-4.1	±1.2	-3.7	±1.8	-1.4	±2.1
negative*	-5.3	±0.3	-1.7	±0.5	-1.6	±2.0

Table 4: Effect of training BART-LARGE with different types of rationales in ECQA. Blue/red cells indicate improvement/worsening in performance compared to the baseline (no rationales, row 1). * indicates length-controlled rationales, *i.e.,* truncation of the negative and freeform rationales to have the same length as the positive rationales. We report means across 5 random seeds, as well as standard errors of the means.

answers, are more topically related to gold answers than negative rationales, while freeflow rationales have topical relevance between those of positive and negative rationales. We observe that the effects of training with these rationale types align with their levels of topical relevance. Future work can further explore how properties like topical relevance influence the utility of rationales.

5 Conclusion

We investigate to what extent training models to rationalize their predictions affects their robustness to spurious correlations. We experiment with encoder-decoder and decoder-only models ranging in size from 140 to 774 million parameters across two tasks—natural language inference and commonsense question-answering—and measure reliance on spurious correlations through both manually-annotated, out-of-domain challenge sets and challenging in-domain subsets of original test sets. We find that the effects of self-rationalization are model- and task-specific: While self-rationalization can improve robustness to spurious correlations in lower-resource settings for some models and tasks, it tends to exacerbate reliance on spurious correlations in higher-resource settings. Furthermore, larger models tend to benefit more from rationales, and rationale content influences rationale utility in improving robustness.

The variability of our results suggests that, despite the appeal of self-rationalization models for increasing model trustworthiness by facilitating debugging and interaction with end-users (Jacovi

et al., 2021a), training models to self-rationalize can have the unintended effect of increasing reliance on spurious features and biases, thereby decreasing the models' trustworthiness. Thus, appropriate care should be taken when training self-rationalization models with the goal of creating trustworthy models. Future work can investigate how to alleviate these harms while retaining the interpretability benefits of models that can rationalize their predictions.

6 Limitations

Conducting the analysis in this work required training over 700 models, particularly because the variability of model robustness requires training multiple models, governed by different random seeds, for every evaluation setting of interest. Thus, a main limitation of replicating this work is its computational demand.

Furthermore, even with the scale of our experiments, we do not exhaustively experiment with all possible evaluation settings of interest. Most notably, we focus our analysis on a standard way of training self-rationalization models—training generation models end-to-end to output rationales after their predictions; future work can investigate how our findings translate to other methods for training with rationales. In addition, while many evaluation sets targeting robustness exist for NLI, they do not for CQA; thus, our evaluation of robustness to spurious correlations for CQA were limited. Future work can develop more tests for evaluating robustness for tasks beyond NLI.

Acknowledgements

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and 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 3050–3065, Online. Association for Computational Linguistics.

David Alvarez-Melis and Tommi S. Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. *NeurIPS*.

Meghana Moorthy Bhat, Alessandro Sordoni, and Subhabrata Mukherjee. 2021. Self-training with few-shot rationalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10702–10712, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness?

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, Melbourne, Australia. Association for Computational Linguistics.

Peter Hase and Mohit Bansal. 2021. When can models learn from explanations? A formal framework for understanding the roles of explanation data.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021a. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 624–635, New York, NY, USA. Association for Computing Machinery.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021b. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L. McClelland, Jane X. Wang, and Felix Hill. 2022. Can language models learn from explanations in context?

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining 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.

Tania Lombrozo. 2016. Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10):748–759.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448,

Florence, Italy. Association for Computational Linguistics.

Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions.

Maxwell I. Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models.

Shruti Palaskar, Akshita Bhagia, Yonatan Bisk, Florian Metze, Alan W. Black, and Ana Marasović. 2022. On advances in text generation from images beyond captioning: A case study in self-rationalization. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, 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, New Orleans, Louisiana. Association for Computational Linguistics.

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2022. Evaluating Explanations: How Much Do Explanations from the Teacher Aid Students? *Transactions of the Association for Computational Linguistics*, 10:359–375.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself!

leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online. Association for Computational Linguistics.

Hendrik Schuff, Hsiu-Yu Yang, Heike Adel, and Ngoc Thang Vu. 2021. Does external knowledge help explainable natural language inference? automatic evaluation vs. human ratings. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 26–41, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. *AAAI*, abs/2104.08142.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Prasetya Utama, Nafise Sadat Moosavi, Victor Sanh, and Iryna Gurevych. 2021. Avoiding inference heuristics in few-shot prompt-based finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9063–9074, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022.

Chain of thought prompting elicits reasoning in large language models.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot in-context learning.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning.

Xinyan Zhao and V. G. Vinod Vydiswaran. 2021. Lirex: Augmenting language inference with relevant explanation. *AAAI*, abs/2012.09157.

	TEST ↑		CAD ↑		HANS ↑		TEST HYP ↑		TEST HARD ↑		Δ TEST SUBSETS ↓	
BART-BASE												
no rationales	75.37	±0.15	57.85	±0.61	51.30	±0.58	59.80	±0.67	56.86	±0.66	25.44	±2.75
original	+0.21	±0.18	+1.04	±0.63	+0.96	±0.62	-0.18	±0.38	+1.41	±0.46	-2.73	±1.65
shuffled	-4.71	±0.83	-5.60	±1.11	-0.21	±0.39	-5.60	±1.05	-4.40	±1.52	+2.10	±3.81
GPT2-MEDIUM												
no rationales	66.39	±0.94	48.84	±0.81	52.66	±0.38	48.70	±1.20	47.59	±2.22	33.39	±2.19
original	+1.12	±0.66	+0.80	±0.73	+1.42	±0.27	+0.59	±1.07	+3.61	±2.16	-3.94	±3.97
shuffled	-1.97	±0.75	-2.45	±0.95	+0.96	±0.55	-3.72	±1.46	0.00	±3.09	+0.66	±4.63
BART-LARGE												
no rationales	84.62	±0.31	71.20	±0.27	50.18	±0.08	73.68	±0.40	73.82	±0.85	14.70	±1.50
original	+1.05	±0.21	+1.75	±0.34	+0.44	±0.19	+1.56	±0.40	+2.04	±0.91	-2.04	±1.93
shuffled	-0.59	±0.37	-2.11	±0.45	+0.86	±0.52	-1.81	±0.50	-2.51	±1.60	+4.15	±2.86
T5-LARGE												
no rationales	84.03	±0.34	71.08	±0.31	51.54	±0.31	74.26	±0.54	75.45	±0.71	10.46	±1.03
original	+0.26	±0.21	+0.62	±0.48	-1.13	±0.17	-0.10	±0.40	-0.42	±1.11	+2.06	±2.33
shuffled	-0.67	±0.41	-0.23	±0.18	+3.02	±0.98	-0.75	±0.70	-2.62	±1.25	+3.27	±2.46

Table 5: Comparison between training with original rationales in ESNLI vs. shuffled rationales with 1K instances for NLI. Blue/red cells indicate improvement/worsening in performance compared to the baseline (no rationales). We report means across 5 random seeds, as well as standard errors of the means.

	TEST ↑		TEST HARD ↑		Δ TEST SUBSETS ↓	
BART-BASE						
no rationales	32.4	±0.3	33.6	±0.5	14.7	±1.4
positive	0.0	±0.4	+1.4	±0.8	-2.9	±2.3
freeflow	-1.4	±0.7	+0.3	±1.2	-0.0	±1.9
freeflow*	-1.7	±0.4	-0.6	±0.8	-0.0	±3.1
negative	-3.5	±1.0	-2.5	±1.2	-2.0	±1.6
negative*	-3.1	±0.8	-2.2	±0.8	+3.8	±1.7
GPT2-MEDIUM						
no rationales	30.6	±0.4	29.2	±1.5	13.3	±3.5
positive	+2.3	±0.3	+0.9	±1.5	+2.6	±2.5
freeflow	+2.6	±0.4	+2.1	±0.6	+2.7	±2.8
freeflow*	+3.4	±0.6	+2.9	±0.6	+1.8	±2.4
negative	+0.9	±0.2	+0.8	±1.1	+1.8	±3.7
negative*	+1.1	±0.4	+0.9	±0.6	+1.3	±1.9
BART-LARGE						
no rationales	48.3	±0.8	49.0	±1.0	19.9	±1.8
positive	+2.2	±1.3	+2.2	±1.0	-1.9	±3.0
freeflow	-2.3	±1.1	-1.7	±1.2	-2.4	±2.5
freeflow*	-3.3	±0.4	-1.1	±2.7	+2.2	±4.3
negative	-4.1	±1.2	-3.7	±1.8	-1.4	±2.1
negative*	-5.3	±0.3	-1.7	±0.5	-1.6	±2.0
T5-LARGE						
no rationales	60.4	±0.7	61.5	±1.5	10.5	±1.9
positive	+1.2	±0.7	+2.0	±1.7	+2.4	±1.3
freeflow	+1.0	±0.9	+1.6	±1.3	+0.3	±1.4
freeflow*	+1.0	±0.7	+1.7	±0.9	+2.1	±1.2
negative	-1.1	±0.6	-1.3	±0.8	+4.5	±3.1
negative*	-1.3	±0.5	-1.6	±0.7	+2.5	±1.4

Table 6: Effect of training with 1K instances using the different types of rationales in ECQA. * indicates length-controlled rationales. Blue/red cells indicate improvement/worsening in performance compared to the baseline (no rationales, row 1). We report means across 5 random seeds, as well as standard errors of the means.