Measuring the Human Utility of Free-Text Rationales in Human-AI Collaboration

Brihi Joshi[♪*] **Ziyi Liu**[♪*] **Zhewei Tong**[♫] **Aaron Chan**[♪] **Xiang Ren**[♪]
[♪]University of Southern California [♫]Tsinghua University
{brihijos, zliu2803, chanaaro, xiangren}@usc.edu[♪] tzw19@mails.tsinghua.edu.cn[♫]

Abstract

Recently, there has been growing interest in using language models (LMs) for human-AI collaboration. To explain their reasoning processes to humans, state-of-the-art LMs have been shown to fluently generate free-text rationales (FTRs) in natural language, *e.g.,* via chain-of-thought prompting. Still, it remains unclear how effectively these generated FTRs can provide *human utility* for human-AI collaboration, *i.e.,* assist humans in solving NLP tasks. To investigate what makes an FTR useful to humans, this paper analyzes the relationships between human utility and various LM/FTR properties. First, although LMs are often finetuned/prompted to jointly generate task labels and FTRs, we find that LMs' task performance has little correlation with human utility, whereas LM size is a positive predictor of human utility. Second, we observe that certain FTR property pairs are strong positive predictors of human utility, *e.g.,* high-utility FTRs tend to both be concise and contain novel information. Third, we show that high-utility FTRs for a given task instance can provide transferable knowledge that helps humans generalize to solving new instances. By shedding light on the nature of FTRs' human utility in practical settings, our findings can help guide future work on designing LMs and FTR generation strategies for stronger human-AI collaboration.

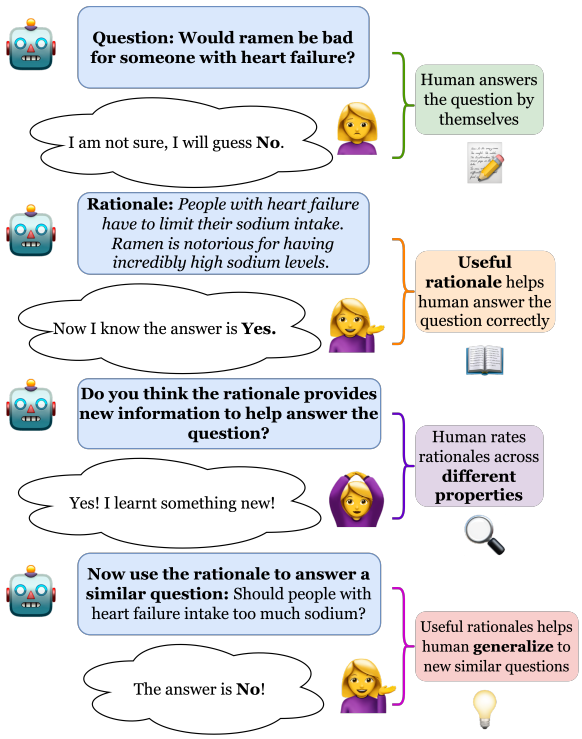

Figure 1: **An illustration of our pipeline to evaluate human utility of free-text rationales:** Humans answer a question without a rationale. An LM then shows them a rationale, that helps them correctly answer the same question, after which they rate the rationale across different axes like *novelty* (shown here). The same rationale also helps them answer a new question which uses a similar reasoning process as the original question.

1 Introduction

In recent years, there has been a surge of interest in using language models (LMs) for human-AI collaboration (Wiegreffe et al., 2022; You and Lowd, 2022; Liu et al., 2022). For example, LMs have played a large role in helping researchers more efficiently construct diverse, high-quality text datasets (Yuan et al., 2021; Liu et al., 2022), helping teachers more efficiently design appropriate reading comprehension exam questions (Yao et al., 2022), and helping social media managers more efficiently identify explicit posts/comments on their platforms (Lai et al., 2022).

However, LMs' opaque reasoning processes pose serious concerns about LMs' role in high-stakes decision-making (Bender et al., 2021; Doshi-Velez and Kim, 2017; Lipton, 2018). To improve LMs' ability to communicate with humans, many works have explored using LMs to generate *free-text rationales* (FTRs). Whereas extractive rationales explain LMs' decisions by highlighting im-

[*]Equal contribution.

portant input tokens (Sundararajan et al., 2017; Li et al., 2016; Chan et al., 2022b), FTRs can fluently provide human-like explanations via natural language (Ehsan et al., 2018; Narang et al., 2020; Rajani et al., 2019a; Camburu et al., 2018; Wei et al., 2022; Majumder et al., 2021). Plus, FTRs can reference things beyond the task input as well as support high flexibility in content, style, and length (Narang et al., 2020; Rajani et al., 2019a; Wiegreffe et al., 2022, 2021; Chan et al., 2022a).

Rationale quality has been studied with respect to properties like faithfulness , plausibility , and LM task performance . Yet, despite the promise of human-AI collaboration, there has been little work on evaluating rationales' *human utility*, defined as the extent to which rationales improve humans' performance on a given task. Moreover, the only prior works exploring rationales' human utility have only considered extractive rationales (Chen et al., 2022; Idahl et al., 2021), whose explanatory scope is restricted to input token scoring. To investigate the human utility of FTRs, we study three research questions, each analyzing the relationship between human utility and a different category of LM/FTR properties.

First, *what is the correlation between an FTR's human utility and its corresponding LM's task performance?* (§3) LMs are often finetuned/prompted to jointly generate task labels and FTRs. Nonetheless, across a wide range of LM architectures and prompt templates, our human-subject studies show little correlation between an FTR's human utility and the task performance of the LM that generated the FTR. Furthermore, we find that larger LMs (*e.g.,* GPT-3) tend to generate FTRs with higher human utility, while LM finetuning is a more important factor in determining the LM's task performance.

Second, *which FTR properties are most predictive of human utility?* (§4) We ask humans to evaluate both LM-generated and gold FTRs with respect to eight FTR properties: grammaticality, validity, coherence, conciseness, leakage, novelty, association, and contrast. Our results suggest that individual FTR properties are not predictive of human utility. On the other hand, we observe that the presence of certain property pairs (*e.g.,* grammaticality+leakage, conciseness+novelty) in a given FTR can be a strong positive predictor of the FTR's human utility.

Third, *to what extent do FTRs for a given task instance help humans generalize to new instances?* (§5) We create new instances (*e.g.,* questions) by either paraphrasing the original instance in a non-trivial manner (rephrase), editing the original instance so that its correct label is changed (counterfactual), or writing an instance that requires a similar reasoning process as the original instance (similar reasoning). Our human-subject studies verify that high-utility FTRs are not limited to explaining their original instances. On the contrary, high-utility FTRs can effectively provide humans general knowledge for solving new instances.

This paper presents the first comprehensive study of why humans perceive some FTRs to be more useful than others. By analyzing the relationships between FTRs' human utility and various LM/FTR properties, we establish a better understanding of how LMs and FTR generation strategies can be designed to yield higher human utility. We believe our findings can help guide future work on developing methods for efficient and reliable human-AI collaboration.

2 Analysis Setup

What is Human Utility? We define human utility of rationales as *the ability of a human to correctly solve a task with a rationale, that they are otherwise unable to* (Idahl et al., 2021; Chu et al., 2020).

More formally, let $\mathcal{F}$ be a *self-rationalizing model* (Wiegreffe et al., 2020) that can generate rationales with its predictions, and a corresponding input-output pair x, y. $\mathcal{F}$ takes in x as an input and generates a prediction y_p, and a rationale that corresponds to this prediction r_p. The task accuracy (which corresponds to accuracy, when aggregated over all the instances) of this instance is given as follows:

$$\textsc{Task Accuracy} = \begin{cases} 1 & y_p = y \\ 0 & otherwise \end{cases}$$

Let $\mathcal{H}$ be a human predictor that first takes in the instance x and predicts a label for that instance, y_h. Then, $\mathcal{H}$ is also shown the rationale r_p and now takes both the instance and rationale x, r_p as an input, and predicts a label y_{hr}.

Therefore, human utility of the rationale r_p is calculated as:

$$\textsc{Human Utility} = \begin{cases} \textsc{Useful} & y_h \neq y \; and \; y_{hr} = y \\ \textsc{Not Useful} & y_{hr} \neq y \\ \textsc{Unsure} & y_h = y \; and \; y_{hr} = y \end{cases}$$

Method Type	Template	Input	Output
In-Context Learning	Chain-of-Thought	*Q:* Demonstration Question 1 *A:* Demonstration Rationale 1. The Predicted Answer is Demonstration Answer 1. .... (repeated based on # of demonstrations) *Q:* Input Question *A:*	Generated Rationale. The answer is Predicted Answer.
	FEB	*Answer the Input Question from the provided choices, and provide a reason why the Predicted Answer is correct.* *Question:* Demonstration Question 1 *Choices:* Yes or No *Answer:* Demonstration Answer 1 *Reason:* Demonstration Rationale 1 .... (repeated based on # of demonstrations) *Question:* Input Question *Choices:* Yes or No *Answer:*	Predicted Answer. *Reason:* Generated Rationale
Fine-tuning	SQuAD-T5	*explain strategyqa Input Question:* Input Question *context:* True, False	Predicted Answer because Generated Rationale.
	Infilling	*explain strategyqa Input Question:* Input Question *choice:* True, False <extra_id_0> because <extra_id_1>	<extra_id_0> Predicted Answer <extra_id_1> Generated Rationale <extra_id_2>
	T5-Like	*explain strategyqa query:* Input Question *entities:* True, False	Predicted Answer because Generated Rationale.
	QA-simple	*explain* Input Question *A) True B) False*	Predicted Answer because Generated Rationale.

Figure 2: **Prompt templates for generating rationales:** Shown here are inputs and outputs of different template variations. Chain-of-Thought templates are taken from publicly released versions by Wei et al. (2022), whereas FEB and Fine-tuning templates are taken from Marasovic et al. (2022).

In other words, rationales are *useful* if a human incorrectly solved the task before, and with the introduction of the rationale, is able to correct their answer. If even after being shown the rationale, the human is still solving the task incorrectly, this implies that the rationale has *not* been useful. However, if the human was correct both before and after being shown the rationale, we cannot conclusively determine the role of the rationale in helping solve the task. We term these rationales are *Unsure*. These category of instances can either be too easy, or it can be the case that the human was already aware of the answer even before being shown the rationale. Of course, this can also imply that the rationale has still been useful in answering the task correctly, however, our definition of utility specifically evaluates cases where rationales are solely responsible for human utility.

Task and Dataset Selection. We refrain from tasks used in existing free-text rationale works (Wiegreffe and Marasović, 2021) like NLI (Camburu et al., 2018) and Commonsense QA (Aggarwal et al., 2021). A primary reason for this is that humans are already able to reason better than models for NLI and Commonsense QA (Nangia and Bowman, 2019; Talmor et al., 2021). Therefore, the objective of machine rationales in this case is just to establish trust or generate faithful rationales. We aim to study rationale utility specifically in cases where the rationales can help with knowledge transfer that helps humans to correctly solve a task. We thus impose the following constraints in our task and dataset selection:

- **Added advantage:** Tasks where machines can provide added advantage and that are not trivial or obvious for humans to solve.
- **Objectivity:** Tasks where the reasoning has a limited scope of subjectivity.
- **Dataset size (of rationale annotations)**: Size of gold rationales is considerably larger in the dataset, so as to provide room for training LMs with those rationales.

In this work, we choose the StrategyQA dataset (Geva et al., 2021), which is an open-domain binary QA benchmark, where questions require implicit reasoning steps to be answered. The StrategyQA dataset consists of an input question, the answer, along with intermediate implicit reasoning steps that are used to answer the questions. The implicit reasoning steps were generated by decomposing the original question into multiple questions. For our project, we combine these implicit reasoning steps and use them as rationales for a given instance.

$\mathcal{F}$	Model	Size	Finetuning setting	Accuracy			
				SQuAD-T5	Infilling	QA-simple	T5-like
Without Rationale	T5	large	full	64.41	62.45	61.35	62.45
		3B	48-shot	55.46 ± 3.47	53.35 ± 2.95	50.95 ± 3.85	52.84 ± 4.51
		3B	128-shot	60.48 ± 0.87	60.11 ± 2.21	52.47 ± 2.21	61.50 ± 2.55
	UnifiedQA	large	full	63.76	61.57	67.90	68.34
		3B	48-shot	54.80 ± 3.64	55.46 ± 4.36	55.97 ± 3.01	55.24 ± 4.22
		3B	128-shot	60.05 ± 3.08	58.22 ± 0.55	61.50 ± 0.55	59.24 ± 5.18
With Rationale	T5	large	full	61.14	65.50	62.45	60.26
		3B	48-shot	51.97 ± 1.00	53.35 ± 1.33	50.94 ± 2.62	50.87 ± 3.28
		3B	128-shot	52.40 ± 2.19	56.70 ± 1.85	53.93 ± 3.61	53.35 ± 1.40
	UnifiedQA	large	full	64.85	65.72	62.45	62.45
		3B	48-shot	53.49 ± 4.36	60.99 ± 2.56	55.38 ± 5.70	55.09 ± 4.63
		3B	128-shot	58.23 ± 3.07	62.08 ± 0.77	59.97 ± 4.94	57.50 ± 1.28

Table 1: **Self-Rationalising Model Results (Fine-tuning)**: Shown here are test set accuracies of LMs of different sizes, and fine-tuned with different number of training examples, for four different templates. Cells highlighted in blue are highest performing templates for each model configuration and red denotes a configuration selected for the rest of our work.

Self-rationalizing Models. We try variations of in-context learning based approaches (Wei et al., 2022), as well as few-shot and full finetuning approaches (Marasovic et al., 2022) to generate rationales. For in-context learning based approaches, we vary the demonstrations based on the number of demonstrations desired, and the selection strategy for these demonstrations. These demonstrations can either be fixed across all instances vs. randomly picked for each instance, from the training set. Demonstrations that are picked randomly can either be six in number (to match a fixed number of demonstrations as per Wei et al. (2022)), or determined by a maximum token length that is specific beforehand (for our experiments, we use 2048 as the maximum token length of an input). For these settings, we implement two input-output templates – where rationales r_p come after (FEB) (Marasovic et al., 2022) or before the prediction y_{hr} respectively (Chain-of-Thought or CoT) (Wei et al., 2022). The LM used for all in-context learning experiments is GPT-3 (Brown et al., 2020). For fine-tuning approaches, we fine-tune two LMs - T5 (Raffel et al., 2019) and UnifiedQA (Khashabi et al., 2020b), with varying sizes - large and 3B. For each of these two LMs, we use four variations of input-output templates (SQuAD-T5, Infilling, T5-Like and QA-simple), as defined by Marasovic et al. (2022). Examples of each of these templates are provided in Figure 2.

As seen in Tables 2 and 1, for the StrategyQA dataset, FEB templates with randomly selected demonstrations provides the highest accuracy for in-context learning approaches, whereas the infilling template consistently outperforms other input-output templates for fine-tuning approaches. For the rest of our work, we select three best performing LM configurations with varying sizes – (1) GPT-3 (with FEB template, and 6 randomly selected demonstrations), (2) T5-large (with infilling template, fine-tuned on the entire training set) and (3) T5-3B (with infilling template and 128-shot fine-tuning).

$\mathcal{F}$	Template	# of demo	Demo Picked	Accuracy
Without Rationale	CoT	6	Randomly	57.11
		max len	Randomly	53.98
		6	Fixed	56.23
	FEB	6	Randomly	52.84
		max len	Randomly	56.33
		6	Fixed	54.80
With Rationale	CoT	6	Randomly	58.51
		max len	Randomly	55.24
		6	Fixed	58.90
	FEB	6	Randomly	60.04
		max len	Randomly	60.04
		6	Fixed	57.42

Table 2: **Self-Rationalising Model Results (In-Context Learning)**: Shown here are test set accuracies of GPT-3, when it is prompted to predict with/without generating rationales. Cells highlighted in blue are highest performing variations, and red denotes a configuration selected for the rest of our work.

3 Does Task Performance Correlate with Human Utility?

Task Performance. For the three selected best performing LM configurations, we note (Tables 2, 1) that task performance increases after the LM is forced to generate rationales. This is also consistent with prior findings (Wei et al., 2022; Marasovic et al., 2022).

Human Utility Annotations. We conduct human-subject studies to evaluate utility of free-text rationales. For each StrategyQA test instance, we ask humans to first provide an answer given the question. We then show them a rationale and ask them to answer the question again. The rationale shown to them is generated by either of the three selected LMs. We use Amazon Mechanical Turk [1] to first curate a set of annotators that understand the task well (via extensive qualification tests). Each instance is answered by five annotators.

We observe that, StrategyQA instances are difficult to annotate by humans, as a lot of them are fact-based, which the human might or might not know beforehand. Therefore, human agreement before the rationale is shown is low (Krippendorf's $\alpha = 0.18$). However, after being shown the rationale, the agreement increases, with Krippendorf's α being 0.47, 0.30 and 0.24 for GPT-3, T5-3B and T5-Large LMs respectively. Examples of rationales annotated into each of the three human utility categories (useful, not useful, unsure) is in the Appendix (Table 10).

Correlation between Task Performance and Human Utility. For each instance, we calculate human utility as defined in §2, where predictions made by five annotators are aggregated by taking a majority vote. Overall, while including annotations for all models combined, we observe that the correlation between task accuracy (whether a given instance was correctly predicted by the self-rationalizing model) and human utility of a rationale (useful, not useful and unsure) was close to none (Theill's $U = 0.0359$). This indicates that while generating rationales might improve overall task performance, there is no guarantee that these rationales useful for humans in solving the task correctly.

In fact, if we look at the correlations for each LM separately, we observe Theill's U for GPT-3, T5-3B and T5-Large were 0.111, 0.034 and 0.005 respectively. This also demonstrates that even though T5-Large, which was fine-tuned on the entire training set had the highest task performance, it has the lowest correlation with human utility.

[1] www.mturk.com

4 What are the Properties of Rationales that are Useful?

As we have observed in §3, higher task performance does not necessarily correlate with the rationales being useful for humans in solving the task. We follow up this discussion by investigating more granular-level properties that associate with human utility of rationales.

Granular-level Rationale Properties. What are some granular-level properties of rationales that are useful? Can such qualities help distinguish between rationales that are useful, to those that are not useful or ones we are unsure about? To answer the above questions, we follow list a set of desirable properties of that useful rationales should satisfy (Wiegreffe et al., 2021, 2022; Golovneva et al., 2022). These properties evaluate rationales along four axes - surface form qualities, support towards predicted labels, informativeness and style. Surface form qualities test whether a rationale is *grammatical* and *factually valid*. *Association* with label and *contrast* between different labels measure the extent to which rationales support the labels that were generated with them. We also evaluate the informativeness of a rationale, which is determined by *novel information* that the rationale provides over the question, along with asking whether it directly *leaks the answer*. Lastly, we also check whether the rationale contains *irrelevant hallucinations* or relevant but *redundant information*. Descriptions and examples of these properties are shown in detail in Figure 3.

Human Annotations. For rationales generated by all three LMs, as well as gold rationales, we conduct human studies to evaluate whether the rationales satisfy the given properties. For each instance, a property is marked on a binary scale (Yes / No), indicating the presence or absence of that property and evaluated by five annotators. Each category of properties is evaluated on a separate HIT, for which instructions have been modified so as to ensure that the annotators understand our definitions of the properties. Given the complex nature of the human study, we make sure that the prop-

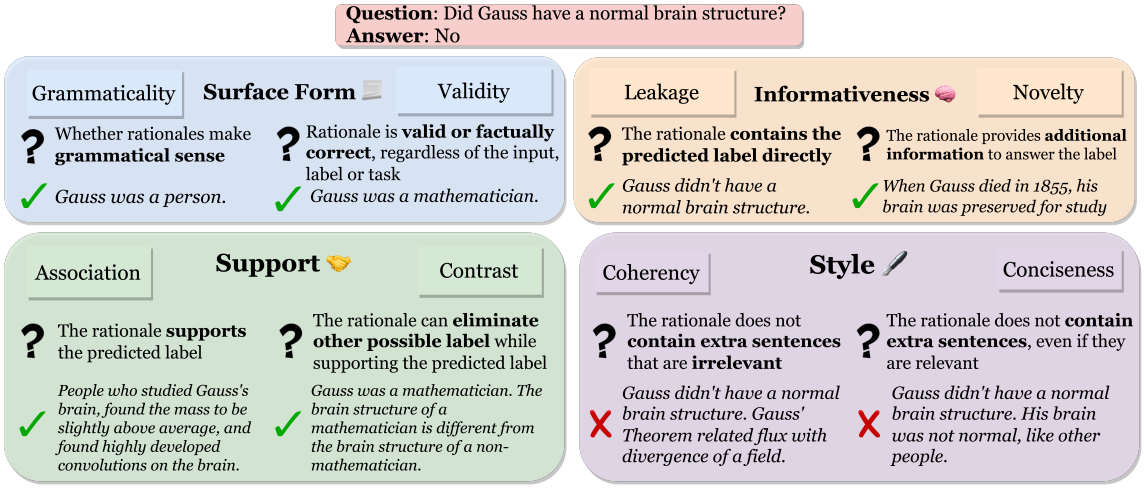

Figure 3: **Granular-level Rationale Properties:** Definitions for properties along each axes (surface form, informativeness, support and style) are shown. For all but style axes, an example of a rationale *satisfying* the property is also shown. For style, we show examples of rationales that do not satisfy the given properties.

erty annotations reach low to moderate agreement across all annotators (Appendix 8).

Property Analysis. We first study the presence of these properties in rationales, without considering the utility of these rationales. Figure 4 plots the distribution of these properties, split by the models that generate these rationales, along with Gold rationales. The distributions are obtained by taking the mean of ratings from five annotators for a given instance, where a higher value indicates a more frequent presence of that particular property in the set of rationales. We observe that Gold rationales, in comparison to other model-generated rationales, have lower scores for leakage and higher scores for other properties. In fact, Gold rationales are always associated with the gold label, which serves as a sanity check, as they are designed to help answer the gold label. While all types of rationales are mostly grammatically correct , T5-large and T5-3B suffer at producing rationales that are factually correct, and T5-large T5-large rationales also tend to hallucinate and produce redundant sentences in rationales more often. While GPT-3 rationales tend be generally better than T5-large and T5-3B for surface-form and stylistic properties, they leak the predicted label more often than them. There is high variation for rationale-label association and contrasting features in rationales for all model-generated rationales, however on average, GPT-3 generated rationales are better on these metrics too.

Properties of Useful Rationales. We use a Generalized Linear Mixed-Effects Model (GLMEM) (similar to Lamm et al. (2020)) to model the correlation of different properties and their interactions with that of high utility. The formula used for modelling the GLMEM is as follows: $\textsc{Response} = (\textsc{Grammaticality} + \textsc{Validity} + \textsc{Coherence} + \textsc{Conciseness} + \textsc{Leakage} + \textsc{Novelty} + \textsc{Association} + \textsc{Contrast})^2 + (1|\textsc{Question ID}) + (1|\textsc{Model ID}) + (1|\textsc{Human Prior})$

The response (dependent variable) is human accuracy after the human was shown the rationale. More formally,

$$\textsc{Response} = \begin{cases} 1 & y_{hr} = \hat{y} \\ 0 & y_{hr} \neq \hat{y} \end{cases}$$

All properties, along with their second-order interactions (implemented using the squared term above) are dependent variables. Furthermore, we try to control for random effects whose variability might influence the response. We control for randomness induced by a particular question, the model generating the rationales or whether the human had correctly answered the question before (Human Prior). More formally,

$$\textsc{Human Prior} = \begin{cases} 1 & y_{h} = \hat{y} \\ 0 & y_{h} \neq \hat{y} \end{cases}$$

Table 3 shows the log odds of a rationale being useful, when a certain property is present or absent, while averaging over other properties. We note that all of the log odds are negative, which means that in

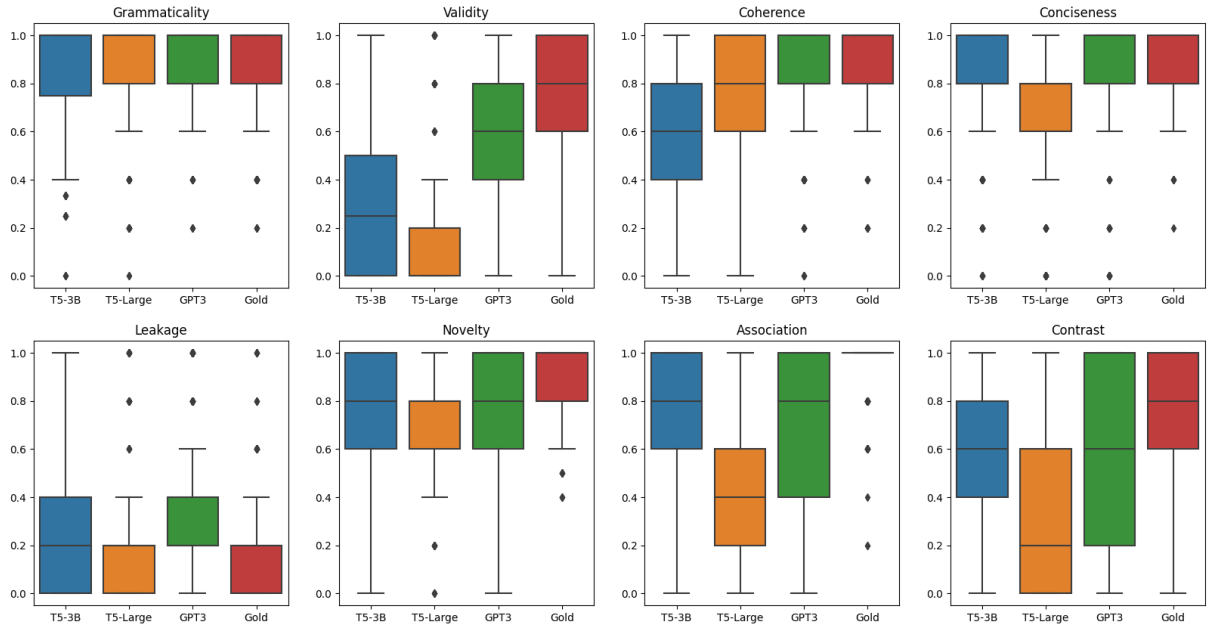

Figure 4: **Distribution of Property Annotations for Different Rationales:** Distribution is generated by aggregating scores of five annotators of each instance. A higher value implies more presence of the property in the rationale generated by the particular LM.

Property	Present	Absent
Grammaticality	-0.568	-0.686
Validity	-0.554	-0.700
Coherence	-0.665	-0.589
Conciseness	-0.540	-0.714
Leakage	-0.616	-0.638
Novelty	-0.712	-0.542
Association	-0.632	-0.622
Contrast	-0.613	-0.641

Table 3: **Influence of individual properties in human utility:** Log odds of a rationale being useful, when a certain property is present or absent.

Parameter	Coefficient (SD)
(Intercept)	-0.724 (0.72)
+ grammaticality + leakage	0.226 (0.55)
+ conciseness + novelty	0.169 (0.32)
+ grammaticality + novelty	0.149 (0.50)
+ coherence + novelty	0.138 (0.23)
+ novelty + contrast	0.136 (0.27)
+ conciseness + contrast	0.119 (0.37)
+ validity + leakage	0.118 (0.19)
+ association + contrast	0.112 (0.54)
+ leakage + contrast	0.098 (0.29)
+ coherence + association	0.095 (0.27)

Table 4: **Pairwise property interactions for rationale utility**: Given an intercept (when a rationale does not satisfy any property), the top ten pairs of properties that lead to an *increase* in the log odds of a rationale being useful from the intercept is shown.

isolation, the presence or absence of any property is does not correlate well with rationales of high utility.

We then look at pairwise interactions. Table 4 shows top ten pairs which lead to an increase in utility log odds from the base level (Intercept), which is when a rationale does not satisfy any property. A grammatically correct rationale that explicitly leaks the answer leads to the highest increase in log odds. This is also intuitive, as leakage is a direct signal to a human to select a given answer, without any reasoning from the human's behalf.

When all possible combinations of properties are considered, presence of all but coherence and association leads to a positive log odds for rationale utility: 0.139. This implies that humans are generally robust to hallucinations that are irrelevant to the question. Furthermore, association of the rationale with its predicted label is also not an important property for rationale utility, as the rationale may not be associated with the correct answer and therefore, mislead the human into making an incorrect choice.

5 Can Rationales with High Utility Help Humans Generalize to New Instances?

As we have defined and shown in previous sections, human utility of rationales is determined by their

Original Question, Gold Rationale and Label	Generalization Question and Label	Generalization Type
Q: Was Iggy Pop named after his father? *R*: Iggy Pop's birth name was James Newell Osterberg Jr. The father of Iggy Pop was James Newell Osterberg Sr. *A:* Yes	*Q*: Was Iggy Pop's name derived from his father? *A*: Yes	Rephrase
Q: Can the Moscow Kremlin fit inside Disney Land? *R*: The Moscow Kremlin is a fortified complex in the middle of Moscow Russia. The Kremlin takes up sixty eight acres. Disney Land is an amusement park in California. Disney Land occupies eighty five acres. *A:* Yes	*Q*: Is the Moscow Kremlin bigger than Disney Land? *A*: No	Counterfactual
Q: Does Julia Roberts lose the prolific acting contest in her family? *R*: As of May 2020, Julia Roberts has acted in 64 projects. Julia Roberts has a brother in acting, Eric Roberts, and a niece in acting, Emma Roberts. As of May 2020, Eric Roberts has acted in 577 projects. *A:* Yes	*Q*: Does Julia Roberts have more acting projects than her brother? *A*: No	Counterfactual
Q: Does Snoop Dogg advocate a straight edge lifestyle? *R*: A straight edge lifestyle requires abstaining from the usage of recreational drugs or alcohol. Snoop Dogg is famous for his chronic usage of marijuana. *A:* No	*Q*: Does Snoop Dogg advocate the use of recreational drugs or alcohol? *A*: Yes	Counterfactual
Q: Can vitamin C rich fruits be bad for health? *R*: Oranges are fruits that are rich in vitamin C. Oranges are very acidic fruits that can wear down tooth enamel. Too much Vitamin C can cause nausea and diarrhea. *A:* Yes	*Q*: Can oranges be bad for health? *A*: Yes	Similar Reasoning
Q: Is the Matrix a standalone movie? *R*: The Matrix ends in a cliffhanger. The story is then resolved in two sequels, making a trilogy. There are also supplemental works adding to the story, such as a video game and the Animatrix. *A:* No	*Q*: Is the Matrix a trilogy? *A*: Yes	Similar Reasoning
Q: Does water have viscosity? *R*: Viscosity is resistance of fluid to deformation. Water is not resistant to deformation. *A:* No	*Q*: Is water resistant to deformation? *A*: No	Similar Reasoning

Table 5: **Examples of generalization questions of each type**: We show the original question, rationale and label triplet, along with GPT-3 generated generalization questions and gold label for the generated question.

ability to guide humans to correctly solve tasks. We follow this up by investigating if humans can generalize to syntactic or semantic perturbations of the original question, while being shown rationales of the original question. This will help us understand if human utility of rationales can also indicate whether rationales help with knowledge transfer for unseen instances.

Types of Generalization Questions. For our study, we consider three distinct types of generalization setups. Firstly, we evaluate the human $\mathcal{H}$'s ability to generalize to non-trivial **rephrases** of the original question. We avoid simple rephrases like changing a preposition, or removing an adverb so as to avoid near duplicates of the original question. Next, we look at **counterfactual** questions. These questions follow the same reasoning steps of the original question, however, they flip the answer of the original question. Lastly, we test $\mathcal{H}$'s ability to understand questions that follow a **similar reasoning** process as the original question, but are not related to the original question. These questions can entail entity swaps, or questions that uses one of the reasoning steps to answer the original question. Examples of each type of generalization question is shown in Table 5.

Generating Generalization Questions. For generating generalization questions as describe above, we follow the Human and AI collaboration paradigm as introduced by Liu et al. (2022). We first start by manually creating templates with instructions for each type of generalization question. We then select six demonstrations for these templates. The selected instructions and demonstrations are in Appendix (Table 9). These demonstrations are fixed for each type (however, may differ across the different types) and are selected from the training set. For every test instance, we insert

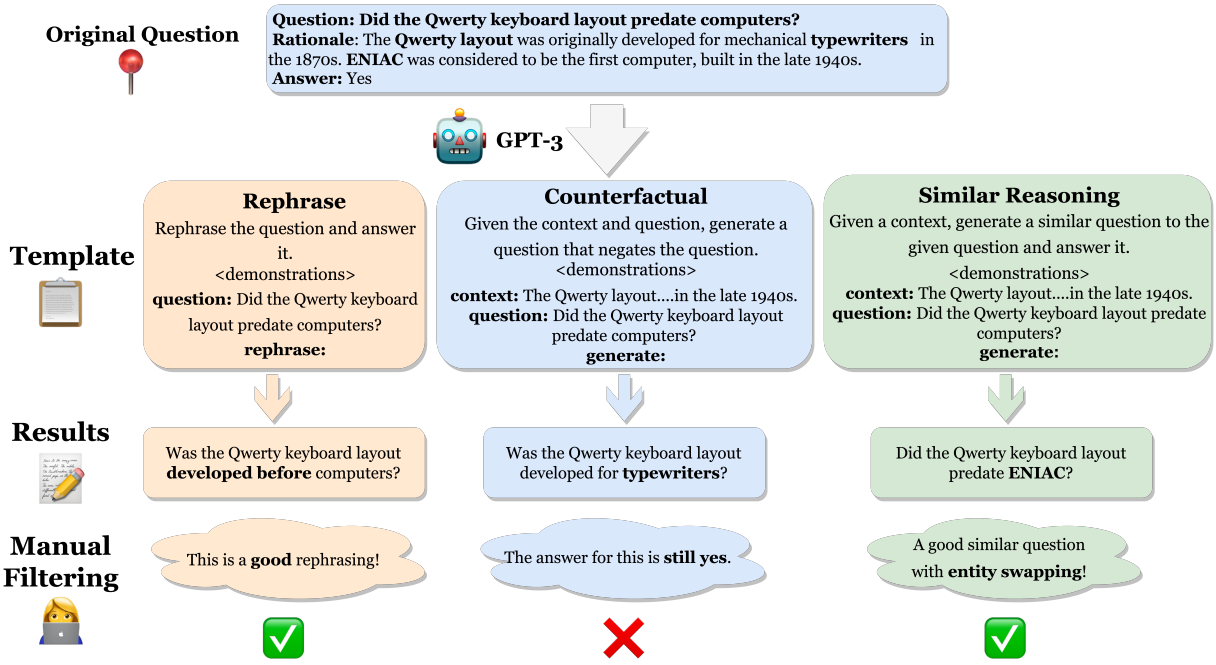

Figure 5: **Generating Generalization Questions:** A question, rationale, answer triplet is fed to GPT-3 using given templates. GPT-3 generates potential generalization question candidates from each template, which are then manually filtered and answered to create gold labels in-house.

it at the end of the corresponding template, which is then used as a prompt for GPT-3 to generate questions. We then manually filter the generated questions in-house and provide gold answers for them, to make sure that the final set of questions are of good quality. Finally, we have a set of 123 rephrase, 102 counterfactual and 171 similar reasoning generalization questions that are of good quality from the test set. The pipeline and examples of the templates are shown in Figure 5.

Human Annotations. Similar to §3, we first ask the annotators to answer a given question with and without the rationale. We then show them a generalization question, which they have to answer while referencing the rationale for the original question. We repeat the experiment above with rationales from the three LMs, along with gold rationales. Each instance is annotated by five annotators. Given that there are no corresponding rationales for the generalization questions, this annotation setup would measure the impact of rationales of the original question towards answering the generalization questions.

Results. Figure 6 shows generalization accuracy of rationales with different human utility, split across different LM's generating them as well as gold rationales. We observe that gold rationales form an upper bound for generalization accuracy, across all types of generalization questions and types of rationale utility. Useful rationales are always significantly better than non-useful rationales while answering generalization questions, which indicates that rationale utility is a strong signal for generalizing to newer instances. Paraphrases of original questions are relatively easier to answer, when compared to counterfactuals and similar reasoning, as it can be seen by both Gold rationale, as well as accuracy of all models combined. We can also note that GPT-3 generated rationales help generalize better to more difficult scenarios like counterfactuals or similar reasoning questions. Rationales that we are unsure of in terms of utility are helpful in generalizing to paraphrases or similar reasoning questions. This indicates the presence of easy-to-answer questions in that bucket of instances, which also hints towards the fact that their paraphrases or similar reasoning counterparts would also be easier to answer. Counterfactuals however show a different trend, where unsure rationales have a lower generalization accuracy when compared to useful rationales. Examples of generalization questions which were answered correctly/incorrectly for rationales that have high or low human utility is shown in the Appendix (Table 11).

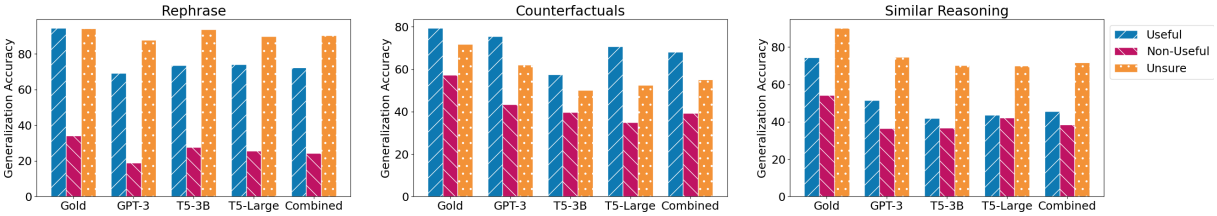

Figure 6: **Generalization Performance:** Shown here is generalization accuracy for different generalization question types, split by LM-generated and gold rationales. The *Combined* group is aggregated across all LM-generated rationale performance.

6 Conclusion and Future Work

In this work, we explore the human utility of LM-generated free-text rationales in human-AI collaborative settings, *i.e.,* , guiding humans to solve a task correctly. We observe that, while generating rationales can help improve downstream task performance for LMs, there is no correlation between task performance and human utility. We further note that larger LMs like GPT-3 often generate more useful rationales, even if their task performance is not optimal like that of a fine-tuned T5-large LM. We also identify properties that are more closely correlated with useful rationales – like conciseness and novelty. Lastly, we observe that high-utility rationales are also good sources of knowledge to help humans generalize to new instances.

The purpose of investigating human utility is to better situate rationales in real-world settings where human decision making can be guided by LMs. While we provide an initial exploration of this paradigm, there are several challenging directions that follow this study. Human utility can be used as a prior to guide LMs to not only generate rationales that improve LMs' task performance but also control for human utility of these rationales. However, evaluating human utility requires an expensive human annotation study. Therefore, another interesting direction is to be able to effectively capture human utility in a reliable automated metric, that can be used directly in LMs. Lastly, our evaluation of utility is constrained to our selected dataset and task. A more thorough investigation of the human utility of the rationales generated for more high-stakes scenarios is also necessary.

7 Related Work

Free-text Rationale Human Utility Outside of the NLP community, extractive explanations have been used to improve human's understanding of the model (Wang and Yin, 2021; Feng and Boyd-Graber, 2018) or detecting errors in model predictions (González et al., 2021). Although prior motivation of generating free-text rationales has been primarily to improve task model performance (Rajani et al., 2019b; Zelikman et al., 2022; Wei et al., 2022; Lampinen et al., 2022), recent works have evaluated human utility of free-text rationales in various ways. Wiegreffe et al. (2022) use human acceptability judgements on over-generated rationales by GPT3 (Brown et al., 2020) to train a downstream model for generating good quality rationales. They also evaluate the rationales across seven axes like grammar, factuality, *etc*. Sun et al. (2022) compared human written rationales with those generated by GPT3 across two axes: rationales that provide new information over the input, and those that leak the label directly. While prior work (Chu et al., 2020) shows that human accuracy is unaffected by showing visual explanations on tasks which are difficult for both models and humans (like age prediction from images), they use extractive explanations like input saliency. Carton et al. (2020) report similar findings, where input attributions do not have a significant impact on human accuracy, but it helps reduce the cognitive burden in understanding the task for humans.

Rationale Generation There are two distinct methods of generating free-text rationales. The first way is to fine-tune an encoder-decoder like model, for example, T5 or it's variations like UnifiedQA (Raffel et al., 2020; Khashabi et al., 2022, 2020a). Finetuning T5 to generate rationales (Narang et al., 2020; Paranjape et al., 2021) entails appending a tag like `explain:` in the input text, to nudge the LM to generate rationales during prediction. The generated text can either be separated by structured tags like `answer:, explanation:`, or it can be unstructured, with the answer followed by a `because` keyword, followed by the rationale. Recent methods have also analysed few-shot prompting of T5

with different input-output templates (Marasovic et al., 2022). Another recent approach of generating free-text rationales is via in-context learning (Wei et al., 2022; Kojima et al., 2022; Marasovic et al., 2022; Wiegreffe et al., 2022). A decoder-only model like GPT-3 or its variants (Brown et al., 2020; Wang and Komatsuzaki, 2021) that are pre-trained on a larger corpora of world-knowledge are prompted with demonstrations (Wei et al., 2022), wherein each example contains its corresponding explanation.

Human Utility of Human Rationales Several works in Psychology and Cognitive Science detail the role that human rationales play for human understanding. These studies have shown that human rationales are inherently incomplete and do not capture the complete deductive reasoning process. (Tan, 2021). These rationales are used to either provide *evidence* or *procedure* behind obtaining a given conclusion for a situation (Lombrozo, 2006). Furthermore, some works have also detailed the utility human rationales have for human understanding. Human rationales have shown to help better generalise to unknown circumstances (Lombrozo and Gwynne, 2014), justify decision-making (Patterson et al., 2015), understand relationships between different world entities (Hummel et al., 2014), diagnose when something went or might go wrong, as well as explain one off events that are bizarre (Keil, 2006).

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.

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*, pages 610–623.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-based explanations don't help people detect misclassifications of online toxicity. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):95–106.

Aaron Chan, Shaoliang Nie, Liang Tan, Xiaochang Peng, Hamed Firooz, Maziar Sanjabi, and Xiang Ren. 2022a. Frame: Evaluating rationale-label consistency metrics for free-text rationales. *arXiv preprint arXiv:2207.00779*.

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. 2022b. Unirex: A unified learning framework for language model rationale extraction. In *International Conference on Machine Learning*, pages 2867–2889. PMLR.

Valerie Chen, Nari Johnson, Nicholay Topin, Gregory Plumb, and Ameet Talwalkar. 2022. Use-case-grounded simulations for explanation evaluation.

Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? a case study in model-in-the-loop prediction.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2018. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 81–87, New York, NY, USA. Association for Computing Machinery.

Shi Feng and Jordan Boyd-Graber. 2018. What can ai do for me: Evaluating machine learning interpretations in cooperative play.

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

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2022. Roscoe: A suite of metrics for scoring step-by-step reasoning.

Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. 2021. Do explanations help users detect errors in open-domain QA? an evaluation of spoken vs. visual explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1103–1116, Online. Association for Computational Linguistics.

John E. Hummel, John Licato, and Selmer Bringsjord. 2014. Analogy, explanation, and proof. *Frontiers in Human Neuroscience*, 8.

Maximilian Idahl, Lijun Lyu, Ujwal Gadiraju, and Avishek Anand. 2021. Towards benchmarking the utility of explanations for model debugging. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 68–73, Online. Association for Computational Linguistics.

Frank C Keil. 2006. Explanation and understanding. *Annu. Rev. Psychol.*, 57(1):227–254.

D. Khashabi, S. Min, T. Khot, A. Sabhwaral, O. Tafjord, P. Clark, and H. Hajishirzi. 2020a. Unifiedqa: Crossing format boundaries with a single qa system.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020b. 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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

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

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

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

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.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends Cogn. Sci.*, 10(10):464–470.

Tania Lombrozo and Nicholas Z. Gwynne. 2014. Explanation and inference: mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, 8.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. 2021. Knowledge-grounded self-rationalization via extractive and natural language explanations.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

Nikita Nangia and Samuel R. Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the glue benchmark.

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

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.

Richard Patterson, Joachim T. Operskalski, and Aron K. Barbey. 2015. Motivated explanation. *Frontiers in Human Neuroscience*, 9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

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

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019a. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019b. 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.

Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. 2022. Investigating the benefits of free-form rationales.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Chenhao Tan. 2021. On the diversity and limits of human explanations.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 318–328, New York, NY, USA. Association for Computing Machinery.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models.

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, Ana Marasović, and Noah A Smith. 2020. Measuring association between labels and free-text rationales. *arXiv preprint arXiv:2010.12762*.

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.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 2022. It is AI's turn to ask humans a question: Question-answer pair generation for children's story books. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–744, Dublin, Ireland. Association for Computational Linguistics.

Wencong You and Daniel Lowd. 2022. Towards stronger adversarial baselines through human-AI collaboration. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 11–21, Dublin, Ireland. Association for Computational Linguistics.

Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. 2021. Synthbio: A case study in human-ai collaborative curation of text datasets.

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

A Appendix

A.1 Self-Rationalising Models Training Details

In the experiements, we mainly used 3 models: T5-3B, T5-Large, and GPT3, the model details and hyperparameters are shown in Table 6. For T5-large, we used full train set to train and for t5-3b, we have 2 settings: 48-shot and 128-shot. We used 3 seeds for generating shots for t5-3b. For GPT3, we only used the OpenAI GPT3 api to do inference.

A.2 MTurk Details

In this section, we demonstrate the MTurk experient setup. The details of MTurk experiemts including how many turkers took the evaluation and average time used on finishing evaluations are shown in Table 7. Each MTurk annotator is paid mini-mum wage. Figure 7, 8 and 9 demonstrate the setup for human utility evaluation. Figure 10,11,12 and 13 demonstrate the setup for property evaluation. Figure 14, 15 and 16 demonstrate the setup for generalization questions evaluation.

A.3 Property Annotation Agreements

Config	Assignment
models	**T5-3b** Number of parameters: 3 billion
	T5-large Number of parameters: 770 million
	GPT3(davinci-instruct-beta) Number of parameters: 175 billion
train batch size	4
eval batch size	4
seed	0
max epochs	25
learning rate	3e-5
learning scheduler	fixed

Table 6: **Self-Rationalising Models Training Details**: Here we show the models we used and hyperparameters we used for T5-3b and T5-large model training.

Tasks	Number of Turkers	Average Time(s)
Human Utility Evaluation	80	37.41
Property Evaluation	137	36.50
Generalization Question	25	35.93

Table 7: **Details of MTurk:** Shown here are number of unique Turkers (annotators) and average time of solving one HIT for each task

A.4 Examples

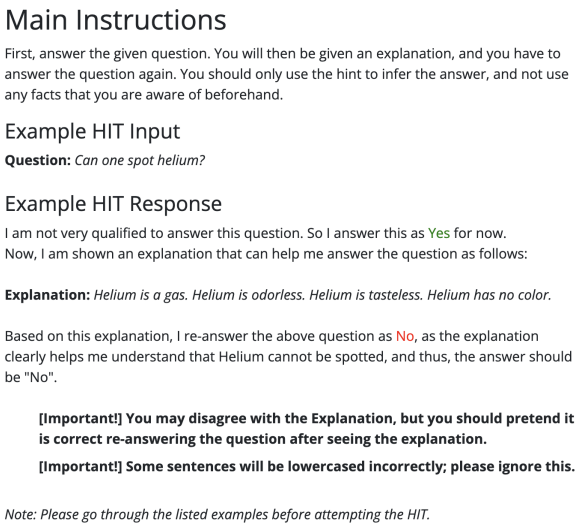

Figure 7: **Instructions for human utility evaluation:** We first show annotators the description of the task and one example of HIT. We also included important notices to make sure annotators will use explanations.

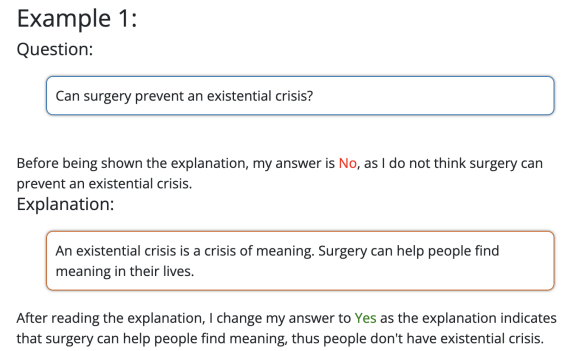

Figure 8: **An example for human utility evaluation:** We then show annotators 5 examples(we only show one of them in this figure). In the example, we will show them the procedure of annotations and how to response.

Rationale	Grammaticality	Validity	Coherence	Conciseness	Leakage	Novelty	Association	Contrast	Average
Gold	0.11	0.18	0.19	0.10	0.24	0.21	0.12	0.24	0.17
GPT-3	0.14	0.18	0.14	0.39	0.25	0.12	0.32	0.42	0.25
T5-3B	0.11	0.22	0.18	0.16	0.27	0.19	0.11	0.15	0.17
T5-Large	0.33	0.51	0.22	0.10	0.24	0.13	0.26	0.33	0.27

Table 8: **Annotation Agreements for Property Ratings**: Shown here are annotation agreements (Krippendorf's α) for each property rating, along with aggregated agreements.

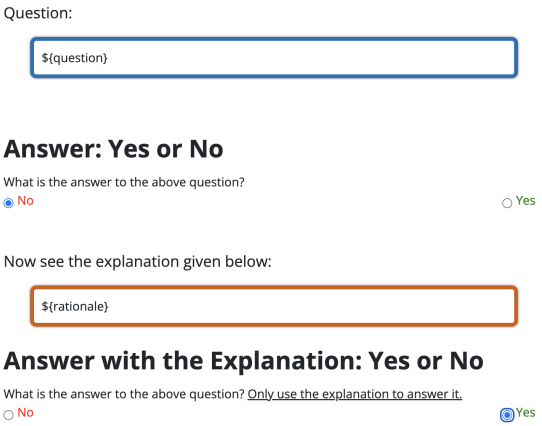

Figure 9: **Questionnaire for human utility evaluation:** Here is the template for evaluation. In the MTurk, the question and rationale will be replaced with real data. We will show the first question in the beginning. When annotators choose yes or no, the explanation and second question will appear.

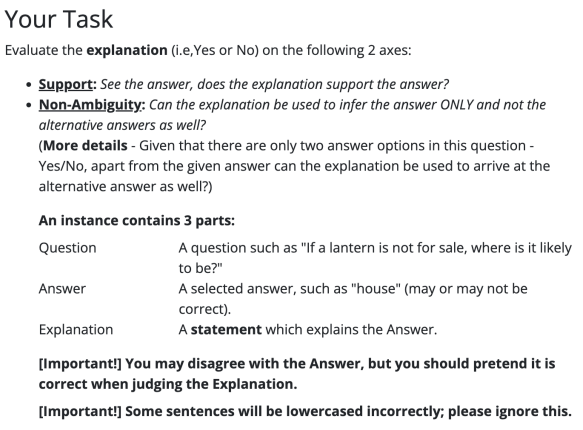

Figure 10: **Instructions for property evaluation:** In this task, we split the property into 4 groups and conduct 4 rounds of annotations.(We show one of the groups - support).We rephrased 'label association' to 'support and 'contrast' to 'non-ambiguity' for easier understanding. In the introduction, we explain the properties and components of instances

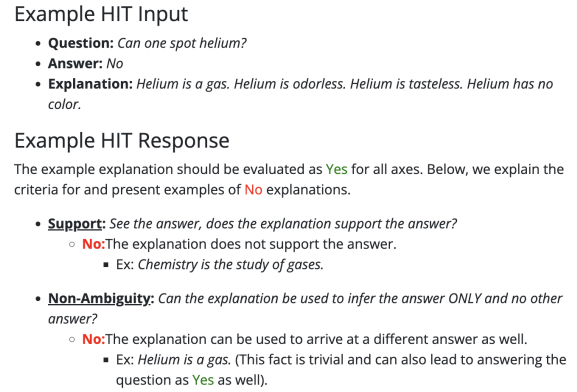

Figure 11: **Instructions for property evaluation:** In the instruction, we also include one HIT example. We explain the properties by showing negative examples.

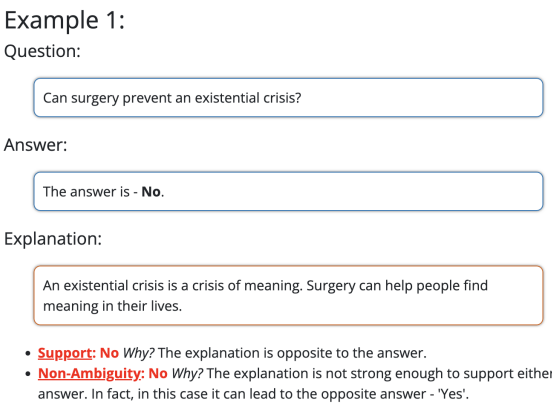

Figure 12: **An example for property evaluation:** We demonstrate 6 examples in the template and we show different combination of results in examples.

Category,Instruction	Demonstrations
Rephrase : Rephrase the question and answer it.	**question**:Are more people today related to Genghis Khan than Julius Caesar? **rephrase**:Do more people today have connection with Genghis Khan than Julius Caesar? **answer**:True.
	question:Would a dog respond to bell before Grey seal? **rephrase**: Would Grey seal respond to bell later than a dog? **answer**:True.
	question:Is a Boeing 737 cost covered by Wonder Woman (2017 film) box office receipts? **rephrase**:Does Wonder Woman box office receipts cover a Boeing 737 cost? **answer**:True.
	question:Is the language used in Saint Vincent and the Grenadines rooted in English? **rephrase**: Does the language used in Saint Vincent and the Grenadines originate from English? **answer**:True.
	question:Are Christmas trees dissimilar to deciduous trees? **rephrase**:Are Christmas trees different from deciduous trees? **answer**:True.
	question:Does Dragon Ball shows and movies fall short of Friday 13th number of projects? **rephrase**:Does Dragon Ball make less shows and movies than Friday 13th? **answer**:True
Counterfactual : Given the context and question, generate a question that negates the question.	**context**:A plum tree is a deciduous tree that bears fruit. Deciduous trees shed their leaves in the autumn. Autumn happens from September until the end of Deember. **question**:Is November a bad time for a photographer to take pictures of a plum tree in bloom? **generate**:Is a plum tree in bloom in the autumn?.
	context:The animals that Yetis are said to look similar to are able to use their hands or toes to grasp items The ability to grasp with hands or other limbs is to be prehensile. **question**:Would a Yeti be likely to have prehensile limbs? **generate**:Is a Yeti able to grasp items with its hands or toes?
	context:Keelhauling was a severe punishment whereby the condemned man was dragged beneath the shipŽ019s keel on a rope. Keelhauling is considered a form of torture. Torture is considered cruel. The Eighth Amendment forbids the use of cruel and unusual punishment **question**:Would keelhauling be a fair punishment under the Eighth Amendment? **generate**:Would keelhauling be considered cruel?
	context:Khanbaliq was the winter capital of the Mongol Empire. Khanbaliq was located at the center of what is now modern day Beijing, China. Moon Jae-In was born in Geoje, South Korea. **question**:Was Moon Jae-in born outside of Khanbaliq? **generate**:Was Moon Jae-in born in Beijing?
	context:Amazonas is mostly tropical jungle. Tropical jungles contain dangerous creatures. Dangerous creatures put people's lives at risk. **question**:Does walking across Amazonas put a person's life at risk? **generate**:Is Amazonas a safe place?
	context:The Los Angeles Memorial Sports Arena had a capacity of 16,740 people. Coachella has had attendance numbers in excess of 99.000 people. Coachella relies on an outdoor set up to accommodate the massive crowds. **question**:Was Los Angeles Memorial Sports Arena hypothetically inadequate for hosting Coachella? **generate**:Would Los Angeles Memorial Sports Arena be too big for Coachella?
Similar reasoning : Given a context, generate a similar question to the given question and answer it	**context**:A plum tree is a deciduous tree that bears fruit. Deciduous trees shed their leaves in the autumn. Autumn happens from September until the end of Deember. **question**:Is November a bad time for a photographer to take pictures of a plum tree in bloom? **generate**:Will the leaves a plum tree fall in the autumn?**answer**:True
	context:The Alamo is located in San Antonio. The Alamo was the site of a major battle during the Texan Revolution against Mexico in 1836. **question**:Was San Antonio the site of a major battle in the 19th century? **generate**:Was the Alamo the site of a major battle in the 19th century?**answer**:True
	context:Filicide is the act of killing a son or a daughter. Marvin Gay Sr. committed filicide in 1984 when he shot his son, singer Marvin Gaye. Isaac's father Abraham, was commanded by God to sacrifice his son Isaac, but was spared by an angel. **question**:Did Isaac's father almost commit similar crime as Marvin Gay Sr? **generate**:Did Isaac's father almost commit filicide?**answer**:True
	context:The animals that Yetis are said to look similar to are able to use their hands or toes to grasp items. The ability to grasp with hands or other limbs is to be prehensile. **question**:Would a Yeti be likely to have prehensile limbs? **generate**:Will a Yeti fail to grasp items with its hands or toes?**answer**:True
	context:Land of Israel was controlled by the Ottoman Empire in 16th century. The religion of Ottoman Empire was Sunni Islam. **question**:Was Land of Israel in possession of an Islamic empire in 16th century? **generate**:Was the Ottoman Empire Islamic once?**answer**:True
	context:Wedding rings are typically made of precious shiny stones such as diamonds. Silicon is a solid rock like element at room temperature that has a natural lustre. Bromine is a liquid at room temperature that is toxic to the touch. **question**:Will silicon wedding rings outsell bromine wedding rings? **generate**:Are silicon wedding rings shiny?**answer**:True

Table 9: **Demonstrations for generating generalization questions**: For each category, we used 6 fixed demonstrations. We used different questions for each category.

Utility	Examples
Useful	**Question**: Could Al Capone have read works from the Harlem Renaissance? **Rationale**: The Harlem Renaissance began in the late 1600's. Al Capone was born in 1888.
	Question: Has Alan Greenspan lived through at least a baker's dozen of president's? **Rationale**: Alan Greenspan was born in 1926. He has lived through the presidencies of Roosevelt, Truman, Eisenhower, Kennedy, Johnson, Nixon, Ford, Carter, Reagan, Bush, Clinton, Bush, Obama, Trump.
Unsure	**Question**: Can a sea turtle play tennis using a tennis racket? **Rationale**: Sea turtles are aquatic animals that cannot swim well under water. Tennis rackets are designed for land animals such as humans.
	Question: Do gorillas fight with panda bears? **Rationale**: Gorillas are native to Africa. Panda bears are native to China.
Not Useful	**Question**: Does 50 Cent get along with Jeffrey Atkins? **Rationale**: Jeff Atkins is a musician who has recorded with 50 Cent. Jeff Atkins has a reputation for being a hard worker and a good friend.
	Question: Do Apollo and Baldur share similar interests? **Rationale**: Apollo is the god of the sun and the arts. Baldur is the god of light and peace.

Table 10: **Examples of rationales for Section 3**:Shown here are examples of useful, unsure, not useful rationale. We included two examples of each type.

Utility	Examples
Useful	**Original Question**:Did Evander Holyfield compete in an Olympics hosted in the western hemisphere? **Rationale**:Evander Holyfield competed in the 1984 Olympics in Los Angeles, California. **Generalization Question**:Did Evander Holyfield compete in an Olympics hosted in the United States?
	Original Question:Is Nine Inch Nails's lead singer associated with David Lynch? **Rationale**:Nine Inch Nails's lead singer is Trent Reznor. Trent Reznor is a producer for David Lynch's film, "Split" (1985). David Lynch produced Split (1985). **Generalization Question**:Is Trent Reznor associated with David Lynch?
Unsure	**Original Question**:Is a beard is moss that grows on a human? **Rationale**:A beard is hair that grows on a human. Moss is a type of plant. **Generalization Question**:Is a beard a type of plant?
	Original Question:Does the Red Sea have biblical significance? **Rationale**:The Red Sea is a body of water in the middle of the desert. The biblical story of Moses crossing the Red Sea is found in Exodus 14:26-27. **Generalization Question**:Is the Red Sea a biblical sea?
Not Useful	**Original Question**:Has a baby ever had a moustache? **Rationale**:Babies are born without facial hair. **Generalization Question**:Has a baby ever had lanugo?
	Original Question:Can Michael Jordan become a professional cook in America? **Rationale**:Michael Jordan was born in 1964 The United States of America was founded in 1776. **Generalization Question**:Can Michael Jordan become a culinary apprentice?

Table 11: **Examples of rationales for Section 5**: For useful and unsure rationales, we selected those that support humans to answer the generalization questions correctly; and for not useful rationales, we selected examples where human failed to give the right answer.

Type of Generalization Questions	Model	Generalization Accuracy		
		Useful	Non-useful	Unsure
Rephrase	Gold	94.68	34.24	94.35
	GPT-3	69.38	18.95	87.90
	T5-3B	73.58	27.82	93.90
	T5-Large	74.11	25.60	90.00
	Combined (Models)	72.31	24.31	90.52
Counterfactuals	Gold	79.50	57.34	71.83
	GPT-3	75.00	43.47	62.11
	T5-3B	57.57	39.72	50.22
	T5-Large	70.66	35.06	52.45
	Combined (Models)	68.20	39.26	55.03
Similar Reasoning	Gold	74.38	54.34	90.27
	GPT-3	51.63	36.61	74.68
	T5-3B	41.93	36.77	70.22
	T5-Large	43.61	42.11	70.00
	Combined (Models)	45.69	38.54	71.77

Table 12: **Generalization Results** - Numbers corresponding to Figure 6.

Figure 13: **Questionnaire for property evaluation:** User will be shown a triplet of question, answer and explanation. Similar as the previous task, user need to answer the first question to get to the second one.

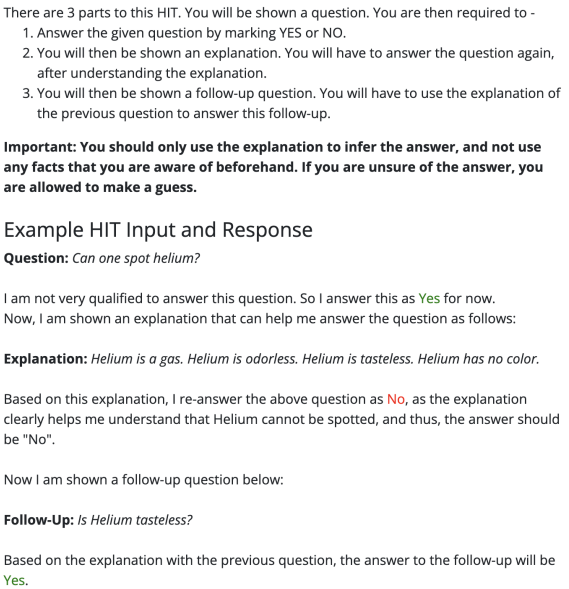

Figure 14: **Instruction for generalization question:** In section 5, generalization questions are divided into 3 types, but in MTurk, we hide this information from annotators. Instruction will help annotators to understand the process and what is follow-up question.

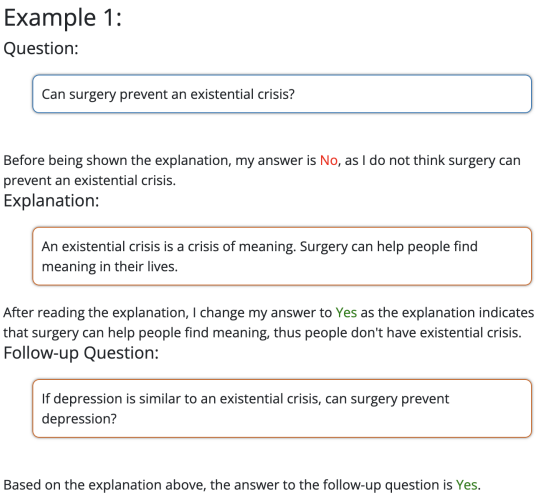

Figure 15: **An example for generalization question:** We demonstration 5 examples in the template. We show how our thinking process change before and after given explanation and how explanation help to answer the follow-up question.

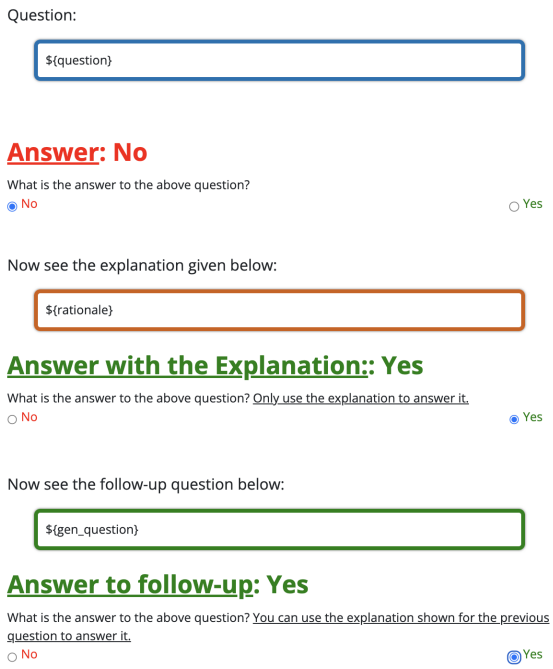

Figure 16: **Questionnaire for generalization question:** In the questionaire, annotators will repeat the steps in human utility evaluations. We repeat it because we cannot make sure annotators took human utility evaluations and annotators took generalization question evaluations will be same group of people. After this, we show them follow-up question and ask them to use the explanation to answer the question.