Are Machine Rationales (Not) Useful to Humans? Measuring and Improving Human Utility of Free-Text Rationales

Brihi Joshi[♣*] Ziyi Liu[♣*] Sahana Ramnath[♣] Aaron Chan[♣] Zhewei Tong[◇]
Shaoliang Nie[♠] Qifan Wang[♠] Yejin Choi[♦♥] Xiang Ren[♣♥]
[♣]University of Southern California [◇]Tsinghua University [♠]Meta AI
[♥]Allen Institute for Artificial Intelligence [♦]University of Washington
{brihijos, zliu2803, sramnath, chanaaro, xiangren}@usc.edu, tzw19@mails.tsinghua.edu.cn
{snie, wqfcr}@meta.com, yejin@cs.washington.edu

Abstract

Among the remarkable emergent capabilities of large language models (LMs) is free-text rationalization; beyond a certain scale, large LMs are capable of generating seemingly useful rationalizations, which in turn, can dramatically enhance their performances on leaderboards. This phenomenon raises a question: can machine generated rationales also be useful for humans, especially when lay humans try to answer questions based on those machine rationales? We observe that human utility of existing rationales is far from satisfactory, and expensive to estimate with human studies. Existing metrics like task performance of the LM generating the rationales, or similarity between generated and gold rationales are not good indicators of their human utility. While we observe that certain properties of rationales like conciseness and novelty are correlated with their human utility, estimating them without human involvement is challenging. We show that, by estimating a rationale's helpfulness in *answering similar unseen instances*, we can measure its human utility to a better extent. We also translate this finding into an automated score, GEN-U, that we propose, which can help improve LMs' ability to generate rationales with better human utility, while maintaining most of its task performance. Lastly, we release all code and collected data with this project.[1]

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). For example, LMs have played a large role in reducing human effort for dataset creation (Bonifacio et al., 2022; Yuan et al., 2021; Liu et al., 2022) and helping humans critique text (Saunders et al., 2022). However, the opaque reasoning processes of these LMs pose serious concerns about their role in high-stakes decision-making (Bender et al., 2021; Doshi-Velez and Kim, 2017). Recently, many works have explored using LMs to generate fluent, human-like *free-text rationales*[2] via natural language (Ehsan et al., 2018; Rajani et al., 2019a) that can explain their decisions. Further, rationales can reference things beyond the task input, and also support high flexibility in content, style, and length (Narang et al., 2020; Wiegreffe et al., 2022, 2021; Chan et al., 2022). However, evaluating if a rationale of a task-instance contains enough knowledge to help lay humans understand and solve that instance correctly is still under-explored.

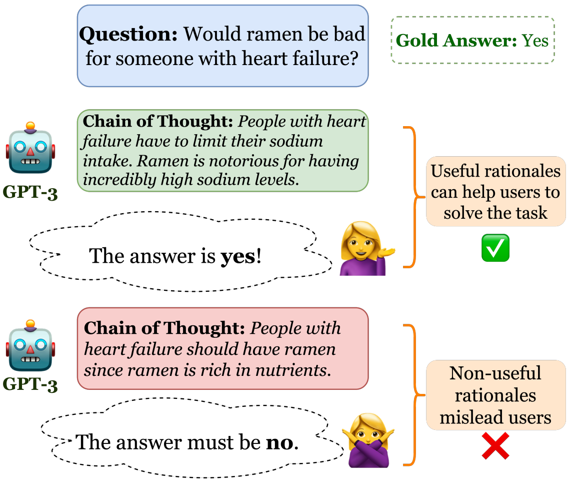

Question: Would ramen be bad for someone with heart failure?
Gold Answer: Yes
GPT-3
Chain of Thought: People with heart failure have to limit their sodium intake. Ramen is notorious for having incredibly high sodium levels.
The answer is yes!
Useful rationales can help users to solve the task
✅
GPT-3
Chain of Thought: People with heart failure should have ramen since ramen is rich in nutrients.
The answer must be no.
Non-useful rationales mislead users
❌

Figure 1: **An illustration of Human Utility of rationales:** Here, we show Chains of Thought (rationales) generated by GPT-3 in two scenarios. The first one is providing knowledge to the human to be able to answer the question, but the second rationale is not useful, and is in fact, misleading the human to answer incorrectly.

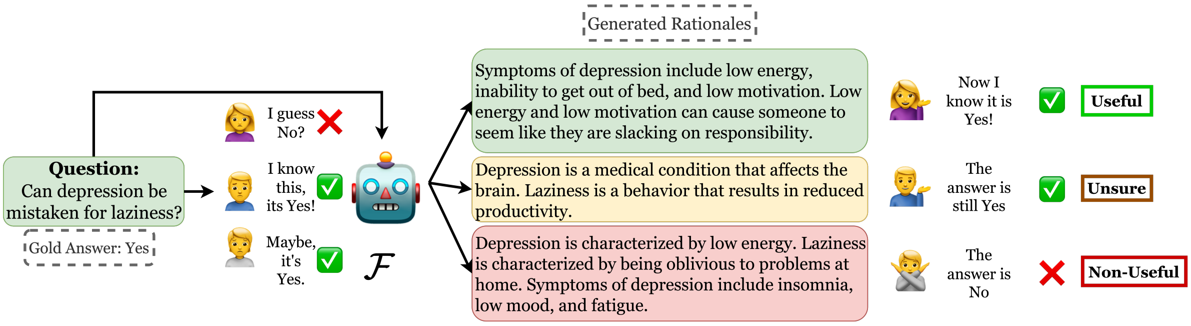
Prior literature for human-AI collaboration has studied plausibility (Wiegreffe and Marasović, 2021). However, plausibility only aims to cap-

[*] Equal contribution.

[1] https://github.com/INK-USC/RationaleHumanUtility

[2] We use the term '*rationales*' throughout the paper to refer to free-text rationales and explanations.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7103–7128
July 9-14, 2023 ©2023 Association for Computational Linguistics



Symptoms of depression include low energy, inability to get out of bed, and low motivation. Low energy and low motivation can cause someone to seem like they are slacking on responsibility.

Depression is a medical condition that affects the brain. Laziness is a behavior that results in reduced productivity.

Depression is characterized by low energy. Laziness is characterized by being oblivious to problems at home. Symptoms of depression include insomnia, low mood, and fatigue.

Figure 2: **An illustration of measuring human utility of machine rationales.** We evaluate whether a human's belief of the answer changes before and after seeing a rationale generated by an LM.

ture human judgement of the rationale supporting LM's predicted label. There has been little work done on evaluating actionable advantages offered by rationales to *lay humans* in understanding a task, despite the promise of human-AI collaboration (Schuff et al., 2022). Studying human utility of rationales is important to not only situate them in real-world use cases beyond the involvement of researchers, but also to bridge the gap between human and AI understanding, specifically in scenarios where AI systems perform better. In this work, we shift the paradigm of rationale evaluation, by investigating *human utility* of rationales in helping lay humans understand and solve a given task correctly.

In our study, we observe that *human utility of current LMs is far from satisfactory (including large LMs like GPT-3)*, with only 20% of generated rationales being actually useful (§2). Given that human evaluations are expensive, we should find a reliable way to measure human utility. We examine the correlation of two straightforward measures like LM task performance and alignment with gold rationales, with human utility and find no usable insights. We also ask humans to evaluate rationales w.r.t eight granular-level properties. While we observe that six out of these eight properties are correlated with human utility, reliably estimating them without human evaluation is still an open question (Golovneva et al., 2022).

In addition to the above observation, we find that *high-utility rationales effectively transfer knowledge to humans for solving new instances.* (§3) We create new instances (*e.g.,* questions) by either paraphrasing the original instance in a nontrivial 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). We observe that useful rationales help humans generalize better to new instances, whereas non-helpful rationales even mislead them to answer incorrectly.

To follow up on the above finding, we show that we can *improve an LM's ability to generate rationales with better human utility.* (§4) We translate this finding into an automated score, GEN-U, that reflects the ability of a rationale to help an LM answer generalization instances, that better correlates with human utility (when compared to other metrics like LMs' task accuracy). We use GEN-U as a reward (Lu et al., 2022) while generating rationales and observe that the updated LM generates 2% more useful rationales and gets rid of 4% misleading rationales than before, via human-subject evaluations, without hindering the LMs' task performance.

This paper presents the first comprehensive study of lay human utility of free-text rationales. By introducing lay humans in the rationale evaluation pipeline, and using their insights into existing LMs, we believe our findings can help guide future work on developing methods for efficient and reliable human-AI collaboration.

2 Human Utility of LM-generated Rationales

We begin by defining human utility, intuitively and formally, and describing the LMs that we use for the rest of the paper. Based on this definition, we conduct human studies to investigate whether existing LMs are capable of generating useful rationales. Finally, we follow this up by identifying granular-level syntactic and semantic properties of rationales can indicate their human utility.

Human Utility of Machine Rationales. We first define human utility of rationales as *the advantage that rationales offer lay humans to solve tasks, that*

Human Utility	Examples
Useful	**Question**: Could Al Capone have read works from the Harlem Renaissance? **Gold Answer**: Yes **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? **Gold Answer**: Yes **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? **Gold Answer**: No **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? **Gold Answer**: No **Rationale**: Gorillas are native to Africa. Panda bears are native to China.
Not Useful	**Question**: Does 50 Cent get along with Jeffrey Atkins? **Gold Answer**: No **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? **Gold Answer**: Yes **Rationale**: Apollo is the god of the sun and the arts. Baldur is the god of light and peace.

Table 1: **Examples of rationales with different human utility from the StrategyQA Dataset:** Shown here are questions, rationales and gold answers, for different rationale types, as evaluated by our human studies.

they are otherwise unable to (Schuff et al., 2022; Idahl et al., 2021; Chu et al., 2020) (Figure 2). In theory, we can estimate human utility of a rationale in a forward simulation-like (Doshi-Velez and Kim, 2017) setup: the difference in human performance of a task, with and without the assistance of a rationale. In this work, we reformulate this definition of utility for a classification task (multi-choice question answering). We use the StrategyQA (Geva et al., 2021) and OBQA (Mihaylov et al., 2018) datasets for our paper. The reason for doing so is to pick tasks where humans are not already better than LMs (unlike NLI and CommonsenseQA (Nangia and Bowman, 2019; Talmor et al., 2021)), and study cases where rationales are capable of knowledge transfer that can help humans. More details about our task and dataset selection reasoning is highlighted in §A.1.

Formal setup for calculating human utility. Let $\mathcal{F}$ be a *self-rationalizing LM* (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.

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:

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

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 as *unsure*. This 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.

Self-rationalizing Models. For our choice of $\mathcal{F}$, we experiment with in-context learning and fine-tuning based approaches. For the rest of our paper, we pick three LM configurations that provide us the best task accuracy for the rest of our experiments in this paper: davinci-instruct-beta (GPT3) (Brown et al., 2020b) with six randomly picked demonstrations, with the FEB (Marasovic et al., 2022) template, where rationales are generated after the predicted answer, T5-large with full fine-tuning and infilling template (Marasovic et al., 2022) and T5-3B with 128-shot fine-tuning and infilling template. Details about prompt templates, experiment settings and model selection are in §A.2.

To what extent do LM-generated rationales provide utility to humans? We conduct human-subject studies to evaluate utility of free-text rationales. We use Amazon Mechanical Turk [3] to

[3] www.mturk.com

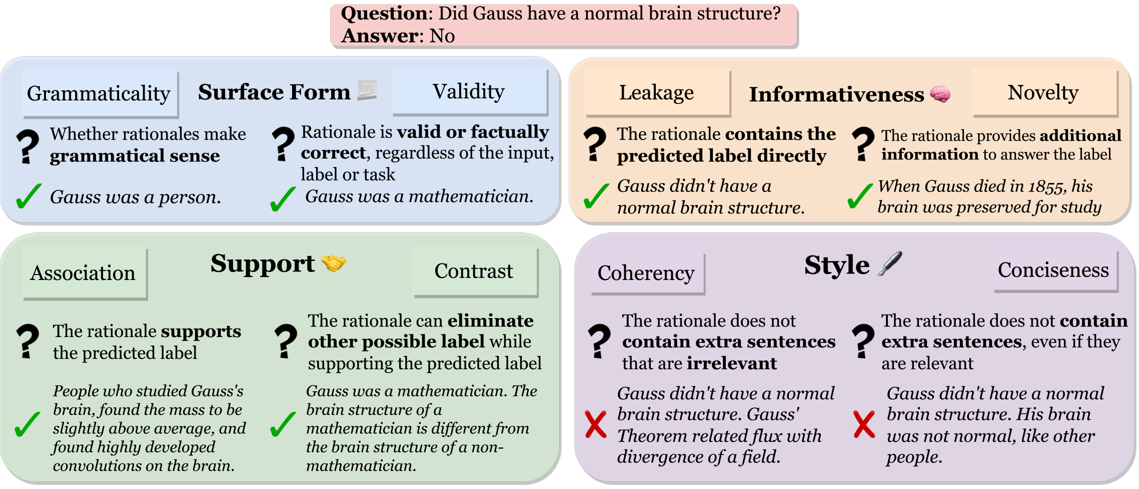

Question: Did Gauss have a normal brain structure?
Answer: No
Surface Form
Grammaticality
Validity
Whether rationales make grammatical sense
Gauss was a person.
Rationale is valid or factually correct, regardless of the input, label or task
Gauss was a mathematician.
Informativeness
Leakage
Novelty
The rationale contains the predicted label directly
Gauss didn't have a normal brain structure.
The rationale provides additional information to answer the label
When Gauss died in 1855, his brain was preserved for study
Support
Association
Contrast
The rationale supports the predicted label
People who studied Gauss's brain, found the mass to be slightly above average, and found highly developed convolutions on the brain.
The rationale can eliminate other possible label while supporting the predicted label
Gauss was a mathematician. The brain structure of a mathematician is different from the brain structure of a non-mathematician.
Style
Coherency
Conciseness
The rationale does not contain extra sentences that are irrelevant
Gauss didn't have a normal brain structure. Gauss' Theorem related flux with divergence of a field.
The rationale does not contain extra sentences, even if they are relevant
Gauss didn't have a normal brain structure. His brain was not normal, like other people.

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.

Dataset	Model	Setting	Test accuracy
STRATEGYQA	T5-LARGE	full-finetuning	67.03
	T5-3B	128-shot	56.70±1.85
	GPT-3-175B	in-context	60.04
OBQA	T5-LARGE	full-finetuning	65.72
	T5-3B	128-shot	56.70±1.85
	GPT-3-175B	in-context	55.60

Table 2: **Self-Rationalising Model Results**: Shown here are the test set accuracies of T5-Large, T3-3B and davinci-instruct-beta (GPT-3) from best settings. We use these three settings for the rest of our work. The results of the complete list of finetuning and in-context learning experiments we performed are shown in Tables 9, 10 and 11.

		% of generated rationales			
Dataset	Type	All	GPT-3-175B	T5-3B	T5-Large
StrategyQA	USEFUL	17.83	20.30	18.12	15.06
	NOT USEFUL	35.00	25.76	35.15	44.10
	UNSURE	47.16	53.93	46.72	40.82
OBQA	USEFUL	15.26	16.06	14.85	14.85
	NOT USEFUL	54.88	54.21	50.60	59.83
	UNSURE	29.85	29.71	34.53	25.30

Table 3: **Distribution of Human Utility of Rationales:** Shown here are the %s of different types of rationales based on their utility, for T5-Large, T5-3B and davinci-instruct-beta (GPT-3), for both StrategyQA and OBQA.

first curate a set of annotators that understand the task well (via extensive qualification tests). Each instance is answered by five annotators. (The annotator agreements are shown in Table 18). For each StrategyQA and OBQA 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. Details about MTurk experiment setup and annotation agreements are in §A.6. For each instance, we calculate human utility as defined above, where predictions made by five annotators are aggregated by taking a majority vote.

We observe that (Table 3) for all the LMs combined only a small amount of rationales generated are actually useful for humans. A large chunk of rationales also mislead humans to select the incorrect answer (NOT USEFUL). In fact, for T5-Large and UnifiedQA-Large, the configuration that led to the best task performance for StrategyQA and OBQA, has the highest % of NOT USEFUL rationales.

		Correlation			
Dataset	Type	Overall	GPT-3-175B	T5-3B	T5-Large
STRATEGYQA	TASK ACCURACY	0.035	**0.111**	0.034	0.005
	BERTSCORE	0.041	**0.021**	0.017	0.002
OBQA	TASK ACCURACY	0.022	**0.092**	0.029	0.016
	BERTSCORE	0.055	0.018	**0.026**	0.017

Table 4: **Correlation between Human Utility of Rationales and Task Performance/BERTScore:** Shown here are the correlation scores between task performance/BERTScore and Human Utility for T5-Large, T5-3B and davinci-instruct-beta(GPT-3). We use Theill's U for Task Performance and Correlation Ration η for BERTScore (Zhang* et al., 2020).

Do existing metrics correlate with human utility? Overall, while including annotations for all models combined, we observe that the correlation

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: 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

Table 5: **Examples of generalization questions of each type from the StrategyQA Dataset**: We show the original question, rationale and label triplet, along with davinci-instruct-beta (GPT-3) generated generalization questions and gold label for the generated question.

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 $U = 0.0359$ and $U = 0.0221$ for StrategyQA and OBQA respectively). This indicates that while generating rationales might improve overall task performance, there is no guarantee that these rationales are 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 (0.092), 0.034 (0.029) and 0.005 (0.016) for StrategyQA (OBQA) respectively (Table 4). 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.

We also compute the similarity between rationales and their corresponding gold rationale using BERTScore (Zhang* et al., 2020) for the test set, and compute their correlation with their human utility (Table 4). For StrategyQA, the Correlation Ratio $\eta = 0.041$ for all three LMs combined, and $\eta = 0.021, 0.017, 0.002$ for GPT-3, T5-3B and T5-Large respectively, whereas for OBQA $\eta = 0.055$ for all three LMs combined, and $\eta = 0.018, 0.026, 0.017$ for GPT-3, T5-3B and T5-Large respectively.

What rationale properties are associated with human utility of rationales? We conduct a case-study for the StrategyQA dataset. We 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.

We use a Generalized Linear Mixed-Effects Model (GLMEM) (similar to Lamm et al. (2020)) to estimate the importance of different properties and their interactions in predicting the human utility of rationales. We observe that while in isolation or pairs, these properties are not sufficient indicators of human utility (§A.3.1), 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.

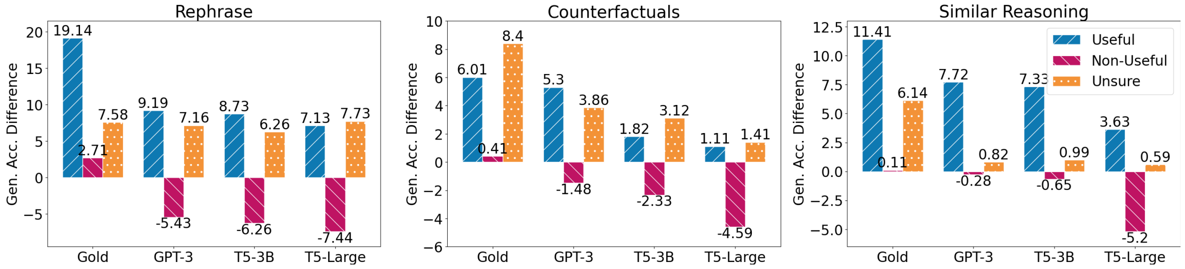


Figure 4: **Generalization Accuracy Difference for the StrategyQA Dataset:** In this Figure, we plot the *difference* in accuracy of generalization questions, after and before a human annotator is shown the original question's rationale.

3 Measuring Rationale Utility by Answering Generalization Questions

As defined in §2, human utility of rationales is determined by their ability to guide humans to correctly solve the task (instances). 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. For all our experiments, we use the StrategyQA Dataset.

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 as 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 use 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 described above, we follow the Human and AI collaboration paradigm for dataset collection 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 21). 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 it at the end of the corresponding template, which is then used as a prompt for GPT-3 to generate questions. To increase the number of good-quality generalization questions, we use GPT-3 to generate 5 generalization questions of each type for a given question, along with their answers. We also vary the temperature (0.7) to control for diversity in generated questions. The generated questions and their answers are then validated by a human study, to make sure that the final set of questions is of good quality (Details in §A.6.2).

In the end, for each original question in the StrategyQA dataset, we obtain generalization questions of three different types, although the number of generalization questions per original question can vary. Overall, we collected 9659, 1164 and 2608 generalization questions for the training, validation and test set, with 5.86, 6.32 and 5.70 generalization questions per original question on average, respectively.

Human generalization is a good indicator of human utility. Similar to §2, we first ask the annotators to answer a generalization question without the rationale. We then show them the rationale of the original question, and ask them to answer the generalization question again, taking the rationale into account. 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.

In Figure 4, we plot the difference between the generalization accuracies after and before being

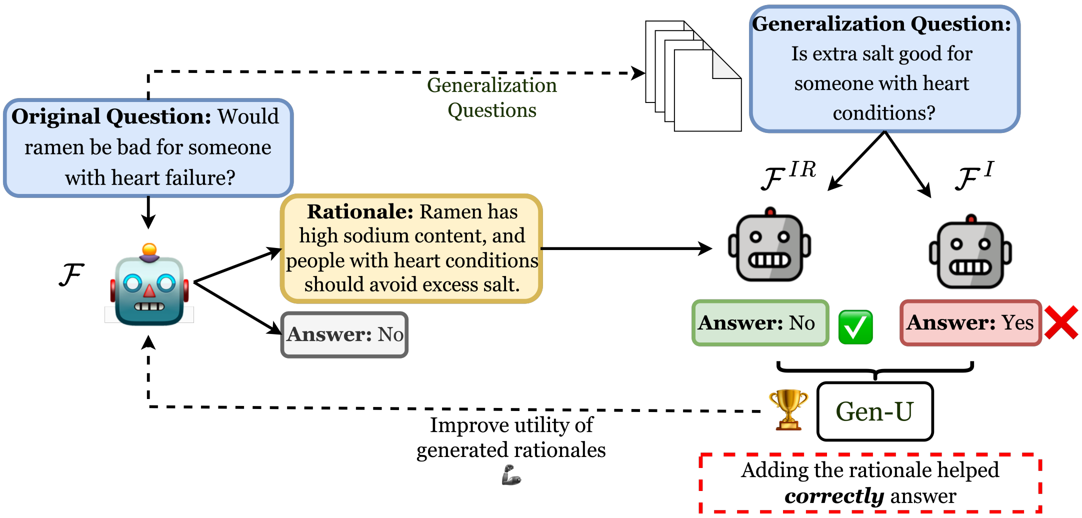

Original Question: Would ramen be bad for someone with heart failure?
Generalization Questions
Generalization Question: Is extra salt good for someone with heart conditions?
$\mathcal{F}$
Rationale: Ramen has high sodium content, and people with heart conditions should avoid excess salt.
Answer: No
$\mathcal{F}^{IR}$
$\mathcal{F}^{I}$
Answer: No ✓
Answer: Yes ✗
Gen-U
Improve utility of generated rationales
Adding the rationale helped *correctly* answer

Figure 5: **Updating self-rationalising LMs with GEN-U:** Based on the generalization ability of two other LMs, we use GEN-U to update $\mathcal{F}$, so as to generate rationales with better utility.

shown the rationale of the original question. We observe that gold rationales form an upper bound for generalization difference, across all types of generalization questions and types of rationale utility. Useful rationales are able to help humans generalize better to new instances, whereas non-useful rationales often *mislead* humans to make incorrect choices, who might have correctly answered the question before, which is indicated by the *negative* plot bars in the Figure. Rationales about which we are unsure are better or close to useful rationales for rephrase and counterfactuals, as these generalization questions are relatively simpler.

However, for similar reasoning questions, they underperform useful rationales. This indicates that for rationales that are unsure, either the human was already aware of the answer or the questions are easier to answer as humans are able to answer rephrases and counterfactuals correctly, but fail in generalizing to questions that follow a similar reasoning process. We can also note that GPT-3 generated rationales help generalize better to more difficult settings like counterfactuals or similar reasoning questions. Examples of generalization questions that were answered correctly/incorrectly for rationales that have high or low human utility is shown in the Appendix (Table 19).

4 Improving Human Utility of Self-Rationalising LMs

Smaller LMs like T5-large have better task accuracy, but lack in generating more useful rationales. It can be observed (§2) that the task performance of a self-rationalizing LM and the human utility of its corresponding generated rationales are not correlated. Based on our insights about how useful rationales can help humans generalization to unseen questions, we propose GEN-U, which simulates a human through an LM: we define and use GEN-U to improve human utility of smaller LMs like T5-large, while aiming to maintain their task accuracy (Figure 5). For all our experiments, we use the StrategyQA Dataset.

LM generalization is a better indicator of rationale's human utility. §3 indicated that generalization to unseen but similar questions via rationales of the original question is a reasonable proxy for human utility of rationales. Based on this insight, we propose GEN-U, which estimates the generalization performance of an LM variant, after and before being shown a rationale generated by a self-rationalizing model.

For a given input-output pair x, y, there exist a set of n generalization questions $X_g, Y_g = \{(x_{g1}, y_{g1}), (x_{g2}, y_{g2}), \ldots, (x_{gn}, y_{gn})\}$ that is created as per §3. Let $\mathcal{F}$ be a self-rationalising LM as defined in §2, for which we want to estimate the score. Let $\mathcal{F}^{\mathcal{I}}$ be an LM that takes in X_g as its input and predicts a set of labels Y_g^I. Similarly, $\mathcal{F}^{\mathcal{IR}}$ be an LM that takes in X_g and the rationale r_p generated for x by $\mathcal{F}$, and predicts a set of labels Y_g^{IR}.

GEN-U for x is defined as:

$$\text{MODE}_{i=1:n}\left(\begin{cases}\left(1-\mathbb{1}(y_{gi}^{I}=y_{gi})\right) & y_{gi}^{IR}=y_{gi}\\ -1 & y_{gi}^{IR}\neq y_{gi}\end{cases}\right)$$

Here, Mode returns the most frequently occurring value from the set (similar to majority voting in a set). In other words, if a generalization question is answered incorrectly after being shown the rationale, Gen-U is -1, otherwise, Gen-U calibrates itself w.r.t the answer before being shown the rationale, to accommodate for cases where the question is easy-to-answer or the LM already contains relevant background knowledge. Then, we pick the majority vote of the scores (depicted by the mode) for all the generalization questions for a given original question as its score.

To validate if Gen-U is indeed usable, we calculate correlations between Gen-U and human utility of the corresponding rationales. We find that Theill's $U = 0.22$, which is indicates that Gen-U is a better estimate that $\mathcal{F}$'s task accuracy or BERTScore similarity between generated and gold rationales (refer Table 6 for correlation scores).

Metric	Gen-U	Task Accuracy	BERTScore
Correlation	0.227	0.035	0.041

Table 6: **Improvement in Correlation Scores for the StrategyQA Dataset:** We observe that Gen-U leads to a better correlation with human utility than Task Accuracy or BERTScore.

Gen-U as a reward for updating LM. We use the Quark (Lu et al., 2022) algorithm with Gen-U to improve the human utility of rationales generated by $\mathcal{F}$. Quark is an RL-inspired training algorithm that uses reward signals as control tokens on the encoder (or decoder) side, to condition the generation of text.

For $\mathcal{F}$, we use the same T5-large setup used in §2. For implementing Gen-U, we use T5-base LMs for $\mathcal{F}^{\mathcal{I}}$ and $\mathcal{F}^{\mathcal{IR}}$, which are both finetuned on the StrategyQA dataset. We begin by first finetuning $\mathcal{F}$ for 25 epochs with supervised learning on the StrategyQA data, after which we continue training with Quark. The final $\mathcal{F}^{'}$ is obtained after finding the best hyperparameter choices based on Gen-U scores for the validation set.

	$\mathcal{F}$	$\mathcal{F}^{'}$ (w/ Quark)	GPT-3-175B
Gen-U	-0.315	-0.26 ↑	-
Task Accuracy	67.03	65.06 ↓	60.04
% Useful	15.06	17.01 ↑	20.30
% Not Useful	44.10	40.20 ↑	25.76
% Unsure	40.82	42.79 ↓	53.93
# of Params	770M	770M	175B

Table 7: **Impact of Gen-U as a reward to update LM using Quark (Lu et al., 2022) algorithm:** On the StrategyQA Dataset, we show the % of different types of rationales for the LM before ($\mathcal{F}$) and after ($\mathcal{F}^{'}$) being updated with generation feedback through the Quark algorithm, using Gen-U as the reward. We also note the % of rationales for davinci-instruct-beta (GPT-3), which is the best performing variant in terms of human utility. Here, ↑ implies improvement seen in $\mathcal{F}^{'}$, and vice versa.

Table 7 demonstrates the Gen-U scores before and after using Quark to update $\mathcal{F}$. On the updated LM $\mathcal{F}^{'}$, we conduct the same human utility evaluations as done in §2 to evaluate the improvement observed by lay humans. We note that the updated LM is able to retain most of the task performance, while improving the % of Useful rationales by 2%. Gen-U also helps in getting rid of 4% of mislead (Not Useful) rationales. We also compare the updated LM with GPT-3, which yielded the best human utility of rationales. Gen-U is able to make the updated LM closer to the human utility of GPT-3, while ensuring the task performance for the updated LM remains better than GPT-3. This indicates that while incorporating human utility while generating rationales is a difficult problem and there is room for improvement, smaller LMs like T5-large are capable of improving, without compromising on the task accuracy that is obtained via fine-tuning.

5 Related Work

Evaluating free-text rationales Extractive explanations have been used to improve human's understanding of the model (Wang and Yin, 2021; Feng and Boyd-Graber, 2018; Carton et al., 2020; Chen et al., 2022b; Idahl et al., 2021; Chu et al., 2020) or detecting errors in model predictions (González et al., 2021). Although prior motivation of generating 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 rationales in various ways. Wiegreffe et al. (2022) use human acceptability judgements on over-generated rationales by GPT-3 (Brown et al., 2020a). They also evaluate the rationales across seven axes like grammar, factuality, *etc*. Sun et al. (2022) measure benefits of rationales to LMs and compared human written rationales with those generated by GPT-3 across two axes: rationales that provide new information over the input, and those that leak the label directly.

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., 2020a; Wang and Komatsuzaki, 2021) that are pretrained 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).

Updating LMs with Generation Feedback There are several ways to update language models with rewards to correct misaligned behaviour that models learn (Chen et al., 2021; Janner et al., 2021). Lu et al. (2022) unlearn these misalignments by fine-tuning the language model on signals of what not to do. Similarly, Zelikman et al. (2022) iteratively leverage a small number of rationale examples to training and only keep good examples. Our method is inspired by several evaluation methods (Chen et al., 2022a; Chan et al., 2022; Wiegreffe et al., 2020; Hase et al., 2020) which discussed how to better evaluate the quality of free-text rationales with regard to labels and contexts.

6 Conclusion and Future Work

In this work, we study human utility of free-text rationales, by measuring how well lay humans are able to solve tasks with their help. Through extensive human evaluations, we show that human utility of rationales generated by current LMs is rather unsatisfactory, and existing available measures do not correlate well with it. We find that generalization ability with rationales as context is a good proxy for human utility, and use it as a reward to improve human utility of LMs.

There are a lot of scopes to improve human utility of self-rationalising LMs, where granular-level properties of rationales can be leveraged directly. Furthermore, evaluation of human utility on other tasks (like closed-book QA) is something that is also worth looking at, given that human annotators cannot 'guess' answers for these tasks, making it harder for LMs and humans alike.

7 Acknowledgments

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract Nos. 2019-19051600007 and 2022-22072200006, Defense Advanced Research Projects Activity (DARPA) No. HR00112220046, NSF IIS 2048211, and gift awards from Google, Amazon, JP Morgan, and Sony. We would like to thank all of our collaborators at USC NLP Group, USC INK Research Lab, Meta AI and AI2, specially Swabha Swayamdipta and Ameya Godbole for their constructive feedback on this work.

Limitations

Estimating human utility is expensive. The core of our work is built on conducting extensive human evaluations, to understand how well lay humans can solve tasks with rationales. In order to replicate these findings to other tasks, one would require the same scale of human evaluations, which are expensive and tedious. These tasks are also difficult to explain to lay crowdworkers, because of which several rounds of turking are required to reach good annotator agreements. Given these shortcomings of human evaluation, a reliable metric that estimates human utility is necessary.

Generating generalization questions is not completely automated. Even though we prompt GPT-3 with varied demonstrations to generate generalization questions of each type, we still have to manually filter them (via crowdsourcing) to obtain a cleaner set of questions. Furthermore, in order to obtain gold answers of these questions, we generate answers by prompting GPT-3 again, which also requires further validation. A completely automated method of generating these questions would lead LM updates to be independent of human involvement.

Even though GEN-U has a better correlation with human utility, the correlation is still low. To train models to produce free-text rationales with more human-utility through Quark (Lu et al., 2022), it is first necessary to have an accurate metric that can serve as a reward function/scoring metric for human utility. In this work, we found that human generalization is good indicator of human-utility. However, given that Quark requires frequent reward scoring, it is infeasible to use human annotations for the same. Our proposed automatic metric GEN-U that simulates human generalization has a good correlation with human utility (better than task accuracy, or BERTScore), but overall, it still has a low correlation with human utility of rationales. Developing a score with better correlation with human utility (perhaps even a stronger version of GEN-U) will decrease the effect of this limitation and lead to training that further increases human utility of generated rationales.

Ethics Statement

Data. All the datasets that we use in our work are released publicly for usage and have been duly attributed to their original authors. Data for all human studies that we conduct is publicly released with this work, with appropriate annotator anonymisations.

Crowdsourcing. All our crowdworkers are from countries where English is the primary language. For all our human studies, the task is setup in a manner that ensure that the annotators receive compensation that is above minimum wage ($15/hour). Since we conduct extensive qualification tasks before annotations, crowdworkers that participate in the qualification are compensated more than the task, given the time taken to read and understand task instructions and examples. Furthermore, we ensure that we correspond with crowdworkers over email to address their queries. Crowdworkers have also been given bonuses for flagging errors in the task, or consistently providing good-quality annotations.

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.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2387–2392, New York, NY, USA. Association for Computing Machinery.

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. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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. 2022. Frame: Evaluating rationale-label consistency metrics for free-text rationales. *arXiv preprint arXiv:2207.00779*.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2022a. Rev: Information-theoretic evaluation of free-text rationales. *arXiv preprint arXiv:2210.04982*.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097.

Valerie Chen, Nari Johnson, Nicholay Topin, Gregory Plumb, and Ameet Talwalkar. 2022b. 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.

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? *arXiv preprint arXiv:2010.04119*.

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.

Michael Janner, Qiyang Li, and Sergey Levine. 2021. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286.

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*.

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?

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.

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022. QUARK: Controllable text generation with reinforced unlearning. In *Advances in Neural Information Processing Systems*.

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.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*.

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.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators.

Hendrik Schuff, Heike Adel, Peng Qi, and Ngoc Thang Vu. 2022. How (not) to evaluate explanation quality. *arXiv preprint arXiv:2210.07126*.

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

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.

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.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 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. We also use the OpenBookQA Dataset (Mihaylov et al., 2018) for validating human utility of rationales for existing LMs. Both of these datasets are available publicly for use, and have been checked manually by authors for toxic/offensive content.

A.2 Self-Rationalising 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,

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 6: **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).

Split	Train	Dev	Test
Number	1648	184	458

Table 8: **Dataset details**: Since the original test set of StrategyQA does not have gold labels, we used only the original train set and validation set in our experiments. Our test set is the original validation set, and our train and validation sets are splits $(90/10\%)$ from the original train set.

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., 2020a). 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 6.

As seen in Tables 9, 10 and 11, for the StrategyQA and OBQA datasets, 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).

Task Performance. For the three selected best performing LM configurations, we note (Tables 9, 10) 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).

A.2.1 Self-Rationalising Models Training Details

In the experiments, we mainly used 3 models: T5-Large, T5-3B, and GPT-3 (model details and hyper-

$\mathcal{F}$	Model	Size	Finetuning setting	Accuracy			
				SQuAD-T5	Infilling	QA-simple	T5-like
Without Rationale	StrategyQA	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
	OBQA	large	full	71	65.8	69	70
		3B	48-shot	64.33 ± 2.30	61.87 ± 3.01	68.40 ± 0.69	63.93 ± 3.63
		3B	128-shot	68.27 ± 4.12	67.27 ± 1.53	71.20 ± 2.11	67.13 ± 0.42
With Rationale	StrategyQA	large	full	61.14	67.03	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
	OBQA	large	full	70.20	70.20	67.20	70.40
		3B	48-shot	62.67 ± 2.34	63.07 ± 2.72	67.93 ± 4.84	66.60 ± 1.64
		3B	128-shot	67.47 ± 3.16	66.07 ± 2.66	70.40 ± 2.31	69.00 ± 0.53

Table 9: **Self-Rationalising Model Results (Fine-tuning)**: Shown here are test set accuracies of LMs (T5) of different sizes (large and 3B), 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.

$\mathcal{F}$	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 10: **Self-Rationalising Model Results (In-Context Learning) for StrategyQA Dataset**: Shown here are test set accuracies of davinci-instruct-beta (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.

$\mathcal{F}$	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	53.60
		max len	Randomly	55.60
	FEB	6	Randomly	40.40
		max len	Randomly	41.20

Table 11: **Self-Rationalising Model Results (In-Context Learning) for OBQA Dataset**: Shown here are test set accuracies of davinci-instruct-beta (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.

parameters are shown in Table 12). For T5-Large, we used the full train set for finetuning. For T5-3B, we trained in 2 settings: 48-shot and 128-shot. We used 3 seeds for generating shots for T5-3B. For GPT-3, we only used the OpenAI GPT-3 API (Brown et al., 2020b) to do inference.

A.3 Property Analysis

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 property annotations reach low to moderate agreement across all annotators (Table 13).

Presence of properties in Gold and LM-generated Rationales We first study the presence of these properties in rationales, without considering the utility of these rationales. Figure 7 plots the distribution of these properties, split by

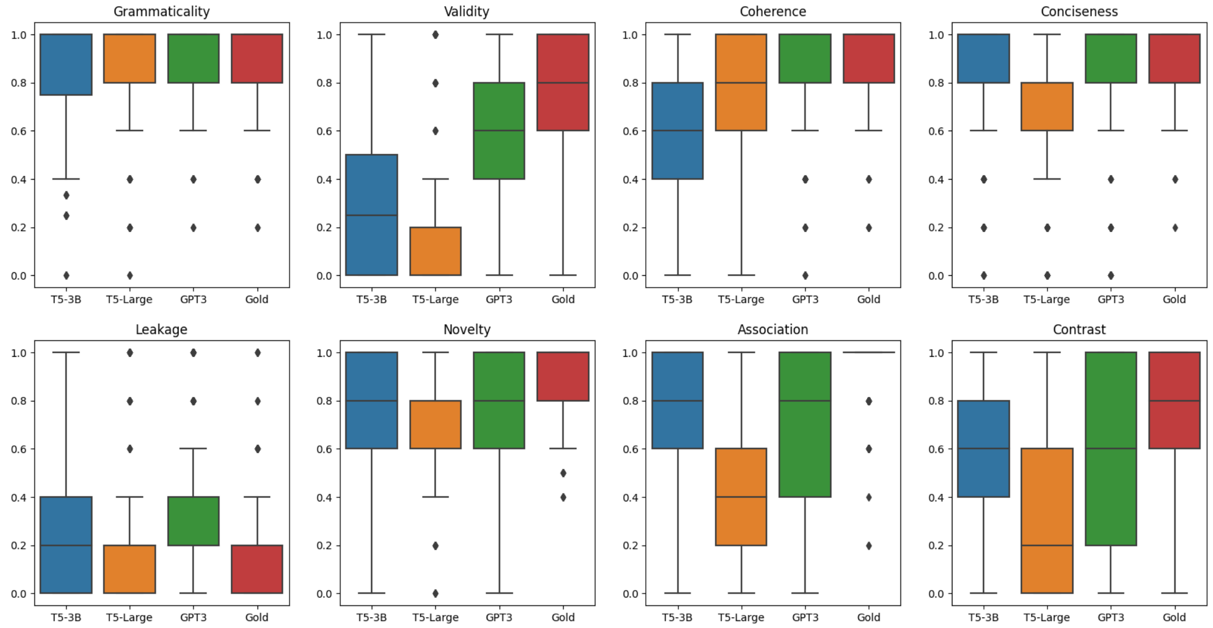


Figure 7: **Distribution of Property Annotations for Different Rationales:** Distribution is generated by aggregating scores of five annotators of each instance. A higher value implies more presence of the property in the rationale generated by the particular LM.

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
GPU	Quadro RTX 6000
Training time	2 hours

Table 12: **Self-Rationalising Models Training Details**: Here we show the models we used and hyperparameters we used for T5-3B and T5-Large model training.

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 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.

A.3.1 Property Correlations with Human Utility

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 human 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.

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

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 14: **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 15: **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.

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

We then look at pairwise interactions. Table 15 shows the 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.

A.4 Quark training details

For the Quark experiments, we used T5-Large as the self-rationalizing LM, and T5-Base for GEN-U. The hyperparameters used for running Quark (Lu et al., 2022) are shown in Table 16.

A.5 Examples

In Table 21 we provide the demonstrations used to generate generalization questions using GPT-3. In Table 19, we provide examples of useful, unsure and non-useful rationales with respect to human generalization. In Table 20 (corresponding to Figure 4) we provide results for the difference in accuracies of human generalization, before and after a human annotator was shown the original question's rationale.

Hyperparameter	Value
Optimizer	Adam
Adam epsilon	$1e$-8
Adam initial learning-rate	$1e$-5
Learning-rate scheduler	linear with warmup
Warmup steps	1000
Gradient clipping	1.0
Gradient accumulation	2 steps
KL-divergence coef.	0.05
Entropy regularization coef.	0.05
Sampling rate	2 samples for every train sample
Frequency of exploration	every 500 steps
Sampling strategy	Top-p (0.7) sampling
Temperature for sampling	1.0
Number of distinct reward-bins	3 (1, 0 and -1)
Train batch-size	4
Eval batch-size	64
Training time	5-6 hours

Table 16: Quark training details

A.6 MTurk Details

In this section, we describe the MTurk experiment setup. The details of MTurk experiments including how many Turkers took the evaluation, and average time used to finish evaluations are shown in Table 17. Each MTurk annotator is paid above minimum wage. Figure 8 demonstrates the setup for human utility evaluation. Figure 9 demonstrates the setup for property evaluation. Figures 10 demonstrates the setup for validating generalization questions. Figure 11 demonstrate the setup for utility evaluation towards generalization questions.

Since the dataset we used is carefully annotated by human, we can assure there is no toxic content and our experiment setup was submitted to IRB for ethical review. We limited our Turkers to English speaking nations - United States, Canada, Australia, New Zealand and United Kingdom.

To ensure the quality of evaluation, we did a round of qualification task before each task which include a small set of evaluations. Turkers need to finish the qualification task first and get results of it, then we will show them the whole task.

A.6.1 Worker Selection and Quality Control

Here, we describe details about how workers are selected and how annotations are ensured to be clean. First, we employ multiple rounds of trials before deploying the actual task so as to get feedback from annotators whether they understand the task correctly. This includes in-house tests, tested via Amazon Turk Sandbox [4] and small batches tested on Turk. Second, we create a set of medium to hard qualification tasks for each task that the annotators have to work on. These tasks are hand curated that cater certain parts of the instruction – whether the annotators are reading the rationale correctly, or whether they are able to make appropriate connectections between the rationale and the question. This weeds out a lot of annotators who do not understand the task or are cheating. We also weed out workers who are too 'fast' (completing the task in less than 5 seconds, which is indicative of potential slacking in the task). Third, we constantly monitor task responses and feedback provided to annotators about their task. We also collect feedback from them which we adapt in new versions of the task.

A.6.2 Turking for Generalization Questions

Each generalization question is validated by 3 annotators each. The validation process includes: checking if the generated question can be answered by the gold rationale, answering the generated question, and checking if the generated question follows the instructions for a given type (being a rephrase, counterfactual or a similar reasoning question). The annotation agreement observed here is high (Krippendorf's α = 0.68).

A.6.3 Annotation Agreements

We observe that StrategyQA instances are difficult to annotate by humans, as many 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 α = 0.18). However, *after* being shown the rationale, the agreement increases, as shown in Table 18. Examples of rationales annotated into each of the three human utility categories (useful, not useful, unsure) is shown in Table 1.

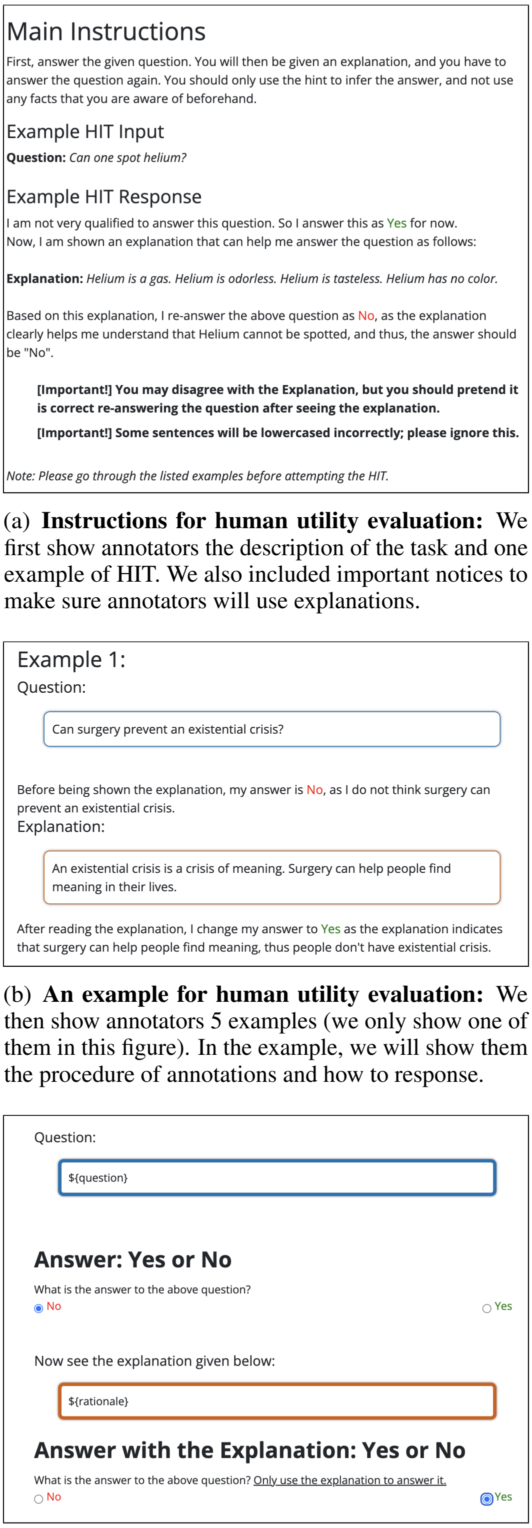
Tasks	Number of Turkers	Average Time(s)
Human Utility Evaluation	80	37.41
Property Evaluation	137	36.50
Generalization Question	25	35.93

Table 17: **Details of MTurk:** Shown here are number of unique Turkers (annotators) and average time of solving one HIT for each task

[4] https://requester.mturk.com/developer/sandbox

Model	GPT-3	T5-3B	T5-Large
Krippendorf's α	0.47	0.30	0.24

Table 18: **Annotators agreement**:Shown here is the annotators agreement. davinci-instruct-beta (GPT-3) has the best agreement even though its task performance is low. Contrastly, T5-Large has highest task performance but a low agreement.



Main Instructions

First, answer the given question. You will then be given an explanation, and you have to answer the question again. You should only use the hint to infer the answer, and not use any facts that you are aware of beforehand.

Example HIT Input

Question: *Can one spot helium?*

Example HIT Response

I am not very qualified to answer this question. So I answer this as Yes for now.
Now, I am shown an explanation that can help me answer the question as follows:

Explanation: *Helium is a gas. Helium is odorless. Helium is tasteless. Helium has no color.*

Based on this explanation, I re-answer the above question as No, as the explanation clearly helps me understand that Helium cannot be spotted, and thus, the answer should be "No".

> **[Important!] You may disagree with the Explanation, but you should pretend it is correct re-answering the question after seeing the explanation.**
>
> **[Important!] Some sentences will be lowercased incorrectly; please ignore this.**

Note: Please go through the listed examples before attempting the HIT.

(a) **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.

Example 1:

Question:

Can surgery prevent an existential crisis?

Before being shown the explanation, my answer is No, as I do not think surgery can prevent an existential crisis.

Explanation:

An existential crisis is a crisis of meaning. Surgery can help people find meaning in their lives.

After reading the explanation, I change my answer to Yes as the explanation indicates that surgery can help people find meaning, thus people don't have existential crisis.

(b) **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.

Question:

${question}

Answer: Yes or No

What is the answer to the above question?

● No ○ Yes

Now see the explanation given below:

${rationale}

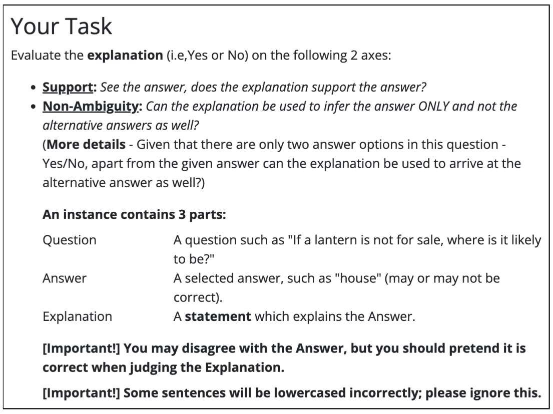
Answer with the Explanation: Yes or No

What is the answer to the above question? Only use the explanation to answer it.

○ No ● Yes

(c) **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 8: **The whole process for human utility evaluation**



Your Task

Evaluate the **explanation** (i.e,Yes or No) on the following 2 axes:

- **Support**: *See the answer, does the explanation support the answer?*
- **Non-Ambiguity**: *Can the explanation be used to infer the answer ONLY and not the alternative answers as well?*
(**More details** - Given that there are only two answer options in this question - Yes/No, apart from the given answer can the explanation be used to arrive at the alternative answer as well?)

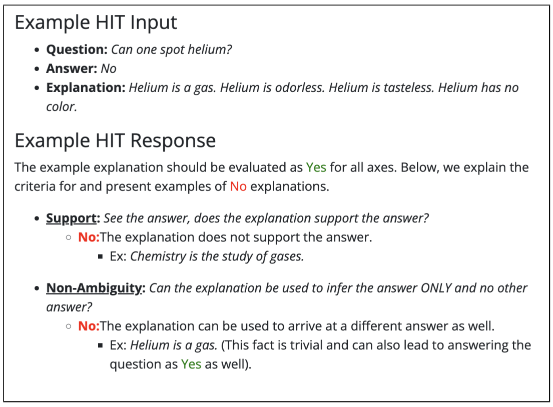
An instance contains 3 parts:

Question	A question such as "If a lantern is not for sale, where is it likely to be?"
Answer	A selected answer, such as "house" (may or may not be correct).
Explanation	A **statement** which explains the Answer.

[Important!] You may disagree with the Answer, but you should pretend it is correct when judging the Explanation.

[Important!] Some sentences will be lowercased incorrectly; please ignore this.

(a) **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



Example HIT Input

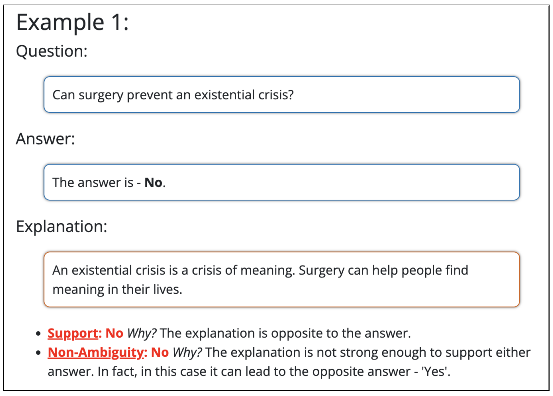
- **Question:** *Can one spot helium?*
- **Answer:** *No*
- **Explanation:** *Helium is a gas. Helium is odorless. Helium is tasteless. Helium has no color.*

Example HIT Response

The example explanation should be evaluated as Yes for all axes. Below, we explain the criteria for and present examples of No explanations.

- **Support**: *See the answer, does the explanation support the answer?*
  - **No:**The explanation does not support the answer.
    - Ex: *Chemistry is the study of gases.*
- **Non-Ambiguity**: *Can the explanation be used to infer the answer ONLY and no other answer?*
  - **No:**The explanation can be used to arrive at a different answer as well.
    - Ex: *Helium is a gas.* (This fact is trivial and can also lead to answering the question as Yes as well).

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



Example 1:

Question:

Can surgery prevent an existential crisis?

Answer:

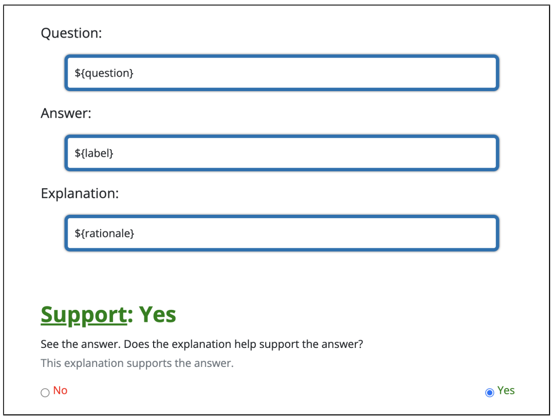
The answer is - **No**.

Explanation:

An existential crisis is a crisis of meaning. Surgery can help people find meaning in their lives.

- **Support**: **No** *Why?* The explanation is opposite to the answer.
- **Non-Ambiguity**: **No** *Why?* The explanation is not strong enough to support either answer. In fact, in this case it can lead to the opposite answer - 'Yes'.

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



Question:

${question}

Answer:

${label}

Explanation:

${rationale}

Support: Yes

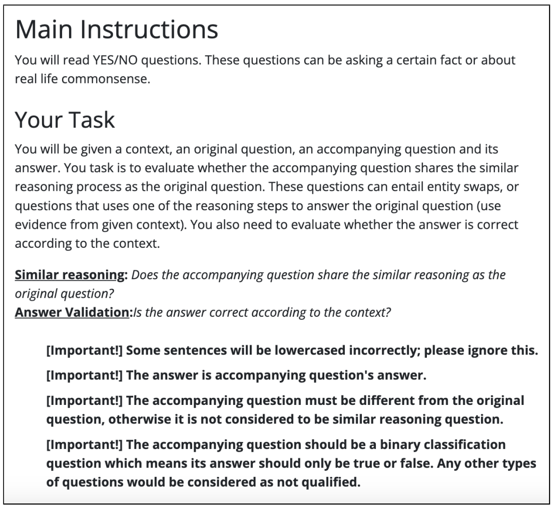
See the answer. Does the explanation help support the answer?

This explanation supports the answer.

○ No ● Yes

(d) **Questionnaire for property evaluation:** Annotators 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 9: **One example of property evaluation questionnaire**: For other properties, they have the similar templates, with different instructions and examples.



Main Instructions

You will read YES/NO questions. These questions can be asking a certain fact or about real life commonsense.

Your Task

You will be given a context, an original question, an accompanying question and its answer. You task is to evaluate whether the accompanying question shares the similar reasoning process as the original question. These questions can entail entity swaps, or questions that uses one of the reasoning steps to answer the original question (use evidence from given context). You also need to evaluate whether the answer is correct according to the context.

Similar reasoning: *Does the accompanying question share the similar reasoning as the original question?*
Answer Validation:*Is the answer correct according to the context?*

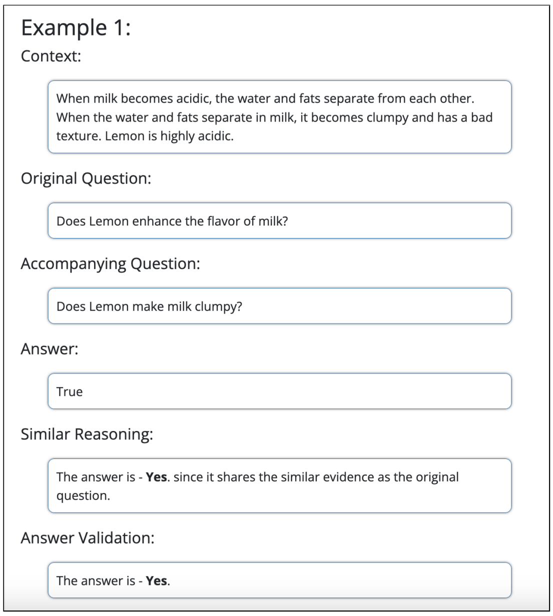
[Important!] Some sentences will be lowercased incorrectly; please ignore this.

[Important!] The answer is accompanying question's answer.

[Important!] The accompanying question must be different from the original question, otherwise it is not considered to be similar reasoning question.

[Important!] The accompanying question should be a binary classification question which means its answer should only be true or false. Any other types of questions would be considered as not qualified.

(a) **Instruction for validation of generalization questions (similar reasoning):** We asked the annotators to validate if the related question is a similar reasoning question.



Example 1:

Context:

When milk becomes acidic, the water and fats separate from each other. When the water and fats separate in milk, it becomes clumpy and has a bad texture. Lemon is highly acidic.

Original Question:

Does Lemon enhance the flavor of milk?

Accompanying Question:

Does Lemon make milk clumpy?

Answer:

True

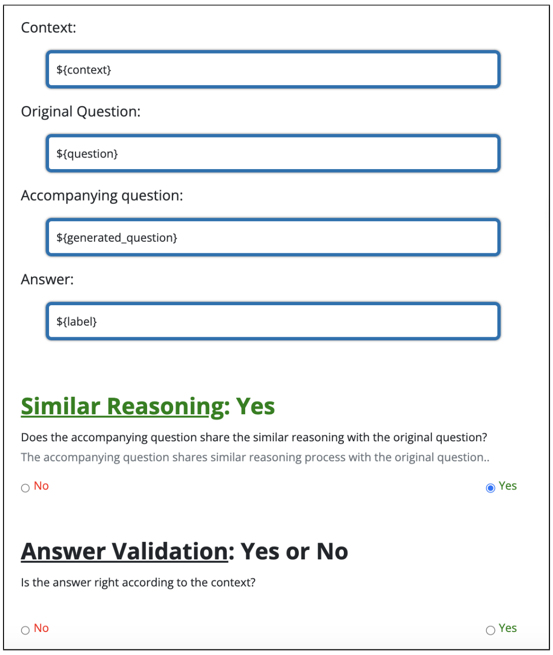
Similar Reasoning:

The answer is - **Yes**. since it shares the similar evidence as the original question.

Answer Validation:

The answer is - **Yes**.

(b) **Example for validation of generalization questions (similar reasoning):** We selected 3 examples in the template to clarify the definition of similar reasoning.



Context:

${context}

Original Question:

${question}

Accompanying question:

${generated_question}

Answer:

${label}

Similar Reasoning: Yes

Does the accompanying question share the similar reasoning with the original question?
The accompanying question shares similar reasoning process with the original question..

○ No ● Yes

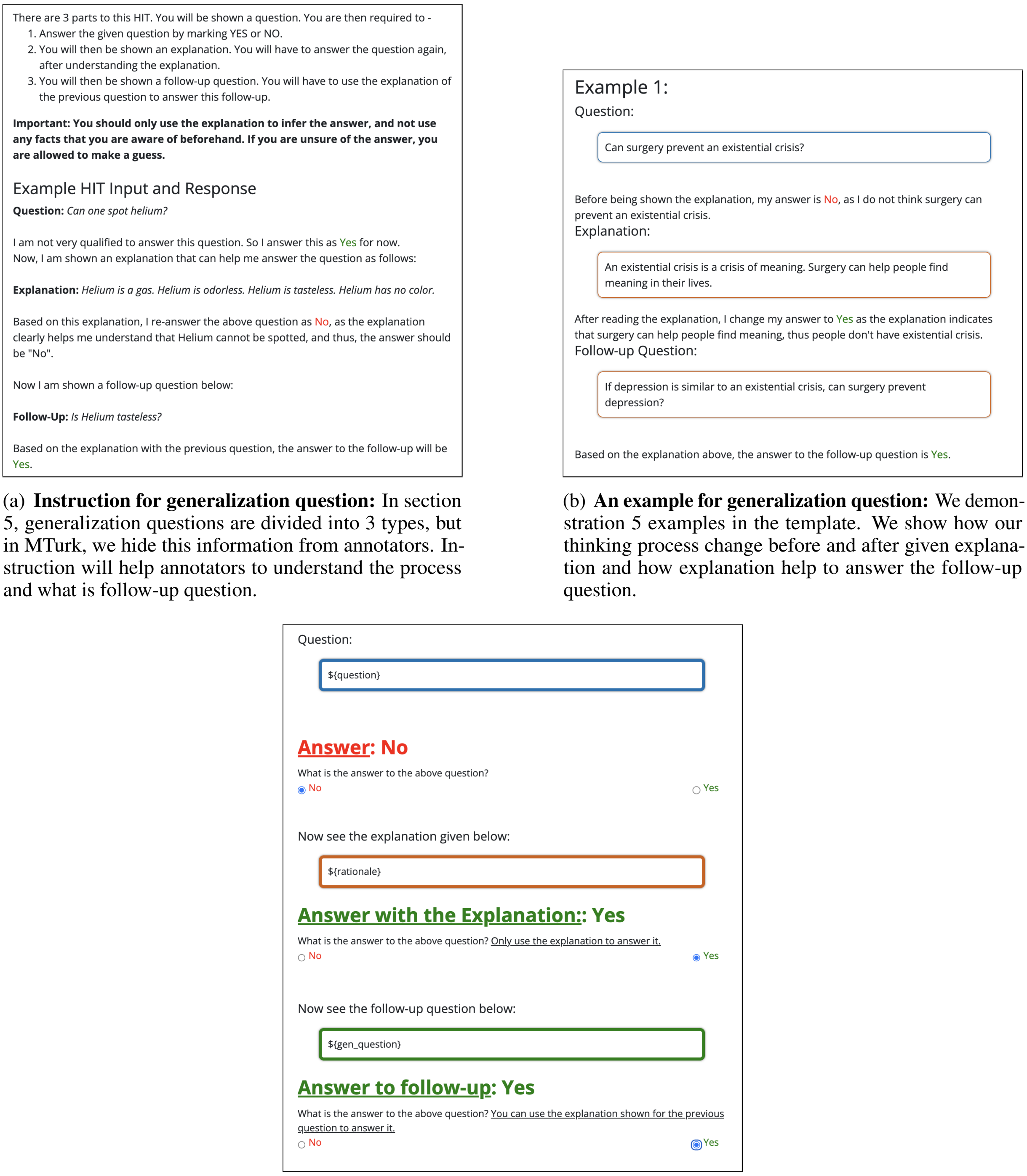
Answer Validation: Yes or No

Is the answer right according to the context?

○ No ○ Yes

(c) **Questionnaire for validation of generalization questions (similar reasoning):** In the questionnaire, annotators need to validate whether the related question is a similar reasoning question then validate the answer of the related question.

Figure 10: **Validation of generalization question**: Rephrase and counterfactual have the similar setup, except for the answer validation. We assume that rephrase questions should have the same answer of original ones and the counterfactual questions should have the opposite answer.



There are 3 parts to this HIT. You will be shown a question. You are then required to -

1. Answer the given question by marking YES or NO.
2. You will then be shown an explanation. You will have to answer the question again, after understanding the explanation.
3. You will then be shown a follow-up question. You will have to use the explanation of the previous question to answer this follow-up.

Important: You should only use the explanation to infer the answer, and not use any facts that you are aware of beforehand. If you are unsure of the answer, you are allowed to make a guess.

Example HIT Input and Response

Question: *Can one spot helium?*

I am not very qualified to answer this question. So I answer this as Yes for now.
Now, I am shown an explanation that can help me answer the question as follows:

Explanation: *Helium is a gas. Helium is odorless. Helium is tasteless. Helium has no color.*

Based on this explanation, I re-answer the above question as No, as the explanation clearly helps me understand that Helium cannot be spotted, and thus, the answer should be "No".

Now I am shown a follow-up question below:

Follow-Up: *Is Helium tasteless?*

Based on the explanation with the previous question, the answer to the follow-up will be Yes.

(a) **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.

Example 1:

Question:

Can surgery prevent an existential crisis?

Before being shown the explanation, my answer is No, as I do not think surgery can prevent an existential crisis.

Explanation:

An existential crisis is a crisis of meaning. Surgery can help people find meaning in their lives.

After reading the explanation, I change my answer to Yes as the explanation indicates that surgery can help people find meaning, thus people don't have existential crisis.

Follow-up Question:

If depression is similar to an existential crisis, can surgery prevent depression?

Based on the explanation above, the answer to the follow-up question is Yes.

(b) **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.

Question:

${question}

Answer: No

What is the answer to the above question?

● No ○ Yes

Now see the explanation given below:

${rationale}

Answer with the Explanation:: Yes

What is the answer to the above question? Only use the explanation to answer it.

○ No ● Yes

Now see the follow-up question below:

${gen_question}

Answer to follow-up: Yes

What is the answer to the above question? You can use the explanation shown for the previous question to answer it.

○ No ● Yes

(c) **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.

Figure 11: **Measuring rationale utility by answering generalization questions**

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 19: **Examples of rationales for Section 3**: 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.

		Generalization Accuracy		
Type of Generalization Questions	**Model**	**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 20: **Generalization Results** - Numbers corresponding to Figure 4.

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 ship2̌019s 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 21: **Demonstrations for generating generalization questions**: For each category, we used 6 fixed demonstrations. We used different questions for each category.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☑ A2. Did you discuss any potential risks of your work?
Section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2,3 Appendix A.2, A.7

☑ B1. Did you cite the creators of artifacts you used?
Appendix A.2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A.2

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix A.2

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Appendix A.2

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix A.7

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix A.2

C ☑ Did you run computational experiments?

Section 2,3,4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A.3.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.3.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 2

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A.3.1

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 2,3,4; Appendix A.7

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A.7

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix A.7

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix A.7

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Appendix A.7

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix A.7