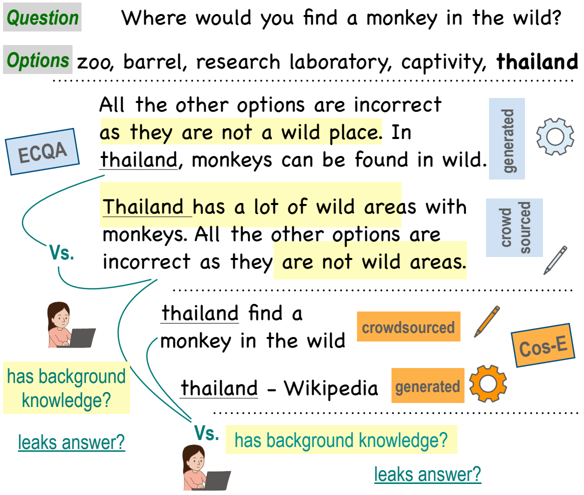
Investigating the Benefits of Free-Form Rationales

Jiao Sun[1] **Swabha Swayamdipta**[1] **Jonathan May**[1,2*] **Xuezhe Ma**[1,2]
[1]University of Southern California
[2]Information Sciences Institute
{jiaosun,swabhas}@usc.edu {jonmay,xuezhema}@isi.edu

Abstract

Free-form rationales aim to aid model interpretability by supplying background knowledge that can help understand model decisions. Popular commonsense QA datasets such as CoS-E and ECQA provide crowdsourced free-form rationales for instances, but their utility remains under-investigated. We present studies which show that 88% of ECQA rationales indeed provide humans additional background information to understand a decision, while 93% of CoS-E rationales do not. Inspired by this finding, we ask: can the additional context provided by free-form rationales benefit models, similar to their effect on human users? We investigate the usefulness of rationales as an additional training signal, by varying the quantity and quality of rationales during training. After controlling for instances where rationales leak the correct answer while not providing additional background knowledge, we find that incorporating only 5% of rationales during training can boost model performance by 47.22% for CoS-E and 57.14% for ECQA during inference. Moreover, we also show that rationale quality matters: compared to crowdsourced rationales, T5-generated rationales provide not only a weaker training signal, but are also not helpful for humans in aiding model interpretability.



Question
Where would you find a monkey in the wild?
Options
zoo, barrel, research laboratory, captivity, **thailand**
ECQA
generated
crowd sourced
Vs.
crowdsourced
Cos-E
generated
has background knowledge?
leaks answer?
Vs.
has background knowledge?
leaks answer?

All the other options are incorrect as they are not a wild place. In thailand, monkeys can be found in wild.

Thailand has a lot of wild areas with monkeys. All the other options are incorrect as they are not wild areas.

thailand find a monkey in the wild

thailand - Wikipedia

Figure 1: Illustration of our investigation into free-form rationales for commonsense QA from CoS-E (Trajanovski et al., 2021) and ECQA (Aggarwal et al., 2021). We conduct *human* studies to understand perceived usefulness of rationales, by asking if they contain background knowledge necessary to answer a question (yellow highlights). We also investigate if rationales leak the answer to *models* that use them as additional training signals. Our work compare rationales from different sources, and finds that ECQA rationales are preferable to CoS-E rationales on various axes. Finally, we find that crowdsourced rationales also offer greater benefits to both humans and models than generated rationales.

1 Introduction

Free-form rationales designed to explain decisions by providing additional world knowledge or commonsense reasoning, are key for interpretability (Kim, 2015; Lipton, 2018; Alvarez-Melis and Jaakkola, 2018) in natural language processing tasks.[1] Free-form rationales come with the promise of being easily interpretable by humans, in contrast to other kinds of explanations, such as extractive rationales in the form of textual highlights (Camburu et al., 2018; Lei et al., 2016), or low-level neuron activations in neural architectures (Hohman et al., 2020). Indeed, there have been increasing efforts to collect corpora containing free-form rationales for task instances, which provide a supervised setting for teaching models to produce rationales for test-time decisions. Such corpora include CoS-E (Rajani et al., 2019) and ECQA (Aggarwal et al., 2021) for commonsense question-answering, e-SNLI (Camburu et al., 2018) for natural language inference, SBIC (Sap et al., 2020) for social bias inference, among others.

However, the benefits of rationales remain un-

[*] Work done prior to JM joining Amazon.

[1] We use the terms "rationale" and "explanation" interchangeably. Please see Wiegreffe and Marasovic (2021) and Jacovi and Goldberg (2021) for more details on terminology.

clear. Do crowdsourced rationales really help human users interpret decisions better, or do they simply provide the right answer without the necessary background knowledge or reasoning? Our work explores this question through two carefully designed human studies. We find that rationales from different corpora have different capabilities: humans find 93% of ECQA rationales provide additional information that can help answer questions, while only 12% of CoS-E rationales do.

Inspired by this finding, we further ask: analogous to the benefit to human users, can crowdsourced rationales *also benefit models* by providing an additional training signal to boost performance? In contrast to prior work that uses rationales as supervision to generate model rationales, we focus on using crowdsourced rationales to simply aid a task model's classification capabilities. Our results indicate that while crowdsourced rationales do indeed boost model performance, they might be doing so trivially, i.e. by simply leaking the correct answer to the model. In response, we experiment with different strategies for altering ECQA and CoS-E rationales to prevent such leakage, and set up a fair test benchmark. We find that, even without leakage, rationales with background knowledge are helpful: including only 5% of high-quality rationales during training can improve model performance by 47.22% at inference time. This finding generalizes to QuaRTz (Tafjord et al., 2019), a dataset for textual relationship inference, in which rationales are designed explicitly to not leak the ground truth.

Finally, we investigate if *automatically generated* rationales provide similar benefits as crowdsourced rationales. Our human studies indicate that the perceived usefulness of generated rationales from T5 (Raffel et al., 2020) is much lower than that of human-written ones. Moreover, we find that these generated rationales are not particularly valuable as training signals. Overall, our results indicate that the quality of free-form rationales in existing popular datasets is paramount for both human interpretability as well as model supervision.[2]

2 Preliminaries

Tasks and Datasets. We explore three large datasets containing crowdsourced free-form natural language rationales. The first two, CoS-E (Rajani et al., 2019) and ECQA (Aggarwal et al., 2021), address commonsense-based question answering (ComQA). The ComQA task is based on answering questions about common situations, from a choice of 3 (CoS-E v1.0) or 5 (CoS-E v1.11) answers, along with providing a free-text explanation for the correct answer.[3] ECQA builds upon and improves the quality of CoS-E v1.11 explanations, in terms of comprehensiveness, refutation completeness and non-redundancy (Aggarwal et al., 2021). In addition, ECQA explanations are contrastive, i.e. they include rationales for choosing the correct option and rejecting other options (see Tables 1, 3, and Table 12 in Appendix D for examples).

Source	Rationale
CoS-E v1.11	People waiting alongside with when you're in a reception area
ECQA	People waits in a reception area. You cant wait along with a motel, hotel, chair or a hospital. These are the people where the reception area is found but people waits together at reception area of such places.
ECQA-shuffle	You cant wait along with a motel, hotel, chair or a hospital. These are the people where the reception area is found but people waits together at reception area of such places. People waits in a reception area.

Table 1: Example annotations from CoS-E v1.11 and ECQA for the question *"What are you waiting alongside with when you're in a reception area?"* with options *1: motel 2: chair 3: hospital 4: people 5: hotels* and the correct option *people*. CoS-E annotation directly combines the question and the correct answer, while ECQA annotation provides additional background knowledge.

We additionally consider an open-domain reasoning task about textual qualitative relationships, via the QuaRTz (Tafjord et al., 2019) dataset, for a subset of our experiments. In this task, each instance contains a triplet: a situated qualitative question, two answer options and a knowledge statement that can help answer the question. For example, for *"Compared to a box of bricks a box of feathers would be (A) lighter (B) heavier"*, the annotated knowledge in QuaRTz is *A given volume of a denser substance is heavier than the same volume of a less dense substance*. In contrast to CoS-E and ECQA, the two options for a question in QuaRTz are orthogonal, which means the knowledge provided to support one option will automatically reject the other option. Furthermore, this general qualitative knowledge statement in QuaRTz is

[2] https://github.com/sunjiao123sun/rationale-utility.

[3] CoS-E does not provide explanations for instances in the test set; we report our results on its validation set.

guaranteed to not leak the correct answer. While not explicitly designed for interpretability, we treat the annotated knowledge in QuaRTz as a rationale that can help understand or derive the correct answer. See dataset stats in Table 11 in App A.

3 Do crowdsourced rationales aid human interpretability?

Free-text rationales purportedly improve human user interpretability by explaining a model's decisions in natural language. We seek to discover which characteristics of the rationales aid users:

Q1 Do rationales provide **additional background knowledge** for understanding decisions? E.g., the rationale: '*Air cannot stay in any object that has a hole in it*' provides additional knowledge for understanding why the answer to '*What would not be true about a basketball if it had a hole in it but it did not lose its general shape?*' should be '*full of air*'.

Q2 Do rationales provide explicit clues to **leak** the correct answer? For ComQA, this might initially seem like a helpful rationale, without really being so.[4] E.g., given a rationale: '*Mexico is one of the largest coffee production country.*', one can guess the correct answer should be '*mexico*', when given the options '*mildred's coffee shop*', '*mexico*', '*diner*', *kitchen*' or '*canteen*', without looking at the question '*In what Spanish speaking North American country can you get a great cup of coffee?*'.

3.1 Preliminary Studies

We investigate Q1 and Q2 via a direct assessment (§3.1.1) by human raters, as well as via proxy questions offering an indirect assessment (§3.1.2) by the raters.

3.1.1 Direct Assessment

We conduct a pilot study where given the question, options, correct answer and rationales from CoS-E and ECQA for a ComQA instance, annotators are tasked to *directly* answer which rationale provides *additional background knowledge* that can help them answer the question. Four options are possible: CoS-E, ECQA, *neither*, or *both*.[5] Simultaneously, we ask annotators if any of the two rationales leaks the correct answer. Concretely, the annotators are required to provide three annotations for each instance:

	$\mathrm{R}_{\mathrm{ECQA}}$	$\mathrm{R}_{\mathrm{CoS\text{-}E}}$	**both**	**neither**
Q1: has bg. knowl.?	65.0%	9.2%	20.8%	5.0%
Q2: leaks answer?	83.3%	43.3%	n/a	n/a

Table 2: Human study directly comparing ECQA and CoS-E rationales on 120 ComQA instances, for the presence of background knowledge, and answer leakage.

- *choose one option for the additional background information* (T1);
- *judge if* ECQA *rationale leaks the correct answer* (T2);
- *judge if* CoS-E *leaks the correct answer* (T3).

We conduct our study on the first 120 rationales in ECQA and CoS-E v1.11 test set via the Amazon Mechanical Turk platform. For each instance, we collect annotations from three independent annotators. Using Fleiss's Kappa (Fleiss and Cohen, 1973), the inter annotator agreement (IAA) for T1, T2 and T3 are 0.43, 0.26, and 0.30, respectively, indicating moderate agreement. We take the majority vote as the final label.[6]

Table 2 shows the results of our human evaluation. We see 85.8% of ECQA rationales provide additional background knowledge to help answer the question, while only 30.0% of CoS-E rationales do the same, indicating greater usefulness of ECQA rationales for human interpretability. Both ECQA and CoS-E rationales leak the correct answers. Indeed, most ECQA rationales provide some background knowledge necessary for humans to understand the decision, while also revealing the correct answer; the same does not hold for CoS-E.

3.1.2 Indirect Assessment

While the previous study asked participants to directly assess the background knowledge of individual rationales, we design two other studies below that use a proxy to extract a human assessment of

[4] While leakage does not reduce the utility of a rationale for human interpretability, it does have implications for utility as model supervision, as we will see in subsequent sections, §4 and §5.

[5] While Aggarwal et al. (2021) provide similar human studies comparing ECQA and CoS-E rationales, they do not specifically ask for additional background knowledge.

[6] Further details on this study are in the Appendix B, including Fig. 2 showing our annotation interface.

	question	options	R^{crowd}	$R^{\text{constructed}}$
CoS-E	Where can a human find clothes that aren't pants?	pants shop, on planet earth, **dress shop**, school, train wreck	dress shop can a human find clothes that aren't pants.	A human can find clothes at dress shop that aren't pants.
ECQA	Where do adults use glue sticks?	classroom, desk drawer, at school, **office**, kitchen-drawer	Glue stick is a solid glue used to stick thin paper materials by adults in offices. Adults don't go to classroom and school, and other options don't have adults.	Adults use glue sticks in their offices. They do not use them at classroom, desk drawer, at school or kitchen drawer.

Table 3: Examples of crowdsourced rationales for CoS-E and ECQA, vs. our manually constructed rationales that declaratively combine the question and the answer without providing any background knowledge or commonsense reasoning.

	R^{crowd}	$R^{\text{constructed}}$	neither	either
CoS-E	3.0%	5.0%	92.0%	0.0%
ECQA	73.0%	9.0%	14.0%	4.0%

Table 4: Results from our human study via indirect assessment to compare 100 pairs of crowdsourced and constructed rationales. The IAA is 0.61.

rationale utility (Tan, 2022), for Q2 and Q1, respectively. Here, we randomly sample 100 ComQA instances from the test set.

For Q2, we ask annotators to guess the correct answer from all options, given only the crowdsourced rationales from CoS-E and ECQA; annotators can also opt for "cannot tell" based on the evidence (see our interface in Appendix B, Fig. 4). We hypothesise that this study will indirectly answer whether the rationale leaks the correct option, if the worker is able to guess correctly. Each instance is provided to three annotators, and we take a majority vote for their ratings. We find that annotators are able to pick the correct answer, given only the rationales (and not questions) in 43.0% of cases for CoS-E and 78.0% of cases for ECQA, with high agreement (IAA 0.73). This confirms our findings from the direct assessment in Table 2.

For Q1, we manually construct rationales to contrast with crowdsourced rationales. Our constructed rationales are designed to simply combine the question and the correct answer, but not provide any additional background knowledge. If a human prefers the crowdsourced rationale, we can indirectly ascertain that it provides some background knowledge to help with human interpretability. For CoS-E, we form a constructed rationale for a question by rephrasing the question as a statement and inserting the correct option in place of the question word. For ECQA, in addition to the CoS-E-style constructed sentence, we add an additional sentence that rephrases the question as a negative statement, replaces some referents with pronoun anaphora, and inserts the incorrect options in place of the question word. We also try to ensure fluency and stylistic consistency with the crowdsourced explanations.

We show two examples of our constructed rationales in Table 3. We provide human subjects with the question, the correct answer, the crowdsourced rationale (from CoS-E or ECQA) and our constructed rationale. We instruct workers to choose the explanation that they would prefer if they need to explain the correct answer to someone who might not have the necessary background knowledge to understand given only the question and set of answer choices (see our interface in Appendix B, Fig. 3). Each instance is provided to three annotators, and we take a majority vote for their ratings.

Results in Table 4 show that human raters overwhelmingly preferred neither our constructed rationales or CoS-E rationales, indicating that neither provides background knowledge necessary for answering the question.[7] On the other hand, raters seem to prefer ECQA rationales over our constructions, indicating that the former might contain background knowledge owing to their rigorous annotation procedure (Aggarwal et al., 2021). Yet, surprisingly, raters picked our constructed rationales 9% of the time over ECQA, while being ambivalent about either rationale for 4% of the cases; moreover, they liked neither for 14% of the cases! This could indicate that some ECQA instances might not provide adequate background knowledge, and / or raters might at times choose simpler (though vacuous) rationales; future work might pursue studying such cases.

3.2 Categorizing Crowdsourced Rationales

CoS-E**.** Although Narang et al. (2020) criticize the quality of CoS-E rationales, CoS-E v1.11 is still widely used for commonsense reasoning (Paranjape et al., 2021), analysis (Majumder et al., 2021; Wiegreffe et al., 2021), and as an additional source of commonsense knowledge (Ye et al., 2019). In order for the community to understand the deficiencies of the crowdsourced CoS-E rationales, we

[7] Surprisingly, raters preferred the constructed rationales slightly over the crowdsourced rationales, which might be because some CoS-E rationales are off-topic; see Appendix B.

Category	**Description**	**Example**	**Distribution**
$R_{\textbf{no-leak-bg}}$	provides additional background knowledge without leaking correct answers.	*Question:* What would not be true about a basketball if it had a hole in it but it did not lose its general shape? *Options:* 1: punctured 2: popular in america 3: **full of air** 4: gone 5: round *Rationale*: Air cannot stay in any object that has a hole in it.	4.83% (59/1221)
$R_{\textbf{leak-bg}}$	leaks the correct answer but contains additional background knowledge that can help answer questions.	*Question:* In what Spanish speaking North American country can you get a great cup of coffee? *Options:* 1: mildred's coffee shop 2: **mexico** 3: diner 4: kitchen 5: canteen *Rationale*: Mexico is one of the largest coffee production country.	6.72% (82/1221)
$R_{\textbf{no-leak-no-bg}}$	neither provides any additional background information, nor leaks the correct answer.	*Question:* why would a person like to have a large house? *Options:* 1: have choice 2: mentally challenged 3: own house 4: obesity **5: lots of space** *Rationale*: This word is most relevant	43.65% (533/1221)
$R_{\textbf{leak-no-bg}}$	leaks the correct answer and does not provide additional background knowledge.	*Question:* where will a cheap book be found? *Options:* 1: bookstore 2: classroom 3: **discount store** 4: school room 5: bedside table *Rationale*: discount shop retail shop	44.80% (547/1221)

Table 5: Our manual four-way categorization of CoS-E v1.11 (dev.) rationales, with examples. Bolded options indicate ground truth. We find that 88.45% of rationales do not provide additional background knowledge.

provide a detailed study of the same, which was missing in Narang et al. (2020).

Building on Q1 and Q2, we aim to categorize CoS-E rationales into 4 categories, to determine if these provided background knowledge and/or leaked the answer. One of the authors manually categorized the rationales in the development set of CoS-E v1.11 into four categories. To validate this categorization, three co-authors annotated a subset of 100 instances independently for the same categorization. We obtained an IAA Fleiss Kappa of 0.65 for *background knowledge* and 0.84 for *leakage*, indicating moderate / high agreement. For these 100 instances, we use the majority vote among the three annotators as the final label. Appendix C provides further details.

Table 5 describes and shows the distribution of the categories, with examples from each picked at random. Rationales that do not provide additional background knowledge make up 88.45% of the entire development set of CoS-E v1.11. Using the development set as a lens, our annotation provides a qualitative and quantitative understanding of the crowdsourced CoS-E rationales. Future research should take into consideration these findings before using CoS-E rationales.

ECQA. Aggarwal et al. (2021) build on CoS-E question-answer pairs and carefully collect detailed rationales. Table 1 compares CoS-E and ECQA rationales, where the former directly combines the correct answer and the question, but the latter contains additional commonsense knowledge that can help answer the question, suggesting a higher quality. Moreover, ECQA rationales are contrastive as they explain, for each option, why it is correct or incorrect. Regardless, we find that all ECQA rationales *start* by explaining the correct option, followed by other options. This ordering introduces a spurious correlation which likely provides a shortcut to a model for predicting the correct answer from the rationale, but for wrong reasons (Geirhos et al., 2020). A random shuffle of the sentences within each ECQA rationale (last row; Table 1) can address this issue.[8]

4 Can Models Benefit from Crowdsourced Rationales?

In §3, we found that crowdsourced rationales from carefully constructed corpora provide additional information to help humans better answer commonsense questions. Now, we seek to answer if these rationales could also help in model learning, by providing an additional training signal to make better decisions, taking into account our findings from the detailed analysis in §3.

Experimental Setup. We use finetuned T5 (Raffel et al., 2020) models throughout our work following prior efforts for analyzing (Wiegreffe et al., 2021) and generating (Narang et al., 2020; Lakhotia et al., 2021) free-text explanations. More specifically, we finetune three model classes based on the T5-base architecture:

- I→O. Predict the label directly from the question and answer options.
- IR→O. Predict the label from the question, answer options *and* the rationale.
- I→R. Predict the rationale from the question and answer options.

For the IR→O model, we experiment with different variations based on the source, and the quantity of the rationales R, provided during training. Since

[8] We use the Spacy sentencizer to split the rationale, and randomly permute sentence ordering, with seed 0.

			$c1$	$c2$	$c3$	$c4$	$c5$	$c6$	$c7$	$c8$
		Test →	**I**	**I+**$R_{\text{CoS-E}}$	**I+**R_{ECQA}		**I+**$R_{\text{CoS-E}}$ (test subsets)			
		$\%R^{\textbf{Tr}}$			**w/o shuffle**	**shuffled**	$R_{\textbf{no-leak-bg}}$	$R_{\textbf{leak-bg}}$	$R_{\textbf{no-leak-no-bg}}$	$R_{\textbf{leak-no-bg}}$
$r1$	I→O	0%	57.00	46.11	53.32	54.95	40.68	46.34	45.97	46.80
$r2$	$\text{IR}^{\text{Tr}}_{\text{CoS-E}}$ →O	5%	$53.78_{1.10}$	$72.53_{2.19}$	$76.50_{2.30}$	$65.57_{2.86}$	$59.89_{5.24}$	$87.40_{4.14}$	$54.72_{1.17}$	$89.03_{2.72}$
$r3$		10%	$54.44_{0.72}$	$76.03_{1.00}$	$80.78_{1.53}$	$63.74_{0.78}$	$70.06_{2.88}$	$89.02_{3.45}$	$56.85_{1.48}$	$93.42_{0.65}$
$r4$		20%	$53.62_{0.23}$	$77.23_{0.47}$	$83.40_{1.41}$	$62.71_{1.80}$	$68.93_{5.59}$	$95.53_{1.15}$	$56.97_{1.08}$	$95.12_{0.09}$
$r5$		30%	$53.12_{0.60}$	$77.43_{0.30}$	$79.17_{3.23}$	$63.56_{1.28}$	$73.55_{5.46}$	$94.71_{0.58}$	$56.72_{1.11}$	$96.49_{0.67}$
$r6$		100%	48.24	78.46	66.01	64.46	71.19	97.56	57.97	96.34
$r7$	$\text{IR}^{\text{Tr}}_{\text{ECQA-shf.}}$ →O	5%	$54.05_{0.95}$	$59.27_{0.91}$	$86.65_{1.10}$	$86.35_{1.54}$	$51.41_{7.10}$	$69.10_{1.52}$	$53.22_{0.49}$	$64.53_{2.35}$
$r8$		10%	$54.05_{1.08}$	$61.72_{2.11}$	$92.55_{0.52}$	$93.01_{0.37}$	$54.80_{5.24}$	$72.76_{4.03}$	$52.53_{1.46}$	$69.77_{3.16}$
$r9$		20%	$53.29_{0.32}$	$66.50_{0.66}$	$95.41_{0.48}$	$94.70_{1.17}$	$64.41_{4.99}$	$83.74_{1.15}$	$55.85_{0.69}$	$74.53_{1.36}$
$r10$		30%	$52.85_{0.67}$	$65.05_{0.78}$	$95.85_{0.34}$	$95.52_{0.51}$	$56.50_{5.24}$	$81.30_{2.30}$	$52.91_{0.41}$	$75.38_{2.15}$
$r11$		100%	38.08	67.32	97.3	96.56	55.93	93.90	39.40	91.77

Table 6: ComQA accuracies under various train (***r***ow) and test (***c***olumn) settings. $r1$ is an I→O T5 baseline without access to rationales during training; the following rows use different amounts (%R$^{\text{Tr}}$) of CoS-E rationales ($r2 - r6$) and shuffled ECQA rationales ($r7 - r11$) for training IR→O T5 models. At inference time, each model predicts the label given no rationale ($c1$), or given the crowdsourced rationales for the entire test set ($c2$-$c4$), or a subset of the CoS-E test set ($c5$-$c8$), selected based on the rationale categories in Table 5. $c4$ and $c3$ report ECQA test set performance, when the test rationales are shuffled or not, respectively. We report accuracies averaged across 3 random seeds (stdev as subscript) for %R selection during training.

			$c1$	$c2$	$c3$
			I	**I+**R_{ECQA}	
		$\%R^{\textbf{Tr}}$		**w/o shuffle**	**shuffled**
$r1$	I→O	-	57.00	53.32	54.95
$r2$	$\text{IR}^{\text{Tr}}_{\text{ECQA}}$ →O	5%	55.45	**93.94**	**76.66**
$r3$		10%	55.36	96.56	73.46
$r4$		20%	54.55	97.21	70.02
$r5$		30%	53.64	97.46	66.91
$r6$		100%	31.44	97.79	76.33

Table 7: The importance of shuffling the order of sentences in ECQA rationales in training. Without shuffling, the model relies on the spurious correlation due to sentence order, as compared to r7-11/c4 in Tab. 6. Accuracies are averaged across 3 random seeds (s.d. as subscript) for %R selection during training, as in Tab. 6.

most of our experiments deal with the first two model classes, we report accuracy of output label prediction. See App. A for details on our T5 model training and I/O formats.

We use rationales for the ComQA training instances to train two different sets of IR→O models, for CoS-E and ECQA respectively. Under each set, we train five different models, randomly selecting different amounts (5%, 10%, 20%, 30% and the full 100%) of CoS-E and shuffled-ECQA rationales for training.[9] During training, we use varying amounts (5%, 10%, 20%, 30% and the full 100%) of CoS-E and shuffled-ECQA rationales, to study how the quantity of rationales affects the model performance. During inference, we provide the IR→O T5 models with rationales under each of the four categories of CoS-E, as in Table 5, as well as all combined together. For ECQA, we report performance for inference with and without shuffled rationales. Finally, we study how rationales from one dataset transfer to the other.

Crowdsourced rationales boost model performance, ruling out leakage. Comparing $c1$ in Table 6 with the columns $c2$-$c8$, we see that rationales help improve the model's ability to make the correct prediction, even when including only 5% of the rationales during training. However, instances that leak the answer make up a large portion of CoS-E. Indeed, when provided at test time, *rationales which neither leak the correct answer nor provide additional background knowledge, cause the least improvement in model performance ($c7$).* Further, with background knowledge, but no leakage, model performance can still be improved ($c5$); after adding 5% of the training data, the model reaches 59.89% accuracy with $C_{\text{no-leak-bg}}$ rationales, which yields 47.2% improvement, compared to 40.68% without rationales.[10] Overall, a close inspection of the rationales is necessary to understand when they can help the model decision for the right reasons (i.e. providing background information, not simply by leaking the answer). In other words, models can benefit from those crowdsourced rationales which

[9] Some training instances receive both the I and R and others receive just I, see Appendix A.1.

[10] Unlike test rationales from other categories, the trends are not monotonic for $R_{\text{no-leak-bg}}$, most likely because this is the smallest (only 4%) subset of the test set (Table 5).

		$\%\mathbf{R}^{\mathbf{Tr}}$	**I**	$\mathbf{I+R}_{\textsf{QuaRTz}}$
r1	I→O	-	70.88	38.27
r2	$\text{IR}^{\text{Tr}}_{\textsf{QuaRTz}}\rightarrow\text{O}$	5%	$66.20_{1.33}$	$67.86_{1.18}$
r3		10%	$67.81_{1.15}$	$70.58_{1.25}$
r4		20%	$67.99_{0.54}$	$69.73_{0.97}$
r5		30%	$67.13_{0.69}$	$71.51_{0.16}$
r6		100%	64.67	81.51

Table 8: QuaRTz model accuracy with and without training with knowledge statements as rationales. We report accuracies averaged across 3 random seeds (s.d. as subscript) for %R selection during training, as in Table 6.

provide utility for human interpretability as well!

Not all rationales are the same. We see benefits from increasing the amount of ECQA rationales in the training data ($r7$-$r11$/ $c4$), even in a transfer setting ($r7$-$r11$/ $c2$). However, this trend is weaker when training with CoS-E ($r2-r6$). This highlights the importance of a rigorous procedure for crowdsourcing rationales (Aggarwal et al., 2021).

Spurious correlations in rationales must be minimized. Recall from §3.2 that ECQA rationales tend to follow an ordering: sentences rationalizing the correct option precede those refuting the incorrect ones. To validate the importance of shuffling sentences in ECQA rationales, we present a baseline in Table 7 which considers unshuffled rationales during training, to be compared to training with shuffled rationales in Table 6. In the unshuffled case, training with only 5% rationales improves the accuracy on unshuffled test rationales from 53.32% to 93.94% ($c2$, Tab. 7). However, when we test the same model using shuffled rationales, the accuracy improves from 54.95% to 76.66% ($c3$). This shows that the model might learn a spurious correlation between the rationale and correct answer, due to ordering. We recommend shuffling ECQA rationales before using them for model training.

		$R^{\textbf{crowd}}$	$R^{\textbf{gen.}}$	**neither**	**both**
CoS-E	Q1: has bg. knowl.?	28.33%	20.00%	34.17%	17.50%
	Q2: leaks answer?	51.67%	40.83%	-	-
ECQA	Q1: has bg. knowl.?	43.44%	22.50%	15.00%	19.17%
	Q2: leaks answer?	89.17%	64.17%	-	-

Table 9: Comparative human studies for direct assessment of annotated vs. generated rationales in CoS-E and ECQA, similar to Table 2. Humans believe that generated rationales less frequently provide additional background knowledge than crowdsourced rationales.

		testing with generated R		testing with crowd R	
	$\%R^{\textbf{gen.}}_{\textbf{Tr.}}$	$R^{\textbf{gen.}}_{\textsf{CoS-E}}$	$R^{\textbf{gen.}}_{\textsf{ECQA}}$	$R^{\textbf{crowd}}_{\textsf{CoS-E}}$	$R^{\textbf{crowd}}_{\textsf{ECQA}}$
$\text{IR}^{\text{gen.}}_{\textsf{CoS-E}}\rightarrow\text{O}$	5%	$44.34_{1.59}$	$45.1_{0.86}$	$68.96_{2.11}$	$58.83_{1.27}$
	10%	$44.94_{0.59}$	$42.89_{0.46}$	$75.98_{0.72}$	$60.53_{0.46}$
	20%	$44.34_{0.71}$	$41.17_{0.74}$	$77.45_{0.54}$	$62.38_{0.27}$
	30%	$44.91_{0.43}$	$39.83_{0.56}$	$77.07_{0.70}$	$64.65_{0.95}$
	100%	**43.90**	35.71	**76.74**	60.11
$\text{IR}^{\text{gen.}}_{\textsf{ECQA}}\rightarrow\text{O}$	5%	$46.33_{0.54}$	$44.64_{1.03}$	$58.97_{1.07}$	$79.31_{1.17}$
	10%	$45.10_{0.34}$	$44.96_{0.30}$	$60.41_{0.47}$	$87.22_{0.40}$
	20%	$46.98_{0.83}$	$45.67_{0.37}$	$62.00_{0.35}$	$89.68_{0.47}$
	30%	$45.81_{0.60}$	$45.51_{0.40}$	$64.51_{1.03}$	$91.51_{0.80}$
	100%	43.16	**44.64**	64.86	**93.37**

Table 10: Performance of IR→O models, trained with different amounts ($\%R^{\text{gen.}}_{\text{Tr.}}$) of generated rationales. The top block indicates training with rationales generated from a CoS-E-trained I→R model, and the bottom block, an ECQA-trained model. The columns indicate rationales provided at test time to the IR→O models. We report accuracies averaged across 3 random seeds (s.d. as subscript) for $\%R^{\text{gen.}}_{\text{Tr.}}$ selection, as in Table 6.

Training with non-leaky rationales is beneficial. Despite taking care to prevent spurious correlations in ECQA, there is still a chance that the models benefit from some amount of leakage of the correct answer, an uninteresting use of rationales to improve model performance. To control for this, we consider the QuaRTz dataset, introduced in §2, using knowledge statements as rationales, which are designed to contain no leakage, but provide the background information. Using a similar setup to our ComQA experiments above, we finetune T5 models for both I→O and IR→O models on QuaRTz. Results in Table 8 show that the non-leaky QuaRTz rationales improve a model's ability to predict the correct answer, consistent with our findings in Table 6. These highlight the generalizability of our conclusions.

5 Can Models Benefit from Generated Rationales?

So far, we have focused on crowdsourced rationales, the primary reason behind collecting which is to train models that generate them automatically, as seen in recent work (Narang et al., 2020; Paranjape et al., 2021). Hence, we now ask: (1) analogous to crowdsourced rationales (§3), can *generated* rationales provide the additional background information necessary for human interpretability, and (2) can *generated* rationales provide additional training signals to improve model performance, similar to §4?

Using crowdsourced CoS-E and ECQA rationales as supervision, we train two different T5-base I→R models, following §2. We generate rationales on all ComQA instances (train as well as test) using these two models.[11]

Human Perception of Generated Rationales. We repeat our comparative human study via direct assessment (on the same 120 test ComQA instances from §3.1.1) on generated vs. annotated (1) ECQA and (2) CoS-E rationales (see Fig. 2 in Appendix B). Table 9 shows that humans believe generated rationales provide background knowledge less often than human-annotated rationales.[12] Generated ECQA rationales have more background knowledge than generated CoS-E rationales, but do leak the correct answer more often as well, reflecting the ECQA training data. These results are consistent with our findings in §3.

Training with Generated Rationales. Next, we use generated rationales for the ComQA training instances to train two different sets of IR→O models, for CoS-E and ECQA respectively. Under each set, we train five different models, randomly selecting different amounts of generated rationales for training. At inference time, we provide the IR→O models, either generated rationales for each ComQA test instance, following the setting from Wiegreffe et al. (2021), or crowdsourced rationales for the same set.

Table 10 shows the results. When testing with generated rationales, we see a reduction in the model's predictive capability compared to testing with crowdsourced rationales.[13] Moreover, training with ECQA-generated rationales seems more beneficial than training with CoS-E-generated rationales. Training with larger quantities of generated rationales keeps boosting the model performance to an extent (from 5% to 30%), consistent to our findings with crowdsourced rationales. However, when each and every instance is paired with its generated rationale during training (rows with 100%), the performance drops when testing with generated rationales as well. We suspect this might be attributed to the generated rationales introducing too much noise, and not providing the model a clear signal. Regardless, under neither setting, do these models do as well as the models trained with crowdsourced rationales, as shown in Table 6.

6 Related Work

Rationales serve interpretability in that they can reveal the "reasoning" behind model decisions, and can be roughly categorized into two broad categories: extractive and free-form rationales. Extractive rationales provide supportive evidence in a grounded context, such as textual highlights within an input document (Lei et al., 2016; DeYoung et al., 2020), sufficient to make a prediction on its own without relying on the rest of the input. Analogous to our work, there has been a line of work that studies the value of extractive rationales as additional training signals to improve model performance (Huang et al., 2021; Carton et al., 2021) or human interpretability (Strout et al., 2019). We use free-text rationales, on the other hand, which employ free-form natural language to fill in the commonsense reasoning or knowledge gaps. Such rationales have been used for language and vision tasks (Hendricks et al., 2016; Kim et al., 2018) but have far less adoption in NLP (Wiegreffe and Marasovic, 2021). Concurrent to our work, Hase and Bansal (2022) provide a formal framework for free-text explanation utility for models, and a synthetic dataset for the same. In addition to the datasets used in this work, e-SNLI (Camburu et al., 2018) provides free-text rationales for the natural language inference task.

Rationale generation models can be roughly categorized into supervised (Lakhotia et al., 2021; Narang et al., 2020; Kumar and Talukdar, 2020; Rajani et al., 2019; Zhao and Vydiswaran, 2021) and unsupervised (Glockner et al., 2020; Brahman et al., 2021). For supervised models, Lakhotia et al. (2021) and Narang et al. (2020) finetune T5 to generate extractive and free-form rationales separately. For unsupervised models, Glockner et al. (2020) propose a differential training framework to create models that output faithful rationales without supervision. Instead of directly generating rationales, Paranjape et al. (2021) use T5 to complete contrastive explanation prompts which explicitly contrast different possible answers (Jacovi et al., 2021). Recent work has relied on few-shot prompting to generate explanations (Wiegreffe et al., 2022;

[11] A qualitative analysis of annotated and generated rationales for both CoS-E and ECQA can be found in Appendix D.

[12] For generated rationales, attributes such as plausibility and faithfulness (Jacovi and Goldberg, 2021) also matter. However, we focus on the same attributes as we use to evaluate crowdsourced rationales (§3), for a contrastive comparison.

[13] See App. §A.2 for an evaluation of these rationales.

Marasović et al., 2022). Our work follows a supervised approach to generate rationales, and further uses both generated and crowdsourced rationales as training signals to produce output labels.

7 Conclusion

We investigated the utility of free-form rationales from both a human and a modeling perspective. Centering our analysis on commonsense QA datasets, we find that humans perceive rationales with more background knowledge as more useful than those which simply combine the question and the answer. We provided a detailed qualitative analysis of CoS-E and ECQA rationales, and found that even small amounts of carefully written rationales are helpful as additional training signals for task models. Our work highlights the importance of inspecting the quality of crowdsourced rationales before using them for additional supervision. We also found that generated rationales are not as useful as crowdsourced rationales for human interpretability or for model supervision. Our investigations shed light on fundamental assumptions about human interpretability in collecting and generating rationales, and call for further deeper investigation into the utility of free-form rationales.

Ethical Consideration

During our manual annotation process, we provide timely warning of potential adult topics and ask workers to return the job if they are under age. Our human studies are labeled as exempt from review by the IRB at the authors' institute.

For modeling, we use T5 throughout our work, which also involves generating rationales. It is well-known that such pretrained language models—trained on massive online texts—capture the biases reflected in the training data (Bender et al., 2021). Therefore, the generation models can be used for malicious purposes and generate rationales that contain toxic content that target at certain groups. We do not have a filtering mechanism that checks the toxicity, bias, or offensiveness of our training data, or that of our generated explanations. Hence, we recommend practitioners interested in using and replicating this work to carefully check the generated content before deployment in any real world application.

The datasets used in our work are all public. These do not contain any explicit detail that leaks information about a user's name, health, negative financial status, racial or ethnic origin, religious or philosophical affiliation or beliefs.

Limitations

This work is subject to several limitations. First, our human studies are pilot studies, where we annotated only 100 instances to understand how much rationales can aid the human interpretability. Although the trend we observed is intuitive and consistent, more data for the human study might improve the quality of the findings. Secondly, we use vanilla T5 models for rationale generation for a fair comparison with previous work. However, there could potentially be rationales of higher quality via more sophisticated and powerful language models, which are beyond the scope of our exploration. Last, this work focuses on question-answering for commonsense knowledge. It still remains unexplored if our conclusions transfer beyond this task.

There are many other ways to evaluate the utility of free-form rationales. Our work focuses on the specific aspect of if a rationale provides additional background information that can help answer the question. Please note that leakage, which is also studied in our work, does not reduce the utility of a rationale for human interpretability. It is natural for an explanation to explicitly lead to the correct answer. However, the specific reason why we include the study of leakage in §3 is that we wanted to bring to the annotator's attention (implicitly) that a rationale might look good simply because it mentions the correct answer even though it might not contain the reasoning for it. We refer interested workers to Jacovi and Goldberg (2020) for discussions about the criteria that constitutes a high-quality interpretation. Future works can explore other aspects that can contribute to or jeopardize the utility of rationales for both humans and models, including but not limited to the factuality, completeness and presence of unnecessary information in rationales.

Acknowledgments

The authors thank anonymous reviewers and the meta reviewer for their constructive feedback and suggestions that helped improve the draft. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement Nos. HR00112290025 and HR00112290056.

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 T. Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *NeurIPS*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2021. Learning to rationalize for nonmonotonic reasoning with distant supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12592–12601.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *NeurIPS*.

Samuel Carton, Surya Kanoria, and Chenhao Tan. 2021. What to learn, and how: Toward effective learning from rationales. *ArXiv*, abs/2112.00071.

Miruna-Adriana Clinciu, Arash Eshghi, and Helen Hastie. 2021. A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2376–2387, Online. Association for Computational Linguistics.

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.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613 – 619.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Max Glockner, Ivan Habernal, and Iryna Gurevych. 2020. Why do you think that? exploring faithful sentence-level rationales without supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1080–1095, Online. Association for Computational Linguistics.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Peter Hase and Mohit Bansal. 2022. When can models learn from explanations? a formal framework for understanding the roles of explanation data. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 29–39, Dublin, Ireland. Association for Computational Linguistics.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367, Online. Association for Computational Linguistics.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. *Lecture Notes in Computer Science*, page 3–19.

Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Chau. 2020. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE Transactions on Visualization and Computer Graphics*, 26:1096–1106.

Quzhe Huang, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2021. Exploring distantly-labeled rationales in neural network 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 5571–5582, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2021. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. 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.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1224–1234. IEEE.

Been Kim. 2015. *Interactive and interpretable machine learning models for human machine collaboration*. Ph.D. thesis, Massachusetts Institute of Technology.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. *Lecture Notes in Computer Science*, page 577–593.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Kushal Lakhotia, Bhargavi Paranjape, Asish Ghoshal, Scott Yih, Yashar Mehdad, and Srini Iyer. 2021. FiD-ex: Improving sequence-to-sequence models for extractive rationale generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3712–3727, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. 2021. Rationale-inspired natural language explanations with commonsense. *ArXiv*, abs/2106.13876.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E. Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of NAACL*.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *ArXiv*, abs/2004.14546.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2021. Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

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.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. 2021. SELFEXPLAIN: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 836–850, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62, Florence, Italy. Association for Computational Linguistics.

Jiao Sun, Xuezhe Ma, and Nanyun Peng. 2021. AESOP: Paraphrase generation with adaptive syntactic control. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5176–5189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. QuaRTz: An open-domain dataset of qualitative relationship questions. 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 5941–5946, Hong Kong, China. Association for Computational Linguistics.

Chenhao Tan. 2022. On the diversity and limits of human explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2173–2188, Seattle, United States. Association for Computational Linguistics.

Stojan Trajanovski, Chad Atalla, Kunho Kim, Vipul Agarwal, Milad Shokouhi, and Chris Quirk. 2021. When does text prediction benefit from additional context? an exploration of contextual signals for chat and email messages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 1–9, Online. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

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.

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

Zhiquan Ye, Qian Chen, Wen Wang, and Zhenhua Ling. 2019. Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. *ArXiv*, abs/1908.06725.

Xinyan Zhao and V.G.Vinod Vydiswaran. 2021. Lirex: Augmenting language inference with relevant explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14532–14539.

Dataset	Train	Test
CoS-E v1.11/ECQA	9,741	1,221
QuaRTz	2,695	783

Table 11: The statistics of 3 datasets in our work.

A Dataset and Implementation Details for Finetuning T5

Table 11 shows the statistics for the datasets used in our work.

We finetune multiple T5 models (Raffel et al., 2020) in our work, and we use HuggingFace (Wolf et al., 2020) throughout our implementation. We use 512 and 256 for the maximum source length and the maximum target length separately. To optimize, we use AdamW (Loshchilov and Hutter, 2019) with a learning rate of 0.0001. We train each model on a NVIDIA RTX 2080 with a batch size of 8 for 30 epochs. During inference, we use beam search as the decoding method with a beam size of 2. The generation of the EOS token or reaching the maximum target length terminates decoding.

A.1 Formatting the Seq2Seq Models.

The formatting of the different models is:

- I→O. Predict the label directly from the question, formatted as: `context:` *{question}* `options:` *{concatenated options}* → *{answer}*.
- IR→O. Predict the label from the question and the rationale, formatted as: `context:` *{question}* `options:` *{concatenated options}* `explanation:` *{rationale}* → *{answer}*.
- I→R. Predict the rationale from the question, formatted as: `explain question:` {question} `answer:` *{concatenated options}* → `explanation:` *{rationale}*.

A.2 Evaluation of I→ R models

Following Wiegreffe et al. (2021), we use *simulatability score* to measure the quality of generated rationales. Simulatability captures the predictive ability a rationale provides over the input:

$$\text{acc}(\text{IR} \rightarrow \text{O}) - \text{acc}(\text{I} \rightarrow \text{O}). \qquad (1)$$

Prior work has shown that simulatability score serves as a reliable measure of rationale quality from the lens of utility to an end user (Hase and Bansal, 2020; Hase et al., 2020; Rajagopal et al., 2021; Poursabzi-Sangdeh et al., 2021; Wiegreffe et al., 2021, *inter alia*), and positively correlates with human judgement of rationale utility. We abstain from reference-based lexical-overlap metrics such as BLEU (Papineni et al., 2002), which are not suited for measuring plausibility (Camburu et al., 2018; Kayser et al., 2021; Clinciu et al., 2021), or faithfulness of rationales (Jacovi and Goldberg, 2020).

Evaluation via simulability of our generated rationales (§5) shows negative simulatability for CoS-E-generated, -13.1 (43.9 - 57.0), and for ECQA-generated rationales, -12.36 (44.64 - 57.00) where acc (I→O) = 57.0 (see Tab. 6). This is consistent with findings from Wiegreffe et al. (2021). Perhaps a better evaluation metric for this task is given by leakage-adjusted simulatability (Hase et al., 2020), where simulatability of non-leaky rationales and those of leaky ones is equally weighted. We leave a detailed investigation of rationale evaluation to future work.

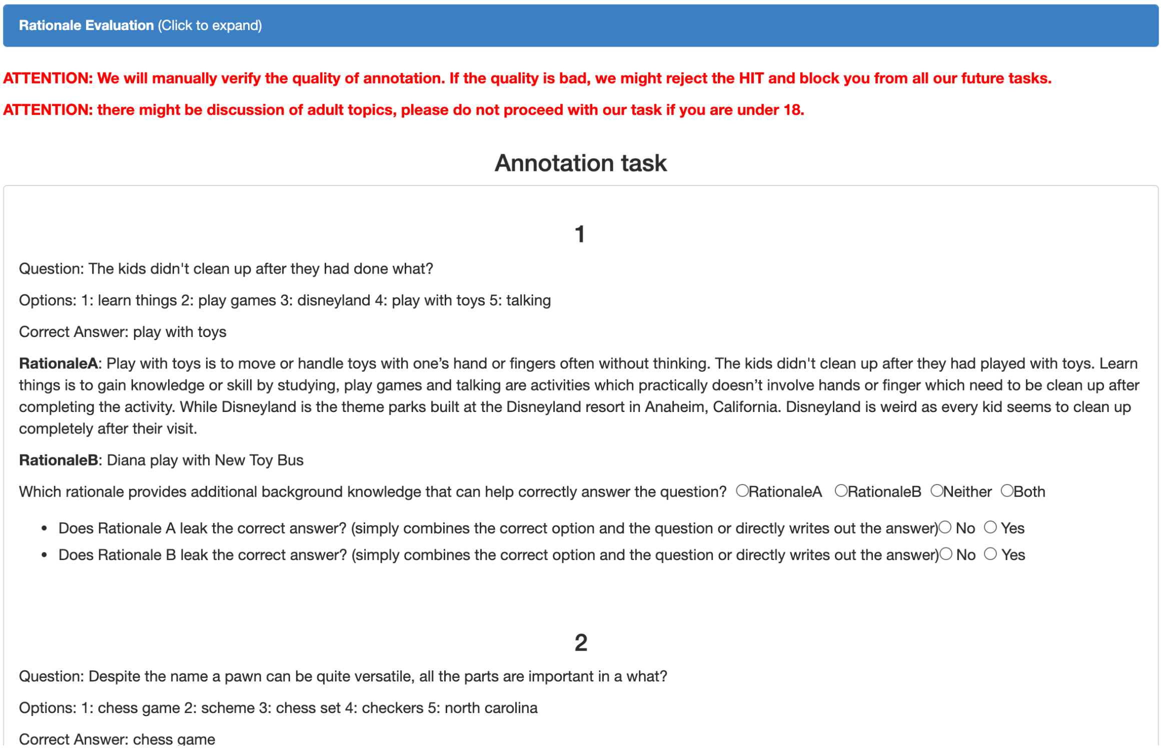
B Human Study Annotation

We recruit workers through our maintained list of qualified workers on Amazon Mechanical Turk (MTurk). These workers have collaboration with us on other projects, e.g., AESOP (Sun et al., 2021). In addition, we require workers to have completed over 1000 HITs with an approval rate over 99% and locate in the United States to qualify for our annotation task. As some of the questions contain discussion of adult topics, we warn workers and ask them to terminate the annotation if they are under 18. Our annotation pays for $1 per HIT.

Please see figures 2, 3, 4 for interfaces to our human studies, discussed in §3 and §5.

C Author Annotation Details

The 100 CoS-E instances that are annotated by three authors are the same ones annotated by workers on MTurk in the additional quality checks. Workers' annotation agrees with ours on 92 out of 100 instances for both background knowledge and leakage annotation, showing a good agreement between us and workers. We do the remaining annotations by ourselves to ensure annotation quality. During the annotation, one example of a $C_{\textit{no-leak-no-bg}}$ rationale is "*Rivers flow trough valleys.*", which occurs in 119 / 1221 instances (9.7% of the entire dev. set), even though it seemed valid for just one dev. instance. We suspect that this ra-



Rationale Evaluation (Click to expand)

ATTENTION: We will manually verify the quality of annotation. If the quality is bad, we might reject the HIT and block you from all our future tasks.

ATTENTION: there might be discussion of adult topics, please do not proceed with our task if you are under 18.

Annotation task

1

Question: The kids didn't clean up after they had done what?

Options: 1: learn things 2: play games 3: disneyland 4: play with toys 5: talking

Correct Answer: play with toys

RationaleA: Play with toys is to move or handle toys with one's hand or fingers often without thinking. The kids didn't clean up after they had played with toys. Learn things is to gain knowledge or skill by studying, play games and talking are activities which practically doesn't involve hands or finger which need to be clean up after completing the activity. While Disneyland is the theme parks built at the Disneyland resort in Anaheim, California. Disneyland is weird as every kid seems to clean up completely after their visit.

RationaleB: Diana play with New Toy Bus

Which rationale provides additional background knowledge that can help correctly answer the question? ○RationaleA ○RationaleB ○Neither ○Both

- Does Rationale A leak the correct answer? (simply combines the correct option and the question or directly writes out the answer)○ No ○ Yes
- Does Rationale B leak the correct answer? (simply combines the correct option and the question or directly writes out the answer)○ No ○ Yes

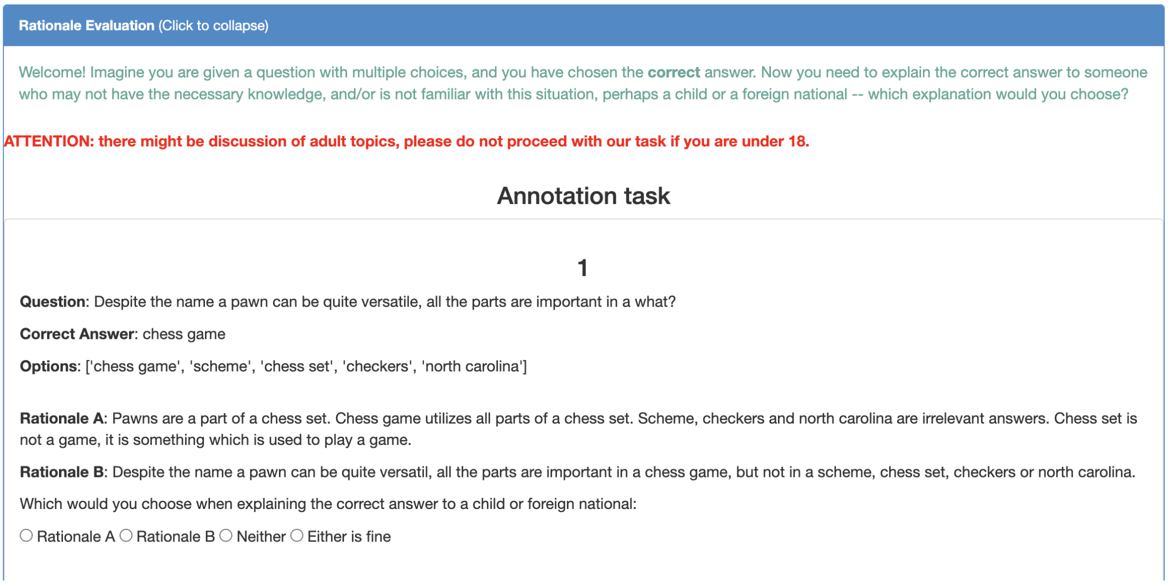
2

Question: Despite the name a pawn can be quite versatile, all the parts are important in a what?

Options: 1: chess game 2: scheme 3: chess set 4: checkers 5: north carolina

Correct Answer: chess game

Figure 2: The annotation interface for the direct assessment user study. We use the same interface for both comparing 1) crowdsourced rationales from CoS-E and ECQA (§3), as well as 2) crowdsourced v.s. generated rationales from those datasets (§5).



Rationale Evaluation (Click to collapse)

Welcome! Imagine you are given a question with multiple choices, and you have chosen the **correct** answer. Now you need to explain the correct answer to someone who may not have the necessary knowledge, and/or is not familiar with this situation, perhaps a child or a foreign national -- which explanation would you choose?

ATTENTION: there might be discussion of adult topics, please do not proceed with our task if you are under 18.

Annotation task

1

Question: Despite the name a pawn can be quite versatile, all the parts are important in a what?

Correct Answer: chess game

Options: ['chess game', 'scheme', 'chess set', 'checkers', 'north carolina']

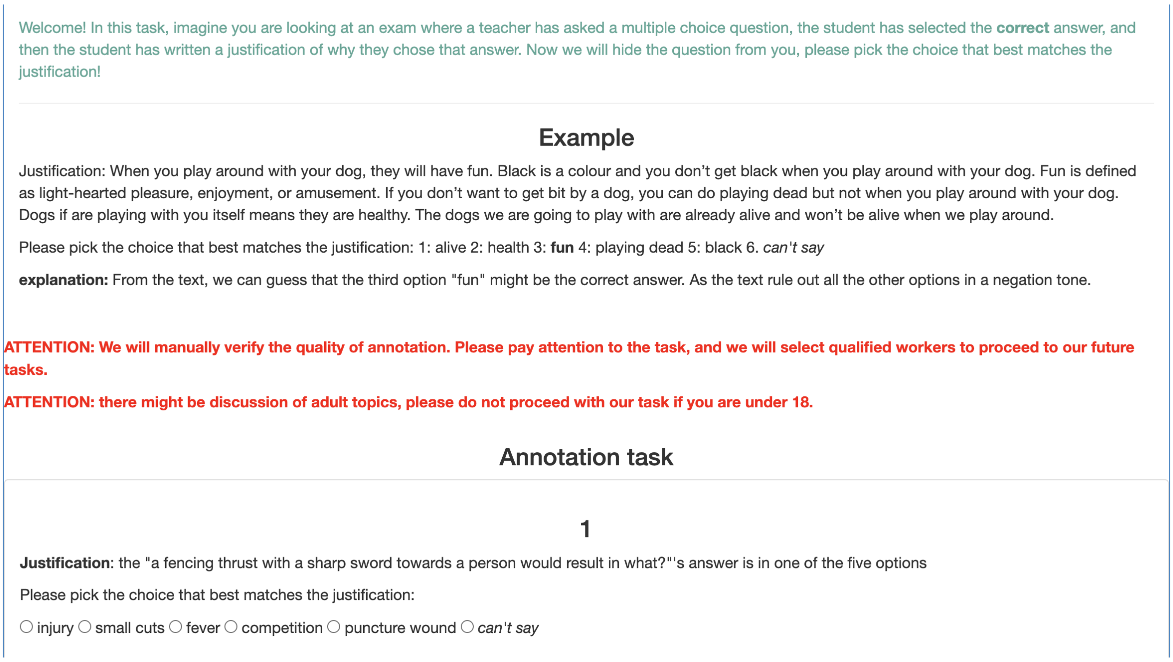
Rationale A: Pawns are a part of a chess set. Chess game utilizes all parts of a chess set. Scheme, checkers and north carolina are irrelevant answers. Chess set is not a game, it is something which is used to play a game.

Rationale B: Despite the name a pawn can be quite versatil, all the parts are important in a chess game, but not in a scheme, chess set, checkers or north carolina.

Which would you choose when explaining the correct answer to a child or foreign national:

○ Rationale A ○ Rationale B ○ Neither ○ Either is fine

Figure 3: The annotation interface for indirect assessment of if an annotated rationale has background information.



Welcome! In this task, imagine you are looking at an exam where a teacher has asked a multiple choice question, the student has selected the **correct** answer, and then the student has written a justification of why they chose that answer. Now we will hide the question from you, please pick the choice that best matches the justification!

Example

Justification: When you play around with your dog, they will have fun. Black is a colour and you don’t get black when you play around with your dog. Fun is defined as light-hearted pleasure, enjoyment, or amusement. If you don’t want to get bit by a dog, you can do playing dead but not when you play around with your dog. Dogs if are playing with you itself means they are healthy. The dogs we are going to play with are already alive and won’t be alive when we play around.

Please pick the choice that best matches the justification: 1: alive 2: health 3: **fun** 4: playing dead 5: black 6. *can't say*

explanation: From the text, we can guess that the third option "fun" might be the correct answer. As the text rule out all the other options in a negation tone.

ATTENTION: We will manually verify the quality of annotation. Please pay attention to the task, and we will select qualified workers to proceed to our future tasks.

ATTENTION: there might be discussion of adult topics, please do not proceed with our task if you are under 18.

Annotation task

1

Justification: the "a fencing thrust with a sharp sword towards a person would result in what?"'s answer is in one of the five options

Please pick the choice that best matches the justification:

○ injury ○ small cuts ○ fever ○ competition ○ puncture wound ○ *can't say*

Figure 4: The annotation interface for indirect assessment of if there is leakage in a rationale.

tionale was used as a default placeholder. We urge future researchers to closely inspect the annotation quality before releasing a dataset.

D Examples of Generated Rationales

Table 12 show examples of generated rationales for CoS-E and ECQA, together with crowdsourced rationales from the original datasets. Based on our human evaluation, the quality of generated rationales is worse than that of the crowdsourced rationales; hence, generated rationales provide weaker signals for training IR→O models.

Question	**Options**	**Crowdsourced** $R_{\text{CoS-E}}$	**Generated** $R_{\text{CoS-E}}$	**Crowdsourced** R_{ECQA}	**Generated** R_{ECQA}
What island country is ferret popular?	own home, north carolina, **great britain**, hutch, outdoors	great britain is the ferret populer	great britain island	Great britain is an island country and Great britain is popular for having ferret as pet. Own home is not island country. North carolina is not a country but state. Hutch is not country but type of cage. Outdoors is not any country.	Ferrets are not found in hutch. Ferrets are not found in own home or outdoors. Ferret is popular in great britain. Great britain is an island country.
Where do adults use glue sticks?	classroom, desk drawer, at school, **office**, kitchen drawer	offices employ only Adults.	glue sticks are used at school	Glue stick is a solid glue used to stick thin paper materials by adults in offices. Adults don't go to classroom and school, and other options don't have adults.	Kids do not use glue sticks in classroom. Desk drawer is a storage space and not a place to store glue sticks. Adults use glue sticks in a classroom. Office is a room or set of rooms or a building where adults work, usually sitting at desks. Kitchen drawer is a storage space and not a place for adults to use glue sticks. School is a place for learning and not a place to use glue sticks.
What could go on top of wood?	lumberyard, synagogue. floor, **carpet**, hardware store	Carpets in Chennai	carpets are made of wood	It is the carpet that could go on top of wood. Lumberyard is a place that cant go anywhere. Floor is at the bottom. Synagogue is a building that could not go on the top of wood. A hardware store can never go on the top of the wood.	Flooring cant go on top of wood. Hardware store or lumberyard or synagogue is not made of wood. Carpet can be made of wood.
Where is a doormat likely to be in front of?	facade, **front door**, doorway, entrance porch, hallway	Your front door makes an entrance	a doormat likely to be in front of entrance	Front door is the main entrance to a house. There is a doormat likely to be in front of the front door. Façade is the principal front and keeping doormat there doesn't solve the issue. Doorway is the entrance to room through a door and not a place where a doormat is likely to be placed. Entrance porch is a covered shelter in front of the entrance and doormat is not likely to be here. Hallway is not likely to have a doormat for wiping feet.	A doormat is likely to be in front of front door. Doorway is the entrance to a room through a door. Doormat cant be found in front of entrance porch or hallway. Façade is not a place where a doormat can be found.
The teacher doesn't tolerate noise during a test in their what?	movie theatre, bowling alley, factory, store, **classrooom**	Rivers flow trough valleys.	a classroom is a learning place	A teacher takes the test in a classroom where she does not tolerate any noise. All the other options are incorrect as a test is not conducted in any of the options.	A classroom is a place where one sits and takes test. All the other options are incorrect as they are not a place where one can sit and take test.
If you have to read a book that is very dry and long you may become what?	have time, boring, learn new, enjoyable, **bored**	a book is world.	boring is when you have to read a book	You may get bored while reading a long and dry book. While reading a long book, you dont have time. A book that seems to be very dry is boring. You cant become learn new. A book which is very dry and long can never be enjoyable.	Dry and long books make us feel bored. Boring comes under having time and the other options are not caused by reading dry and long books. So if you have to read a book that is very dry and long you may become bored.
If you are awaking multiple times throughout the night because a lot is on your mind, what is a likely cause?	irritability, **depression**, getting out of bed, happiness, discomfort	Depression indicates mental distress.	mental or emotional states	Depression is an illness of intense sadness and having a lot in mind while not letting one to sleep is likely caused by depression. The other options do not cause one to have a lot in mind.	Depression is not caused by waking up multiple times throughout the night. Getting out of bed is not a cause. Happiness cant be caused by waking up multiple times throughout the night. By waking multiple times throughout the night, a lot of things are on your mind.Discomfort is a likely cause of irritability.
What do you want someone to do when you illustrate point?	did not understand, accepting, make clear, understood, **understanding**	we need a understanding.	make clear what do you want someone to do when you illustrate	To illustrate is to make something clearer and more visible. We want someone to understand when we illustrate point. Did not understand being opposite of what you want someone to do, accepting and Make clear does not relate to and while understood is a past tense.	When someone illustrates point, they want to make clear about it. To understand something is not appropriate here. If someone did not understand the point illustrated then he will not understand it. Someone cant accept or understand the point if he illustrates it.

Table 12: Examples of crowdsourced and generated rationales from CoS-E v1.11 and ECQA. The ground truth options are **in bold**. As can be clearly seen, the crowdsourced CoS-E rationales are often ungrammatical, and off-topic, and do not provide the background knowledge necessary to understand the ground truth answers. In contrast, crowdsourced ECQA rationales are grammatical, and provide the necessary background knowledge for human interpretability. Moreover, generated rationales are often factually incorrect, such as "*carpets are made of wood*", and also lack much of the commonsense reasoning necessary for rationales. All types of rationales leak the correct answer.