Investigating the Benefits of Free-Form Rationales

Anonymous ACL submission

Abstract

Free-form rationales aim to aid model interpretability by supplying the background knowledge that can help understand model decisions. Crowdsourced rationales are provided for commonsense QA instances in popular datasets such as CoS-E and ECQA, but their utility remains under-investigated. We present human studies which show that ECQA rationales indeed provide additional information to understand a decision, while 70% 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 human users? We investigate the utility of rationales as an additional source of supervision, by varying the quantity and quality of rationales during training. After controlling for instances where rationales leak the correct answer, we find that incorporating only 5% of rationales during training can boost model performance by 16.89%. Moreover, we also show that rationale quality matters: compared to crowdsourced rationales, T5-generated rationales provide not only much weaker supervision to models, but are also not helpful for human users in aiding model interpretability.

1 Introduction

Interpretable natural language processing (NLP) benefits from faithful rationales that are accurate representations of model's decision process (Alvarez-Melis and Jaakkola, 2018). These rationales aim to explain decisions by providing additional world knowledge or commonsense reasoning, necessary for most language understanding tasks.[1] Free-form rationales also come with the promise of being easily interpretable by humans, as opposed to other kinds of explanations, such as extractive rationales in form of textual highlights (Camburu et al., 2018a), or low-level neuron activations in neural architectures (Hohman et al., 2020).

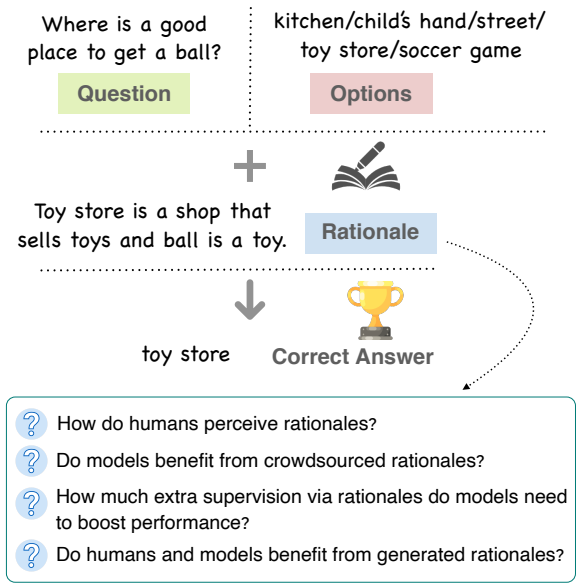

Figure 1: Our work studies the benefits of rationales. We conduct *human* studies to understand perceived utility of rationales and study if *models* can benefit from crowdsourced rationales as extra supervision. Furthermore, we study how much extra supervision models need to boost the performance. Finally, we compare generated rationales with human-annotated rationales. Here we use commonsense QA as an example for illustration.

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., 2018a) for natural language inference, SBIC (Sap et al., 2020) for social bias inference, among others; most of these corpora contain crowdworker-written free-form rationales (Wiegreffe and Marasović, 2021). Despite the relative ease of crowdsourcing, collecting high-quality free-form rationales is challenging: for instance,

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

Narang et al. (2020) find that the rationales in CoS-E are of lower quality, which might defeat the purpose behind collecting such corpora. ECQA (Aggarwal et al., 2021) builds on CoS-E dataset and re-annotates for better quality.

However, the utility of rationales is still unclear. 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 a carefully designed comparative human study for commonsense question answering. We find that rationales from different corpora have different capabilities: humans find ECQA rationales provide additional information that can help answer questions, while only 30% of CoS-E rationales do.

Inspired by this finding, we further ask: analogous to the benefit provided to human users, can crowdsourced rationales benefit *models* by providing additional supervision that results in a performance boost? 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 models' classification capabilities. Our results indicate that while crowdsourced rationale 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. Under this setting, we find that including only 5% of rationales during training can improve model performance by 16.89% at inference time with good rationales. This finding generalizes to Quartz (Tafjord et al., 2019), a dataset for textual relationship inference, in which the provided background knowledge is intentionally designed to contain no leakage.

Finally, we investigate if automatically generated rationales provide similar benefits as crowdsourced rationales. Our human studies indicate that the perceived utility of generated rationales is much lower than that of human-written ones. Moreover, we find that generated rationales from T5 (Raffel et al., 2020) are not ready yet to serve as supervision signals and cannot help improve model performance in the distant supervision setting. These results indicate that the quality of rationales is paramount for both human interpretability and model supervision.

To summarize, our work focuses on understanding the utility of free-form rationales in existing popular datasets, from both the human and modeling perspectives. To the best of our knowledge, none of previous works have quantitatively examined the utility of free-form rationales.[2]

Dataset	Train	Test
CoS-E v1.0	7,610	950
CoS-E v1.11/ECQA	9,741	1,221
QuaRTz	2,695	783

Table 1: The statistics of 3 datasets in our work.

2 Preliminaries

Tasks and Datasets. We explore three large datasets containing crowdsourced free-form natural language rationales. The first two address the commonsense-based question answering (Talmor et al., 2019, ComQA) task: CoS-E (Rajani et al., 2019), and ECQA (Aggarwal et al., 2021). 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. 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. including rationales for choosing the correct option and rejecting other options.

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. 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*. Each instance contains a triplet: a situated qualitative question, with two answer options and a knowledge statement, helpful to answer the question. 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 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

[2]We will publicly release our annotated data and code.

understand or derive the correct answer. The statistics of the three datasets are in Table 1.[3]

Models. We use finetuned T5 (Raffel et al., 2020) models throughout our work following prior efforts for analysing (Wiegreffe et al., 2021) and generating (Narang et al., 2020; Lakhotia et al., 2021) free-text explanations. More specifically, we finetune three T5-base models for each dataset, with ground-truth labels and rationales.[4]

- I→O. We directly finetune a model to predict the correct option, and we format the I/O of this T5 model as `context:` *{question}* `options:` *{concatenated option string}* for input and *{correct option}* for output.
- IR→O. Different from the I→O model, we add rationales into input and have: `context:` *{question}* `options:` *{concatenated option string}* `explanation:` *{rationale}* for input, the output is *{correct option}*.
- I→R. We finetune a model supervised by human-annotated rationales to generate rationales. The format of I/O is `explain question:` {question} `answer:` *{concatenated option string}* for input and `explanation:` *{rationale}* for output.

For the IR→O model, we experiment with different variations based on the source of the rationales, R (e.g. R_{CoS-E}). In addition, rationales used in our analysis could be either from human annotation (R_{crowd}) or model generations ($R_{generated}$).

Evaluation. We use accuracy (*acc*) to evaluate the performance of both I→O and IR→O models. We measure the benefit of rationales as extra supervision by using simulatability score, which reflects the utility of rationales in terms of improving model's performance:

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

We do not report lexical-overlap metrics as our primary evaluation metric because these are not suited for measuring plausibility (Camburu et al., 2018b; Kayser et al., 2021; Clinciu et al., 2021) or faithfulness of rationales (Jacovi and Goldberg, 2020). In contrast, simulatability score (Eq. 1) from humans (Doshi-Velez and Kim, 2017) has been serving as a reliable measure of rationale quality from the lens of utility to an end user in prior literature (Hase and Bansal, 2020; Hase et al., 2020; Rajagopal et al., 2021; Poursabzi-Sangdeh et al., 2021; Wiegreffe et al., 2021, i.a.). Simulatability additionally measures the predictive ability a rationale provides over the input, unlike lexical-overlap metrics. Hase et al. (2020) also verify that simulatability scores positively correlates with human judgement for the rationale utility.

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

[4] See Appendix A for details on our T5 model training.

	ECQA	CoS-E	neither	both
has background knowledge?	65.0%	9.2%	5.0%	20.8%
leaks answer?	83.3%	43.3%	n/a	n/a

Table 2: Human study results on the perceived utility of rationales, on 120 ComQA instances with rationales from ECQA and CoS-E. Most ECQA rationales provide additional information to help humans answer the questions, but frequently leak the correct answer. 70% CoS-E rationales do not provide any additional background knowledge, and 40% leak the correct answer.

3 How do Humans Perceive Rationales?

Free-text rationales purpotedly improve human interpretability by explaining the model decisions in natural language for the benefit of human users. However, how successful are current crowdsourced rationales in providing the additional background knowledge to this end?

We conduct a human study to understand how humans perceive the utility of rationales. For each instance, annotators are presented the question, options, correct answer and rationales from two crowd-annotated sources, $\mathcal{A}$ and $\mathcal{B}$. Annotators are tasked to answer which rationale provides *additional background knowledge* that can help them answer the question. Four choices are given: rationale $\mathcal{A}$, rationale $\mathcal{B}$, neither and both (Q1).[5]

We are additionally interested in whether the rationales simply leak the answer by revealing it in the rationale, regardless of whether they provide additional background knowledge. To this end, annotators are asked if each rationale (independently) leaks the answer (Q2 and Q3).

We use first 120 annotated rationales in both ECQA and CoS-E v1.11 in our study. Rationale $\mathcal{A}$ and rationale $\mathcal{B}$ represent ECQA annotation and CoS-E annotation separately. We conduct our study on Amazon Mturk and each instance requires an-

[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.

	Zero-Shot UnifiedQA	Finetuned UnifiedQA	Finetuned T5
I→O	60%	**66%**	65%
$IR_{\text{CoS-E v1.0}}$ →O	70%	88%	**89%**
I→O	45%	**56%**	**56%**
$IR_{\text{CoS-E v1.1}}$ →O	54%	76%	**78%**
IR_{ECQA} →O	86%	**98%**	**98%**

Table 3: Model accuracy improves with rationales from CoS-E v1.0, CoS-E v1.11 and ECQA as additional supervision signals. We see great improvements under both a zero-shot setting with UnifiedQA, as well as after finetuning UnifiedQA and T5. All rationales R here are from crowdsourcing, hence based on the gold label.

notation from three independent annotators.[6] Post collection, we calculate the inter annotator agreement (IAA) with Fleiss's Kappa (Fleiss and Cohen, 1973). The IAA for Q1, Q2 and Q3 are 0.43, 0.26, 0.30 separately, indicating a moderate agreement. We take the majority vote of users as the final label.

Table 2 shows the result for human evaluation of if annotated rationales can provide additional knowledge. About 85.8% ECQA rationales can provide additional background knowledge to help answer the question, while only 30% CoS-E rationales can achieve the same, confirming the higher quality of ECQA annotations for human interpretability (Aggarwal et al., 2021). However, both ECQA and CoS-E rationales leak correct answers. Indeed, ECQA rationales *often reveal* the correct answer, in addition to providing the background knowledge necessary for humans to understand the decision.

4 Can Models Benefit from Crowdsourced Rationales?

In the previous section (§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 the same information can also benefit machine learning models, by providing them additional supervision, to make better decisions.

4.1 Rationales as Model Supervision

As a first empirical investigation, we use rationales from CoS-E v1.0 and CoS-E v1.11 as additional supervision to task models, following the IR→O set up, as defined in §2. As a baseline, we use the I→O set up, which does not have access to any rationales at either train or test time. We consider three models: UnifiedQA (Khashabi et al., 2020) in a zero-shot setting, as well as UnifiedQA[7] and T5 (Raffel et al., 2020), finetuned on the respective CoS-E benchmarks. We choose UnifiedQA, a state-of-the-art T5-based question answering model, because it not only performs well across twenty QA datasets, but also shows great generalization to out-of-domain data.

Our results in Table 3 show that under each setting, using crowdsourced rationales both during training and inference, greatly improve model performance, even under a zero-shot setting. With finetuning, T5-base performs comparably to UnifiedQA.[8] Most remarkably, ECQA performance is almost perfect after including rationales during finetuning. However, as our human study in §3 suggested, the improvement may come from the direct leakage of correct answer. Thus, we next investigate workarounds to address the leakage problem in CoS-E and ECQA.

4.2 Examining Crowdsourced Rationales

CoS-E: Although Narang et al. (2020) criticize the quality of CoS-E annotation, they do not provide a detailed study of the various deficiencies in the rationales. Nevertheless, CoS-E v1.11 is still widely used for additional commonsense knowledge (Ye et al., 2019), analysis (Majumder et al., 2021; Wiegreffe et al., 2021) and commonsense reasoning (Paranjape et al., 2021). Therefore, it is imperative for the community to understand the quality of annotated rationales in CoS-E.

Motivated by their utility for model supervision, rationales can be categorized as:

C_{leak}: simply state the answer, or combine the correct answer with the question,

$C_{no\text{-}bg}$: neither provide any additional background information, nor leak the correct answer, and

C_{bg}: do not leak correct answers but provide additional helpful background knowledge.

With these criteria in mind, one of the authors annotated 1,221 instances in the development set of CoS-E v1.11 dataset into 3 categories. Table 4 shows the distribution of the categories, examples from each category picked at random, together with

[6] See more annotation details in Appendix B.

[7] We use released models and instructions for finetuning from https://github.com/allenai/unifiedqa.

[8] This justifies our choice of T5 for later experiments. Additionally, given the higher difficulty of the CoS-E v1.11 task (5 answer choices, vs. 3 in v1.0), we use CoS-E v1.11 for the rest of our analysis.

	Example	Reason	Ratio
C_{leak}	Question: Who is a police officer likely to work for? Options: 1: beat 2: direct traffic 3: **city** 4: street 5: president ***Rationale***: a police officer likely to work for city	Directly combines the question and the correct option	38.08% (465/1221)
$C_{\text{no-bg}}$	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 relavant	Rationales are generic and do not provide additional background information	41.03% (501/1221)
C_{bg}	Question: If I want to watch a movie without leaving my home what might I use? Options: 1: drive in movie 2: drive in movie 3: **television** 4: video store 5: show ***Rationale***: The common watching device at home is a tv set	Provides background information, without leaking the answer	20.88% (255/1221)

Table 4: Our manual categorization of 1,221 CoS-E v1.11 (dev.) instances into 3 categories, with corresponding examples. Options in bold are correct options. C_{leak} and $C_{\text{no-bg}}$ make up over 79% of the development set of CoS-E rationale annotation.

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 5: Example annotations from CoS-E v1.11 and ECQA for 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 correct option *people*. CoS-E annotation directly combines the question and the correct answer, while ECQA annotation provides additional background knowledge.

the reason why we annotated them as the corresponding categories. Rationales under the C_{leak} and $C_{no\text{-}bg}$ categories make up over 79% of the entire development set of CoS-E v1.11.[9] Using the development set as a lens, our annotation provides a qualitative and quantitative understanding of the crowdsourced rationales in CoS-E. Future research should be careful when using rationales from CoS-E as additional knowledge or explanations.

ECQA: Aggarwal et al. (2021) build on CoS-E question-answer pairs and reannotate the rationales. Table 5 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 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* with the rationale for the correct option, followed by all other incorrect options. This ordering introduces a spurious correlation which likely provides a shortcut to the model for predicting the correct answer, but for wrong reasons. To address this issue, we randomly shuffle the rationales for different answer choices within each ECQA instance.[10]

4.3 Revisiting Rationales as Supervision

Taking into account our findings from the detailed analysis above (§4.2), we revisit including rationales as supervision for task models (following §4.1), but with a finer-grained understanding of these rationales. 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 performance. During inference, we provide the T5 models with rationales under each of the three categories of CoS-E, as dicussed above, as well as all combined together. For ECQA, we report performance for inference with and without shuffled rationales. Finally, we also study how supervision from one dataset affects another, in a transfer learning setting.

Table 6 (r: row number, c: column number) shows the accuracy of T5 models under all the above settings, showing the mean and standard deviation under three random initializations. We summarize our findings from Table 6 below.

Rationales boost model performance. First, comparing $c1$ with the rest of columns ($c2$-$c7$), rationales can help improve model's ability to make the correct prediction. After adding 5% of training data, the model reaches 60.88% accuracy with C_{bg}

[9] One example of a $C_{no\text{-}bg}$ 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 rationale was used as a default placeholder for annotators.

[10] We use the sentencizer in Spacy (https://spacy.io/) and random permute their order, with seed 0.

			$c1$	$c2$	$c3$	$c4$	$c5$	$c6$	$c7$
		%R	**I**	**I+R**$_{\text{CoS-E}}$	I+R$_{\text{ECQA}}$		I+R$_{\text{CoS-E}}$		
					w/o shuffle	**shuffled**	$C_{\textbf{leak}}$	$C_{\textbf{no-bg}}$	$C_{\textbf{bg}}$
r1	-	0%	57.00	47.09	53.32	54.95	55.43	46.41	52.10
r2	CoS-E	5%	$53.78_{1.10}$	$73.03_{2.01}$	$76.50_{2.30}$	$65.57_{2.86}$	$89.97_{3.14}$	$54.48_{0.73}$	$60.88_{1.57}$
r3		10%	$54.44_{0.72}$	$76.14_{1.07}$	$80.78_{1.53}$	$63.74_{0.78}$	$94.30_{1.20}$	$55.61_{0.97}$	$64.87_{1.50}$
r4		20%	$53.62_{0.23}$	$77.18_{0.58}$	$83.40_{1.41}$	$62.71_{1.80}$	$95.56_{0.23}$	$54.78_{0.59}$	$70.06_{2.13}$
r5		30%	$53.12_{0.60}$	$77.40_{0.20}$	$79.17_{3.23}$	$63.56_{1.28}$	$95.94_{0.16}$	$54.11_{1.04}$	$72.06_{1.23}$
r6		100%	48.24	78.46	66.01	64.46	96.88	55.16	73.65
r7	ECQA-shuffl.	5%	$54.05_{0.95}$	$59.43_{0.78}$	$86.65_{1.10}$	$86.35_{1.54}$	$65.41_{1.43}$	$53.06_{1.49}$	$54.69_{1.13}$
r8		10%	$54.05_{1.08}$	$61.80_{2.24}$	$92.55_{0.52}$	$93.01_{0.37}$	$69.57_{3.22}$	$52.84_{1.25}$	$57.49_{2.94}$
r9		20%	$53.29_{0.32}$	$66.20_{0.76}$	$95.41_{0.48}$	$94.70_{1.17}$	$74.67_{1.28}$	$55.91_{1.32}$	$62.88_{1.76}$
r10		30%	$52.85_{0.67}$	$65.11_{0.91}$	$95.85_{0.34}$	$95.52_{0.51}$	$76.31_{1.68}$	$52.62_{0.56}$	$57.68_{2.69}$
r11		100%	38.08	67.24	97.3	96.56	90.46	37.22	62.87

Table 6: Model accuracy under I→O / IR→O settings, w.r.t. fine-grained understanding of the quality and quantity of rationales. At inference time, we use the full set of annotated rationales, or by category for CoS-E. All reported numbers are 3-seed average accuracy with the standard deviation in the subscript. ***r***ows and ***c***olumns are settings for training and testing separately. $r1$ is the baseline performance without having rationales during training.

			$c1$	$c2$	$c3$
	Source	**Quantity**	**I**	**I+R_** ECQA	**I+R_** ECQA **_shuffle**
r1	-	0	57.00	53.32	54.95
r2	ECQA	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 ECQA rationales. Without shuffling, the model relies on the spurious correlation due to sentence order. The accuracy soars to 93.94% with only 5% training data without shuffling on unshuffled test data, but can only have 76.66% on shuffled test data. Therefore, we use ECQA with shuffling for our experiments.

rationales, which yields 16.89% improvement compared to 52.10% without rationales. Please note that we exempt R1 from comparison because the model does not have any supervision from rationales during training, making it hard to understand the utility of rationales during the inference time.

Rationales help transfer learning. Second, adding more rationales to training will help boost model's performance for data from the same distribution. $r3$-$r6$ under $c2$ and $r7$-$r11$ under $c4$ are two examples. Adding rationales from another type can help model improve performance to an extend (e.g., $r3$-$r6$ under c4 and $r7$-$r11$ under $c2$). However, models would perform worse when adding more rationales (20%→30%) from another type of rationale. We suspect this is because the model overfits to one specific rationale type and raise the distributional shift issue.

Rationale quality matters. Last, the quality of rationales affects performance for both cases: 1) adding rationales to training, and 2) use rationales for model inference. The former is supported by the comparison between $r2$-$r6$ under $c2$ and $r7$-$r11$ under $c4$: using ECQA rationales will yield a better performance. Meanwhile, $r2$ to $r6$ under $c6$ supports the latter argument. Adding poor quality rationales does not help model reasoning.

To validate the importance of shuffling ECQA rationales, we add randomly-picked 5% ECQA rationales without shuffling into training and test the model performance on both shuffled and unshuffled ECQA rationales. Comparing $c2$ and $c3$ in Table 7, we see that with only 5% unshuffled ($r2$), the accuracy rises from 53.32 to 93.94. However, when we test the model on shuffled data, the accuracy is 76.66%. The experimental result suggests that the model learns spurious correlation between the rationale and correct answer before shuffling. Therefore, we shuffle the order of sentences in the ECQA annotation to prevent the model from learning this spurious correlation.

Non-leaky rationales still boost model performance. Despite taking care to prevent spurious correlations in ECQA, there is still a chance models benefit from some amount of leakage of the correct answer. To control for this, we consider the QuaRTz dataset, introduced in §2, using knowl-

		%R	I	I+R$_{\text{QuaRTz}}$
r1	I→O	-	70.88	38.27
r2	IR$_{\text{QuaRTz}}$ →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 supervision from knowledge statements as rationales. Even perfectly non-leaky rationales improve model performance, showing the generalizability of our conclusions. All reported numbers show the average with standard deviation with three random initializations.

Setting		$\mathcal{RA}$	$\mathcal{RB}$	Neither	Both
$\mathcal{RA}$: ECQA ann. $\mathcal{RB}$: ECQA gen.	useful	43.44%	22.50%	15%	19.17%
	leakage	89.17%	64.17%	n/a	n/a
$\mathcal{RA}$: CoS-E ann. $\mathcal{RB}$: CoS-E gen.	useful	28.33%	20%	34.17%	17.5%
	leakage	51.67%	40.83%	n/a	n/a

Table 9: Human evaluation in comparison studies for (1) ECQA annotated rationales v.s. ECQA generated rationales and (2) CoS-E annotated rationales v.s. CoS-E generated rationales. Humans perceive fewer generated rationales to provide additional background knowledge than human-annotated rationales. Meanwhile, ECQA rationales have better quality than CoS-E rationales for both the generated and the human-annotated.

		R$_{\text{CoS-E}}$ generated	R$_{\text{ECQA}}$ generated
R_CoS-E generated	5%	$44.34_{1.59}$	$45.1_{0.86}$
	10%	$44.94_{0.59}$	$42.89_{0.46}$
	20%	$44.34_{0.71}$	$41.17_{0.74}$
	30%	$44.91_{0.43}$	$39.83_{0.56}$
	100%	**43.9**	35.71
R_ECQA generated	5%	$46.33_{0.54}$	$44.64_{1.03}$
	10%	$45.10_{0.34}$	$44.96_{0.30}$
	20%	$46.98_{0.83}$	$45.67_{0.37}$
	30%	$45.81_{0.60}$	$45.51_{0.40}$
	100%	43.16	**44.64**

Table 10: Use generated rationales as extra supervision. We add different amount of generated rationales into training. All reported numbers show the average with standard deviation with three random initializations.

edge statements as rationales, which are guaranteed to contain no leakage. We use the same modeling strategy as before and finetune T5 models for both I→O and IR→O models on QuaRTz. Without rationales, T5 model performs comparably to the BERT-PFT (IR) model which scores 73.7, one of the most promising models reported in (Tafjord et al., 2019). It shows that QuaRTz dataset is hard, and finetuning T5 is a feasible modeling strategy to apply for QuaRTz dataset.

Table 8 shows the experimental results of the distant supervision setting for QuaRTz, which validates our previous finding that rationales help to improve model's ability to predict correct answer, here adding 30% of rationales can bring about 0.89% accuracy improvement. Meanwhile, keep adding more rationales further boosts model's performance. The consistency with our previous findings shows the generalizability of our conclusions.

5 Benefits of Generated Rationales

So far, we have focused on crowdsourced rationales, written by humans. However, there has been a lot of research on generating free-form rationales using T5 (Narang et al., 2020; Paranjape et al., 2021). Based on this, we ask: 1) can generated rationales provide the additional background information necessary for humans to interpret and answer questions, similar to §3, and 2) can generated rationales provide additional supervision to improve model's prediction accuracy?

Human Perception of Generated Rationales? We repeat our studies in §3 on two new sets of comparison studies: (1) generated rationales of ECQA v.s annotated rationales in ECQA and (2) generated rationales of CoS-E v.s. annotated rationales in CoS-E. Table 9 shows the annotation result. We find that human perceive fewer generated rationales to provide additional background knowledge than human-annotated rationales. Meanwhile, ECQA rationales have better quality than both the human-annotated and generated rationales from CoS-E.

Generated Rationales as Model Supervision. First, we use annotated rationales to train I→R models for both CoS-E and ECQA following steps from §2. Then, we use the I→R models to generate rationales and add generated rationales to IR→O model training with various amount. During the inference, we also use generated rationales from the I→R models[11].

Table 10 shows the experimental result of the distant supervision setting with generated rationales. We share the same finding with Wiegreffe et al. (2021) that using generated rationales does not help improve model's performance in terms of predicting correct option, leading to negative simulatability scores, which are -13.1 (43.9-57) for CoS-E generated rationales and -12.36 (44.64-57) for ECQA generated rationales. This finding addresses the

[11] We show an example of annotated and generated rationales for both CoS-E and ECQA in Appendix D.

importance of having good quality rationales for the I→R model training, as we have concluded that ECQA annotations are of better quality than CoS-E. Under the distant supervision setting, although generated rationales do not help improve model performance for prediction compared to the vanilla I→O model, using more generated rationales in training keeps boosting the model performance to an extend (from 5% to 30%), which is consistent with our previous conclusion using human annotated rationales. However, using 100% generated rationales leads to a performance drop compared to only using 30%, we suspect that this is because the generated rationales introduce too much noise and the model fails to learn a clear pattern. These conclusions are consistent with our human studies.

6 Related Work

Types of Explanations. Rationales can be roughly categorized into two broad categories: extractive rationales and free-form rationales. Extractive rationales serve interpretability in that they can reveal the "reasoning" behind model outputs. Therefore, extractive rationales are usually grounded in a specific context (e.g., such as a paragraph) as supportive evidence. For example, in information extraction (IE) tasks, a rationale can be extracted as a subset of the input and is sufficient to make a prediction on its own without relying on the rest of the input. DeYoung et al. (2020) introduce ERASER that comprises 8 datasets and 9 tasks with human annotation of extractive rationales. Free-text rationales take the form of free-form natural language to fill in the reasoning or knowledge gap. There have been fewer datasets focusing on introducing free-form rationales compared to extractive rationales. In addition to e-SNLI (Camburu et al., 2018a), CoS-E (Rajani et al., 2019), QuaRTz (Tafjord et al., 2019) and the most recent ECQA (Aggarwal et al., 2021) provide necessary knowledge for answering questions.

Rationale Generation. From the modeling perspective, rationale generation models can be roughly categorized into supervised and unsupervised models. 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) propose to utilize T5 to complete contrastive explanation prompts that explicitly contrast different possible answers in its explanation. Under the line of contrastive explanation generation, Jacovi et al. (2021) manipulate the latent space, differentiate two potential decisions and construct explanations to answer for which the given label is useful. Following prior work, we also finetune T5 models to generate rationales.

Learning From Rationales. There has been limited work studying the problem of training models to learn from human-annotated free-form rationales. Wiegreffe et al. (2021) investigate how free-form rationales and model predicted labels are associated, and use it as a property to evaluate the faithfulness of rationales; in contrast our work provides a much more detailed study across multiple datasets. Carton et al. (2021) leverage extractive rationales and show a consistent trend that using rationales can improve model performance. Most similar to our work, Huang et al. (2021) noticed that the quality of rationales would have a huge impact and explore the utility of extractive rationales in the distant supervision setting. Our work has a similar motivation to Huang et al.'s and follows a similar setting for free-form rationales.

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 higher quality rationales are helpful as additional supervision sources for task models. Our work highlights the importance of inspecting the quality of human-annotated rationales before using them for additional model supervision. We also found that generated rationales are not as useful for human interpretability or for model supervision, as opposed to crowdsourced rationales. Our investigations shed light on fundamental assumptions about human interpretability in collecting and generating rationales, and calls 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. The data collection in this work has been approved by the IRB board in our institute. For modeling, we utilizes T5 throughout our work, which also involves generating rationales. Trained on massive online texts, it is well-known that such pretrained language models could capture the bias reflecting the training data. Note that our released models might be used for malicious purposes because we do not have a filtering mechanism that checks the toxicity, bias, or offensiveness of source sentences from the input. We suggest interested parties carefully check the generated content before using our trained models in any real-world applications. Three datasets in our works are all public datasets. 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.

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*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018a. e-snli: Natural language inference with natural language explanations. *CoRR*, abs/1812.01193.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018b. 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.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*.

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.

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, 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.

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. *ArXiv*, abs/2105.03761.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, 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.

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.

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.

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.

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.

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.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. NeurIPS 2021 (neurips 2021 datasets and benchmarks track).

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.

A Implementation Details for Finetuning T5

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 the inference time, we use beam search as the decoding method with a beam size of 2. The generation of EOS token or reaching the maximum target length will terminate the decoding.

B Human Study Annotation

We conduct our three human studies in our work. Figure 2 and Figure 3 show the interfaces we build on Amazon Mturk for annotation guideline and workspace separately. 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.

C $C_{no\text{-}bg}$ Example in CoS-E

D Examples of Generated Rationale

Table 11 shows an example of generated rationales for CoS-E and ECQA, together with annotated rationales in the original dataset. Based on our human evaluation, the quality of generated rationales are worse than annotated rationales and cannot provide proper supervision signals for model training.

Source	Rationale
CoS-E annotation	nourishment is a work
CoS-E generation	the dog needs lots of attention
ECQA annotation	Lots of attention is a special care or notice taken of someone or something, the regarding of someone or something as interesting or important. Aside from water and nourishment, our dog needs lots of attention. Bone is a treat that our dog will enjoy, and bone is nourishment and not what our dog needs the most. Charm is a quality of fascinating others and our dog doesn't have to fascinate others. Petted is to show affection and not a care that a dog requires. Walked is moved at a regular pace and that a dog can do by his own and not what he needs from you.
ECQA generation	Aside from water and nourishment, your dog need lots of attention. Bone is the part of human body which provides nourishment and rest. Petted is to be taken care of while lots of attention is not. Walked is done by a dog and not an animal. Charm is not related to nourishment and water.

Table 11: Exampled of annotated and generated rationales from CoS-E v1.11 and ECQA for question *"Aside from water and nourishment what does your dog need?"* with options *1: bone 2: charm 3: petted 4: lots of attention 5: walked* and correct option *lots of attention*.

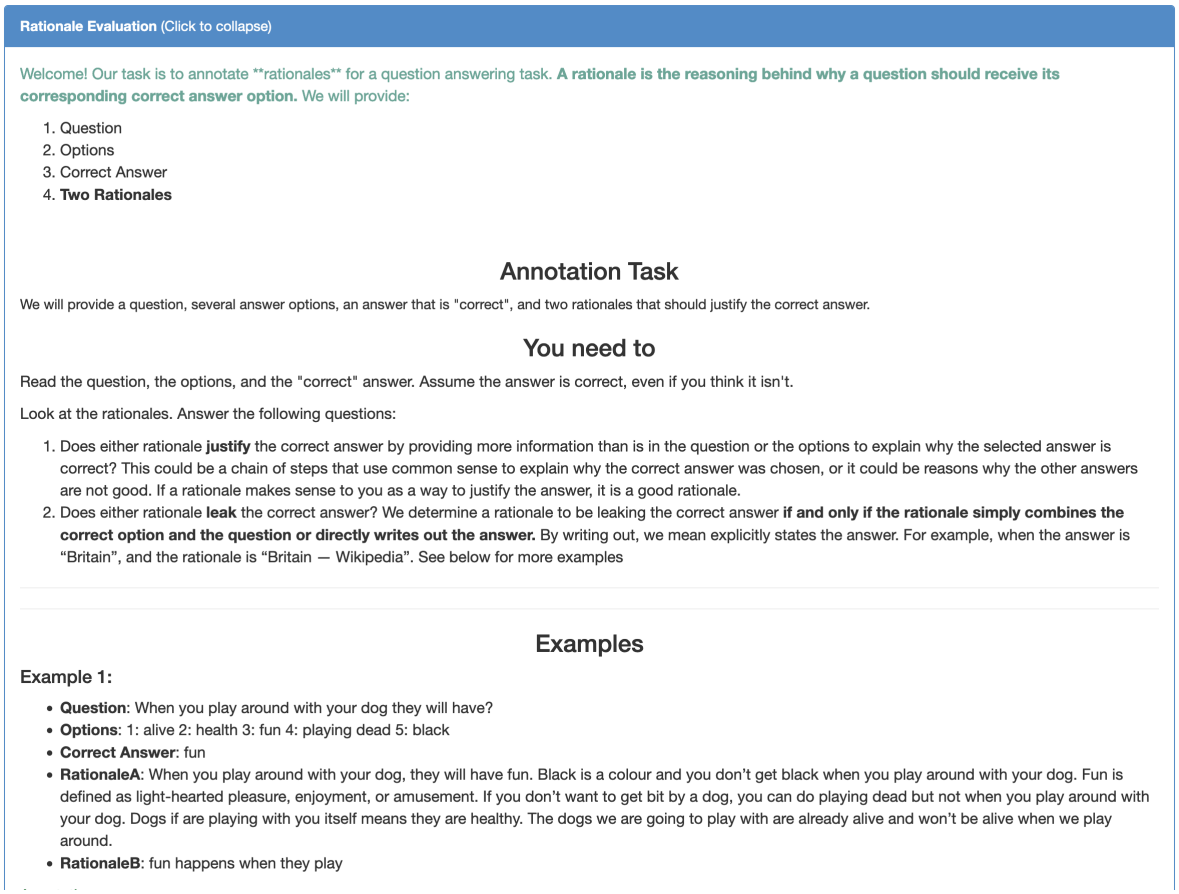

Figure 2: Part of the annotation guideline on Amazon Mturk.

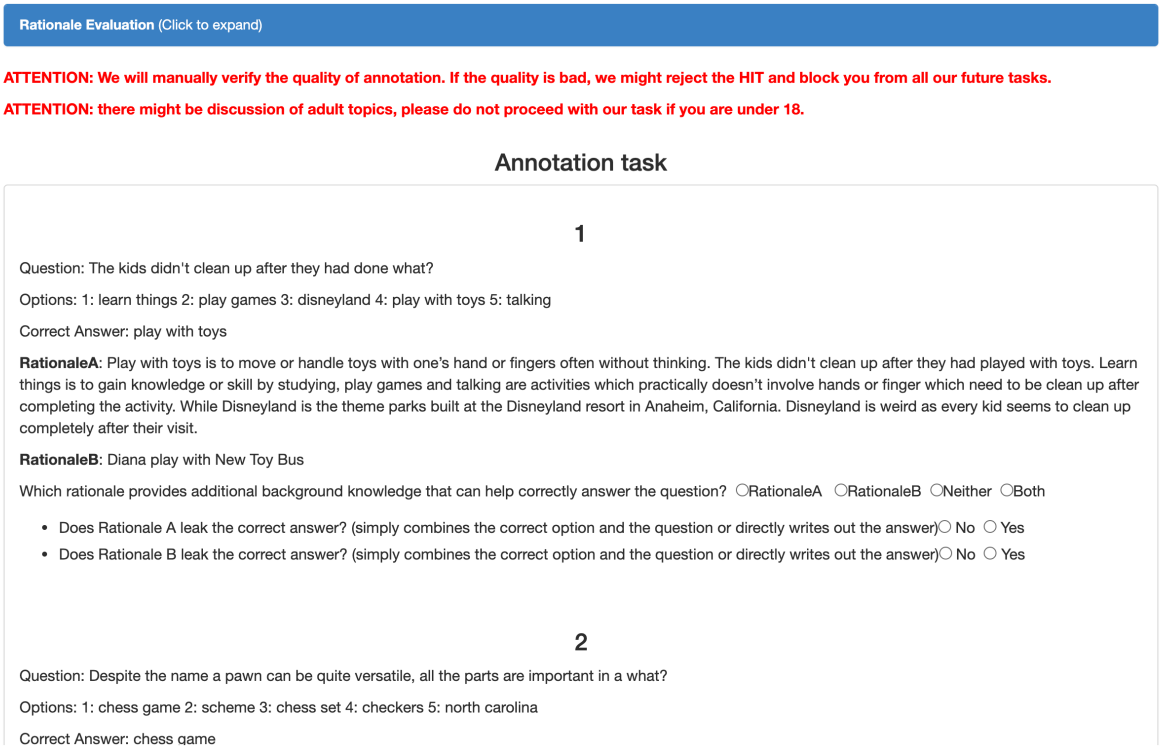

Figure 3: Part of the annotation interface.